On the uniform continuity of the rate-distortion function

Hari Palaiyanur and Anant Sahai
Department of Electrical Engineering and Computer Sciences
University of California at Berkeley
Berkeley, CA 94720
{hpalaiya, sahai}@eecs.berkeley.edu

***Abstract*— It is well known that the rate-distortion function for a finite alphabet IID source with distribution p, denoted $R(p, D)$, is uniformly continuous in its arguments. We prove an explicit bound on $|R(p, D) - R(q, D)|$ for distributions p, q in terms of the variational distance $\|p-q\|_1$. A simple and elementary proof shows that $|R(p, D) - R(q, D)| = O(-\|p - q\|_1 \log \|p - q\|_1)$, with constants depending on the distortion measure. The uniform continuity of the rate-distortion function has the same behavior as the uniform continuity of entropy in the order sense. The bounds are used for several applications. First, an algorithm is presented to compute the rate-distortion function for an arbitrarily varying source to within a given accuracy. Second, we comment on the problem of approximating the rate-distortion function for an unknown IID source to within a desired precision.**

I. Introduction

Shannon [1] showed the necessary and sufficient rate to code an IID source with distribution p to distortion D is quantified by the rate-distortion function $R(p, D)$. For a finite alphabet source with distribution p under an average distortion measure, $R(p, D)$ is a relatively well understood function of the distortion D. For a fixed p, $R(p, D)$ is convex $\cup$, monotonically decreasing and differentiable almost everywhere as a function of D [2]. For a fixed D, $R(p, D)$ is continuous in p and because the set of distributions on a finite set is compact, $R(p, D)$ is also uniformly continuous. Intuitively, continuity of the rate-distortion function allows a coding system to deal with uncertainty in the underlying distribution in a principled way when a certain target fidelity must be met. This is in contrast to studying the continuity of the distortion-rate function, where we think of having a fixed rate 'budget' and would like to bound how the distortion changes because of uncertainty in the distribution.

As a function of p, one expects $R(p, D)$ to behave like entropy, but this intuition is not always correct. For example, Ahlswede [3] shows that $R(p, D)$ may not be concave in the distribution, while entropy is. In fact, $R(p, D)$ may even have multiple local maxima with different values. However, we show that the uniform continuity of $R(p, D)$ behaves like the uniform continuity of entropy, namely for two distributions p, q with $\|p - q\|_1$ small, $|R(p, D) - R(q, D)| = O(-\|p - q\|_1 \ln \|p - q\|_1)$.

An explicit bound on the uniform continuity of $R(p, D)$ is useful in the computation of the rate-distortion function for an arbitrarily varying source (AVS) ([4], [5]). For several adversarial models of lossy source coding, the rate-distortion function has the form

$$R(D) = \max_{p \in \mathcal{Q}} R(p, D), \tag{1}$$

where $\mathcal{Q}$ is a convex set of distributions. Since $R(p, D)$ is not generally quasi-concave in p, standard convex optimization techniques cannot be used to compute $R(D)$. A brute force approach would be to sample a large number of points in $\mathcal{Q}$, compute $R(p, D)$ for them and take the maximum to give an approximation for $R(D)$. The explicit bound on the uniform continuity of $R(p, D)$ translates readily to a bound on the number of distributions to sample to provably compute $R(D)$ to within some additive error $\epsilon > 0$.

The uniform continuity bound also is useful in the problem of estimating the rate-distortion function for an unknown finite alphabet IID source from its samples. Here, we wish to guarantee that the estimated rate-distortion function with n samples is within $\epsilon > 0$ of the true rate-distortion function with probability at least $1 - \tau$. The question is: How many samples are required and how does the number of samples required scale with ϵ and τ as they go to zero? We give a sufficient number of samples based on the uniform continuity bound.

The paper is organized as follows. Definitions are given in Section II and the main result is stated in Section III. We present the applications of the continuity result to AVS rate-distortion computation and rate-distortion function estimation in Sections IV-A and IV-B respectively. Proofs of the main result are given in Sections V and VI.

II. Definitions

Let $\mathcal{X}$ be a finite source alphabet and $\widehat{\mathcal{X}}$ a finite reproduction alphabet. Let $\mathcal{P}(\mathcal{X})$ be the set of distributions (probability mass functions) on $\mathcal{X}$. Let the entropy of a distribution p be $H(p) = -\sum_{x \in \mathcal{X}} p(x) \ln p(x)$. For $p, q \in \mathcal{P}(\mathcal{X})$, the $\mathcal{L}_1$ or variational distance between p and q is $\|p - q\|_1 = \sum_{x \in \mathcal{X}} |p(x) - q(x)|$.

