Estimation of the Rate-Distortion Function

Matthew T. Harrison, *Member, IEEE* and Ioannis Kontoyiannis, *Senior Member, IEEE*

Abstract

Motivated by questions in lossy data compression and by theoretical considerations, we examine the problem of estimating the rate-distortion function of an unknown (not necessarily discrete-valued) source from empirical data. Our focus is the behavior of the so-called "plug-in" estimator, which is simply the rate-distortion function of the empirical distribution of the observed data. Sufficient conditions are given for its consistency, and examples are provided to demonstrate that in certain cases it fails to converge to the true rate-distortion function. The analysis of its performance is complicated by the fact that the rate-distortion function is not continuous in the source distribution; the underlying mathematical problem is closely related to the classical problem of establishing the consistency of maximum likelihood estimators. General consistency results are given for the plug-in estimator applied to a broad class of sources, including all stationary and ergodic ones. A more general class of estimation problems is also considered, arising in the context of lossy data compression when the allowed class of coding distributions is restricted; analogous results are developed for the plug-in estimator in that case. Finally, consistency theorems are formulated for modified (e.g., penalized) versions of the plug-in, and for estimating the optimal reproduction distribution.

Index Terms

Rate-distortion function, entropy, estimation, consistency, maximum likelihood, plug-in estimator

I. Introduction

Suppose a data string $x_1^n := (x_1, x_2, \ldots, x_n)$ is generated by a stationary memoryless source $(X_n\,;\, n \geq 1)$ with unknown marginal distribution P on a discrete alphabet A. In many theoretical and practical problems arising in a wide variety of scientific contexts, it is desirable – and often important – to obtain accurate estimates of the entropy $H(P)$ of the source, based on the observed data x_1^n; see, for example, [35] [26] [30] [29] [32] [31] [8] and the references therein. Perhaps the simplest method is via the so-called **plug-in estimator**, where the entropy of P is estimated by $H(P_{x_1^n})$, namely, the entropy of the empirical distribution $P_{x_1^n}$ of x_1^n. The plug-in estimator satisfies the basic statistical requirement of consistency, that is, $H(P_{X_1^n}) \to H(P)$ in probability as $n \to \infty$. In fact, it is strongly consistent; the convergence holds with probability one [2].

A natural generalization is the problem of estimating the rate-distortion function $R(P, D)$ of a (not necessarily discrete-valued) source. Motivation for this comes in part from lossy data compression, where we may need an estimate of how well a given data set could potentially be compressed, cf. [10], and also from cases where we want to quantify the "information content" of a particular signal, but the data under examination take values in a continuous (or more general) alphabet, cf. [27].

The rate-distortion function estimation question appears to have received little attention in the literature. Here we present some basic results for this problem. First, we consider the simple **plug-in estimator** $R(P_{X_1^n}, D)$, and determine conditions under which it is strongly consistent, that is, it converges to $R(P, D)$ with probability 1, as $n \to \infty$. We call this the **nonparametric estimation problem**, for reasons that will become clear below.

A shorter version of this paper is to appear in *IEEE Transactions on Information Theory*.

MH was supported in part by a National Defense Science and Engineering Graduate Fellowship. IK was supported in part by a Sloan Research Fellowship from the Sloan Foundation, NSF grant #0073378-CCR and USDA-IFAFS grant #00-52100-9615. The material in this paper is preceded by two technical reports [18], [19]. Preliminary results were presented at [21] and [22].

MH is at the Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 (email: mtharris@cmu.edu).

IK is at the Department of Informatics, Athens University of Econ & Business, Patission 76, Athens 10434, Greece (email: yiannis@aueb.gr).

At first glance, consistency may seem to be a mere continuity issue: Since the empirical distribution $P_{X_1^n}$ converges, with probability 1, to the true distribution P as $n \to \infty$, a natural approach to proving that $R(P_{X_1^n}, D)$ also converges to $R(P, D)$ would be to try and establish some sort of continuity property for $R(P, D)$ as a function of P. But, as we shall see, $R(P_{X_1^n}, D)$ turns out to be consistent under rather mild assumptions, which are in fact *too mild* to ensure continuity in any of the usual topologies; see Section III-E for explicit counterexamples. This also explains our choice of the empirical distribution $P_{x_1^n}$ as an estimate for P: If $R(P, D)$ was continuous in P, then any consistent estimator $\hat{P}_n$ of P could be used to make $R(\hat{P}_n, D)$ a consistent estimator for $R(P, D)$. Some of the subtleties in establishing regularity properties of the rate-distortion function $R(P, D)$ as a function of P are illustrated in [11] [1].