This work was supported by an NSF Graduate Research Fellowship.

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

$$\widetilde{d} = \min_{(x,\widehat{x}):d(x,\widehat{x})>0} d(x, \widehat{x}). \tag{2}$$

Also, for a $p \in \mathcal{P}(\mathcal{X})$, let $D_{\min}(p) = \sum_{x\in\mathcal{X}} p(x) \min_{\widehat{x}\in\widehat{\mathcal{X}}} d(x, \widehat{x})$. We say that condition (Z), meaning zero distortion, holds for the distortion measure if

$$(Z)\ \forall\ x \in \mathcal{X}, \exists\ \widehat{x}_0(x) \text{ s.t. } d(x, \widehat{x}_0(x)) = 0.$$

In words, (Z) is the condition that the distortion matrix has a zero in every row.

Let $\mathcal{W}$ be the set of probability transition matrices (channels) from $\mathcal{X}$ to $\widehat{\mathcal{X}}$. For $p \in \mathcal{P}(\mathcal{X})$ and $W \in \mathcal{W}$, let

$$d(p, W) = \sum_{x,\widehat{x}} p(x)W(\widehat{x}|x)d(x, \widehat{x}) \tag{3}$$

be the average distortion of source p across channel W. Then, $R(p, D)$ is defined to be[1]

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

where $\mathcal{W}(p, D) = \{W \in \mathcal{W} : d(p, W) \le D\}$ and $I(p, W)$ is the mutual information (in nats)

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

If (Z) holds, we have $R(p, D) < \infty$ for all $D \ge 0$, $p \in \mathcal{P}(\mathcal{X})$. We let

$$W^*_{p,D} = \arg \min_{W\in\mathcal{W}(p,D)} I(p, W) \tag{5}$$

and note that $W^*_{p,D}$ may not be unique (see Problem 3 of Section 2.3 of [6]). For our purposes, it will be sufficient that at least one $W^*_{p,D}$ exists, which is true by the continuity of $I(p, W)$ in W and the fact that $\mathcal{W}(p, D)$ is closed.

III. Uniform Continuity of $R(p, D)$

In [7], the following bound on the uniform continuity of entropy is given.

Lemma 1: For $p, q \in \mathcal{P}(\mathcal{X})$, if $\|p - q\|_1 \le 1/2$,

$$|H(p) - H(q)| \le \|p - q\|_1 \ln \frac{|\mathcal{X}|}{\|p - q\|_1} \tag{6}$$

Using Lemma 1, one can prove bounds on the uniform continuity[2] of $R(p, D)$. First, we state the bound for distortion measures when condition (Z) holds.

Lemma 2: Suppose condition (Z) holds. For $p, q \in \mathcal{P}(\mathcal{X})$, if $\|p - q\|_1 \le \frac{\widetilde{d}}{4d^*}$, for any $D \ge 0$,

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

Proof: See Section V. ■

Now let $d(\cdot, \cdot)$ be an arbitrary distortion measure with maximum distortion d^* for which (Z) may not hold. We can form a new distortion measure,

$$d_0(x, \widehat{x}) = d(x, \widehat{x}) - \min_{\widehat{x}'\in\widehat{\mathcal{X}}} d(x, \widehat{x}'). \tag{8}$$

We let $\widetilde{d}_0$ be the minimal nonzero distortion for $d_0(\cdot, \cdot)$. We have defined $d_0(\cdot, \cdot)$ so that (Z) holds and hence Lemma 2 applies to it. The rate-distortion functions with respect to the two different distortion measures, $d(\cdot, \cdot)$ and $d_0(\cdot, \cdot)$, are related in a simple way, as discussed in Section VI. If we are careful about when the rate-distortion function is infinite, this allows us to extend Lemma 2 to the following lemma by taking on a larger constant.

Lemma 3: For $p, q \in \mathcal{P}(\mathcal{X})$ with $\|p - q\|_1 \le \widetilde{d}_0/4d^*$, if $D \ge \max\{D_{\min}(p), D_{\min}(q)\}$,

$$|R(p, D) - R(q, D)| \le \frac{11d^*}{\widetilde{d}_0} \|p - q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p - q\|_1}. \tag{9}$$

Proof: See Section VI. ■

These lemmas give an upper bound on the behavior of $|R(p, D) - R(q, D)|$ as $\|p - q\|_1$ goes to 0, namely $O(-\|p - q\|_1 \ln \|p - q\|_1)$. For binary distributions $(1-p, p)$ and $(1-q, q)$ with $p, q \in [0, 1/2]$, the rate-distortion function under Hamming distortion measure is $R((1 - p, p), D) = h_b(p) - h_b(D)$ for $D \in [0, p]$ where $h_b(\cdot)$ is the binary entropy function. Hence, for $D \le \min\{p, q\}$,

$$|R((1 - p, p), D) - R((1 - q, q), D)| = |h_b(p) - h_b(q)|. \tag{10}$$

While Lemmas 2 and 3 are loose in the constants, equation (10) and Lemma 1 shows that they are correct in the order sense.

For use in the next section, we let $f(\gamma) = \gamma \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\gamma}$ for $\gamma \in [0, 1/2]$, with $f(0) = 0$ by continuity. Also, we let $g : f([0, 1/2]) \to [0, 1/2]$ be the inverse function of f (which exists because f is monotonically increasing and continuous on $[0, 1/2]$).

IV. Applications

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

In [4] and [5], the problem of lossy source coding is considered for an arbitrarily varying source. An AVS is composed of finite alphabet subsources with known distributions. One of the subsources outputs a letter at each time that must be encoded, and the choice of which subsource outputs the letter varies arbitrarily over time. In [4], it was shown that if the variation was chosen in a way that did not depend on the realizations of the subsources, the rate-distortion function for such a source is

$$R(D) = \max_{p\in\overline{\mathcal{G}}} R(p, D), \tag{11}$$

where $\overline{\mathcal{G}}$ is the convex hull of the distributions on the subsources. In [5], it was shown that if the variation is allowed to

[1] By convention, $R(p, D) = \infty$ if $\mathcal{W}(p, D) = \emptyset$.

[2] Note to reviewers: We have not been able to find this kind of bound in the Western literature or standard textbooks. We feel it may be known somewhere, and if you are aware of such a result, please let us know.

depend on the realizations of the subsources, the rate-distortion function of the AVS is

$$R(D) = \max_{p\in\mathcal{D}} R(p,D), \tag{12}$$

where $\mathcal{D}$ is a convex set of distributions (actually a polytope) with $\overline{\mathcal{G}} \subset \mathcal{D}$. While $R(p,D)$ can be computed by the Arimoto-Blahut algorithm or convex optimization techniques, $R(D)$ is not easily computed by standard algorithms in either (11) or (12). We now give a simple brute-force algorithm that provably computes $R(D)$ to within any desired precision $\epsilon > 0$ when $R(D)$ is of the form

$$R(D) = \max_{p\in\mathcal{Q}} R(p,D), \tag{13}$$

where $\mathcal{Q}$ is a convex set defined by a finite number of linear inequalities.

Without loss of generality, assume $\mathcal{X} = \{1,2,\ldots,|\mathcal{X}|\}$. Let $\gamma \in (0,1)$ and let $\gamma\mathbb{Z}^{|\mathcal{X}|-1}$ be the $|\mathcal{X}|-1$ dimensional integer lattice scaled by γ. Let $\widetilde{\mathcal{O}} = [0,1]^{|\mathcal{X}|-1}\bigcap\gamma\mathbb{Z}^{|\mathcal{X}|-1}$. Now, define

$$\mathcal{O} = \left\{ q\in\mathcal{P}(\mathcal{X}) : \begin{array}{c} \exists\ \widetilde{q}\in\widetilde{\mathcal{O}}, \\ q(i) = \widetilde{q}(i), i = 1,\ldots,|\mathcal{X}|-1, \\ q(|\mathcal{X}|) = 1 - \sum_{i=1}^{|\mathcal{X}|-1}\widetilde{q}(i) \geq 0 \end{array} \right\}. \tag{14}$$

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

1) Fix a $q\in\mathcal{O}$. If $\min_{p\in\mathcal{Q}}\|p-q\|_1 \leq 2|\mathcal{X}|\gamma$, compute $R(q,D)$ using the Arimoto-Blahut algorithm [7], otherwise do not compute $R(q,D)$. Repeat for all $q\in\mathcal{O}$.
2) Let $R_a(D)$ be the maximum of the computed values of $R(q,D)$, i.e.

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

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

Lemma 4: The preceding algorithm computes an approximation $R_a(D)$ such that $|R_a(D) - R(D)| \leq \epsilon$ if

$$\gamma = \frac{1}{2|\mathcal{X}|}g\left(\frac{\epsilon\widetilde{d}_0}{11d^*}\right). \tag{15}$$

The number of distributions for which $R(q,D)$ is computed to determine $R(D)$ to within accuracy ϵ is at most[3]

$$N(\epsilon) \leq \left(\frac{2|\mathcal{X}|}{g\left(\frac{\epsilon\widetilde{d}_0}{11d^*}\right)} + 2\right)^{|\mathcal{X}|-1}. \tag{16}$$

Proof: The bound on $N(\epsilon)$ is clear as there are at most $(\lceil 1/\gamma\rceil + 1)^{|\mathcal{X}|-1}$ points in $\widetilde{\mathcal{O}}$.

Now, we prove $|R_a(D) - R(D)| \leq \epsilon$. First, for all $p\in\mathcal{Q}$, there is a $q\in\mathcal{O}$ with $\|p-q\|_1 \leq g\left(\frac{\epsilon\widetilde{d}_0}{11d^*}\right) = 2|\mathcal{X}|\gamma$. To see this, let $\widetilde{q}(i) = \lfloor\frac{p(i)}{\gamma}\rfloor\gamma$ for $i = 1,\ldots,|\mathcal{X}|-1$. Then $\widetilde{q}\in\widetilde{\mathcal{O}}$, and we let $q(i) = \widetilde{q}(i)$ for $i = 1,\ldots,|\mathcal{X}|-1$. It is readily checked that $q\in\mathcal{O}$ and $\|p-q\|_1 \leq 2\gamma|\mathcal{X}|$. By Lemma 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 3 again gives $R(q,D) \leq \max_{p\in\mathcal{Q}} R(p,D) + \epsilon$. ■

B. Estimating $R(p,D)$ for an unknown IID source

Recently, Harrison and Kontoyiannis [8] have studied the problem of estimating the rate-distortion function of the marginal distribution of an unknown source. Let $p_{\mathbf{x}^n}$ be the (marginal) empirical distribution of a vector $\mathbf{x}^n \in \mathcal{X}^n$. They show that the 'plug-in' estimator, $R(p_{\mathbf{x}^n}, D)$, the rate-distortion function of the empirical marginal distribution of a sequence, is a consistent estimator for a large class of sources beyond just IID sources with known alphabets. However, if the source is known to be IID with alphabet size $|\mathcal{X}|$, estimates of the convergence rate (in probability) of the estimator can be provided using the uniform continuity of the rate-distortion function.

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

Lemma 5: Suppose (Z) holds. For any $p\in\mathcal{P}(\mathcal{X})$, $\tau\in(0,1)$, and $\epsilon\in(0,\ln|\mathcal{X}|)$,

$$P(|R(p_{\mathbf{x}^n},D) - R(p,D)| \geq \epsilon) \leq \tau \tag{17}$$

if

$$n > \frac{2}{g\left(\frac{\epsilon\widetilde{d}}{7d^*}\right)^2}\left(\ln\frac{1}{\tau} + |\mathcal{X}|\ln 2\right). \tag{18}$$

Proof: From Lemma 2, we have

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

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

The last line follows from Theorem 2.1 of [9]. 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. Setting $\delta = \tau$ and 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 (18) depends only on the distortion measure, alphabet sizes $|\mathcal{X}|$ and $|\widehat{\mathcal{X}}|$, desired accuracy ϵ and 'estimation error' probability τ.

V. Proof of Lemma 2

Proof: The quantity of interest is

$$|R(p,D) - R(q,D)| = |I(p, W^*_{p,D}) - I(q, W^*_{q,D})|. \quad (19)$$

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

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

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

$$\begin{aligned} R(p, d(p, W^*_{q,D})) &\leq I(p, W^*_{q,D}) \\ &\leq I(q, W^*_{q,D}) + \\ &\qquad |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| \\ &= R(q, D) + \\ &\qquad |I(p, W^*_{q,D}) - I(q, W^*_{q,D})|. \quad (20) \end{aligned}$$

Expanding mutual informations yields

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

Above, for a distribution p on $\mathcal{X}$ and channel W from $\mathcal{X}$ to $\widehat{\mathcal{X}}$, $H(pW)$ denotes the entropy of a distribution on $\widehat{\mathcal{X}}$ with probabilities $(pW)(\widehat{x}) = \sum_x p(x) W(\widehat{x}|x)$. $H(p, W)$ denotes the entropy of the joint source on $\mathcal{X} \times \widehat{\mathcal{X}}$ with probabilities $(p, W)(x, \widehat{x}) = p(x) W(\widehat{x}|x)$. It is straightforward to verify that $\|pW - qW\|_1 \leq \|p - q\|_1$ and $\|(p, W) - (q, W)\|_1 \leq \|p - q\|_1$. So using Lemma 1 three times, we have

$$\begin{aligned} |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| &\leq \|p - q\|_1 \ln \frac{|\mathcal{X}|}{\|p - q\|_1} + \\ &\quad \|p - q\|_1 \ln \frac{|\widehat{\mathcal{X}}|}{\|p - q\|_1} + \\ &\quad \|p - q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p - q\|_1} \\ &\leq 3\|p - q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p - q\|_1}. \end{aligned}$$

Now, we have seen $d(p, W^*_{q,D}) \leq D + d^* \|p - q\|_1$. We will use the uniform continuity of $R(p, D)$ in D to bound $|R(p, D) - R(p, D + d^*\|p - q\|_1)|$. This will give an upper bound on $R(p, D) - R(q, D)$ as seen through equation (20), namely,

$$\begin{aligned} R(p, D) - R(q, D) &\leq |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| + \\ &\quad R(p, D) - R(p, d(p, W^*_{q,D})) \quad (21) \\ &\leq |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| + \\ &\quad R(p, D) - R(p, D + d^*\|p - q\|_1), \end{aligned}$$

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

$$|R(p, D_1) - R(p, D_2)| \leq |R(p, 0) - R(p, |D_2 - D_1|)|. \quad (22)$$

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

$$\begin{aligned} D &\geq \sum_x p(x) \sum_{\widehat{x}} W(\widehat{x}|x) d(x, \widehat{x}) \quad (23) \\ &= \sum_x p(x) \sum_{\widehat{x}:\ d(x,\widehat{x})>0} W(\widehat{x}|x) d(x, \widehat{x}) \quad (24) \\ &\geq \widetilde{d} \sum_x p(x) \sum_{\widehat{x}:\ d(x,\widehat{x})>0} W(\widehat{x}|x). \quad (25) \end{aligned}$$

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

$$W_0(\widehat{x}|x) = V(\widehat{x}|x) + \frac{\sum_{\hat{x}':\ d(x,\hat{x}')>0} W(\hat{x}'|x)}{|\{\hat{x}':\ d(x, \hat{x}') = 0\}|}. \quad (26)$$

For all $(x, \widehat{x})$ with $d(x, \widehat{x}) > 0$, define $W_0(\widehat{x}|x) = 0$. So, W_0 is a valid channel in $\mathcal{W}(p, 0)$. For a fixed $x \in \mathcal{X}$, it can easily be checked that

$$\sum_{\widehat{x}} |W(\widehat{x}|x) - W_0(\widehat{x}|x)| = 2 \sum_{\widehat{x}:\ d(x,\widehat{x})>0} W(\widehat{x}|x).$$

Therefore, using (25)

$$\sum_x p(x) \sum_{\widehat{x}} |W(\widehat{x}|x) - W_0(\widehat{x}|x)| \leq \frac{2D}{\widetilde{d}}. \quad (27)$$

So, for $W = W^*_{p,D}$, there is a $W_0 \in \mathcal{W}(p, 0)$ with the above 'modified $\mathcal{L}_1$ distance' with respect to p between W and W_0

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

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

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

$$\begin{aligned} \|pW_0 - pW^*_{p,D}\|_1 &= \sum_{\widehat{x}} \Bigg| \sum_x p(x) W_0(\widehat{x}|x) - \\ &\qquad p(x) W^*_{p,D}(\widehat{x}|x) \Bigg| \qquad (28) \\ &\leq \sum_x p(x) \sum_{\widehat{x}} |W_0(\widehat{x}|x) - W^*_{p,D}(\widehat{x}|x)| \\ &\leq \frac{2D}{\widetilde{d}}. \qquad (29) \end{aligned}$$

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

Now, assuming $D \leq \widetilde{d}/4$, we can again invoke Lemma 1 to get

$$\begin{aligned} |R(p,0) - R(p,D)| &\leq \frac{2D}{\widetilde{d}} \ln \frac{\widetilde{d}|\mathcal{X}|}{2D} + \frac{2D}{\widetilde{d}} \ln \frac{\widetilde{d}|\mathcal{X}||\widehat{\mathcal{X}}|}{2D} \\ &\leq \frac{4D}{\widetilde{d}} \ln \frac{\widetilde{d}|\mathcal{X}||\widehat{\mathcal{X}}|}{2D}. \qquad (30) \end{aligned}$$

Going back to (21), we see that if $\|p-q\|_1 \leq \frac{\widetilde{d}}{4d^*}$,

$$\begin{aligned} \epsilon &\triangleq |R(p, D + d^*\|p-q\|_1)) - R(p,D)| \qquad (31) \\ &\leq \frac{4d^*\|p-q\|_1}{\widetilde{d}} \ln \frac{\widetilde{d}|\mathcal{X}||\widehat{\mathcal{X}}|}{2d^*\|p-q\|_1} \qquad (32) \\ &\leq \frac{4d^*\|p-q\|_1}{\widetilde{d}} \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1}. \qquad (33) \end{aligned}$$

The last step follows because $\widetilde{d}/d^* \leq 1$. Substituting into equation (21) gives

$$\begin{aligned} R(p,D) - R(q,D) &\leq 3\|p-q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1} + \\ &\qquad 4\frac{d^*}{\widetilde{d}}\|p-q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1} \qquad (34) \\ &\leq \frac{7d^*}{\widetilde{d}}\|p-q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1}. \qquad (35) \end{aligned}$$

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.

$$R(q,D) - R(p,D) \leq \frac{7d^*}{\widetilde{d}}\|p-q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1}. \qquad (36)$$

■

VI. Proof of Lemma 3

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

$$d_0(x,\widehat{x}) = d(x,\widehat{x}) - \min_{\widetilde{x} \in \widehat{\mathcal{X}}} d(x,\widetilde{x}) \qquad (37)$$

so that Lemma 2 applies to d_0. Let $R_0(p,D)$ be the IID rate-distortion function for $p \in \mathcal{P}(\mathcal{X})$ at distortion D with respect to distortion measure $d_0(x,\widehat{x})$. By definition, $R(p,D)$ is the IID rate-distortion function for p with respect to distortion measure $d(x,\widehat{x})$. From Problem 13.4 of [7], for any $D \geq D_{\min}(p)$,

$$R(p,D) = R_0(p, D - D_{\min}(p)). \qquad (38)$$

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

$$\begin{aligned} \beta &\triangleq |R(p,D) - R(q,D)| \qquad (39) \\ &= |R_0(p, D - D_{\min}(p)) - R_0(q, D - D_{\min}(q)| \qquad (40) \\ &\leq |R_0(p, D - D_{\min}(p)) - R_0(p, D - D_{\min}(q))| + \\ &\quad |R_0(p, D - D_{\min}(q)) - R_0(q, D - D_{\min}(q))|. \qquad (41) \end{aligned}$$

Now, we note that $|D_{\min}(p) - D_{\min}(q)| \leq d^*\|p-q\|_1$. The first term of equation (41) can be bounded using equation (30) and the second term of (41) can be bounded using Lemma 2. The first term can be bounded if $\|p-q\|_1 \leq \widetilde{d}_0/4d^*$ and the second can be bounded if $\|p-q\|_1 \leq \widetilde{d}_0/4d^*_0$. Since $d^*_0 \leq d^*$, we only require $\|p-q\|_1 \leq \widetilde{d}_0/4d^*$.

$$\begin{aligned} \beta &\leq \frac{4d^*}{\widetilde{d}_0}\|p-q\|_1 \ln \frac{\widetilde{d}_0|\mathcal{X}||\widehat{\mathcal{X}}|}{2d^*\|p-q\|_1} + \\ &\qquad \frac{7d^*_0}{\widetilde{d}_0}\|p-q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1} \qquad (42) \\ &\leq \frac{11d^*}{\widetilde{d}_0}\|p-q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1}. \end{aligned}$$

■

References

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

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

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

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

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

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

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

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

[9] 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