Another advantage of a plug-in estimator is that $P_{x_1^n}$ has finite support, regardless of the source alphabet. This makes it possible (when the reproduction alphabet is also finite) to actually compute $R(P_{x_1^n}, D)$ by approximation techniques such as the Blahut-Arimoto algorithm [7] [3] [12]. When the reproduction alphabet is continuous, the Blahut-Arimoto algorithm can still be used after discretizing the reproduction alphabet; the discretization can, in part, be justified by the observation that it can be viewed as an instance of the *parametric estimation problem* described below. Other possibilities for continuous reproduction alphabets are explored in [33] [5].

The consistency problem can be framed in the following more general setting. As has been observed by several authors recently, the rate-distortion function of a memoryless source admits the decomposition,

$$R(P, D) = \inf_{Q} R(P, Q, D), \tag{1}$$

where the infimum is over all probability distributions Q on the reproduction alphabet, and $R(P, Q, D)$ is the rate achieved by memoryless random codebooks with distribution Q used to compress the source data to within distortion D; see, e.g., [36] [15]. Therefore, $R(P, D)$ is the best rate that can be achieved by this family of codebooks. But in the case where we only have a restricted family of compression algorithms available, indexed, say, by a family of probability distributions $\{Q_\theta\,;\, \theta \in \Theta\}$ on the reproduction alphabet, then the best achievable rate is:

$$R^\Theta(P, D) := \inf_{\theta \in \Theta} R(P, Q_\theta, D). \tag{2}$$

We also consider the **parametric estimation problem**, namely, that of establishing the strong consistency of the corresponding **plug-in estimator** $R^\Theta(P_{X_1^n}, D)$ as an estimator for $R^\Theta(P, D)$. It is important to note that, when Θ indexes the set of all probability distributions on the reproduction alphabet, then the parametric and nonparametric problems are identical, and this allows us to treat both problems in a common framework.

Our two main results, Theorems 4 and 5 in the following section, give regularity conditions for both the parametric and nonparametric estimation problems under which the plug-in estimator is strongly consistent. It is shown that consistency holds in great generality for all distortion values D such that $R^\Theta(P, D)$ is continuous from the left. An example illustrating that consistency may actually fail at those points is given in Section III-D. In particular, for the nonparametric estimation problem we obtain the following three simple corollaries, which cover many practical cases.

Corollary 1: If the reproduction alphabet is finite, then for any source distribution P, $R(P_{X_1^n}, D)$ is strongly consistent for $R(P, D)$ at all distortion levels $D \geq 0$ except perhaps at the single value where $R(P, D)$ transitions from being finite to being infinite.

Corollary 2: If the source and reproduction alphabets are both equal to $\mathbb{R}^d$ and the distortion measure is squared-error, then for any source distribution P and any distortion level $D \geq 0$, $R(P_{X_1^n}, D)$ is strongly consistent for $R(P, D)$.

Corollary 3: Assume that the reproduction alphabet is a compact, separable metric space, and that the distortion measure $\rho(x, \cdot)$ is continuous for each $x \in A$. Then (under mild additional measurability assumptions), for any source distribution P, $R(P_{X_1^n}, D)$ is strongly consistent for $R(P, D)$ at all distortion levels $D \geq 0$ except perhaps at the single value where $R(P, D)$ transitions from being finite to being infinite.

Corollaries 1 and 3 are special cases of Corollary 6 in Section II. Corollary 2 is established in Section III, which contains many other explicit examples illustrating the consistency results and cases where consistency may fail. Section V contains the proofs of all the main results in this paper.

We also consider extensions of these results in two directions. In Section IV-A we examine the problem of estimating the optimal reproduction distribution – namely, the distribution that actually achieves the infimum in equations (1) and (2) – from empirical data. Consistency results are given, under conditions identical to those required for the consistency of the plug-in estimator. Finally, in Section IV-B we show that consistency holds for a more general class of estimators, which arise as modifications of the plug-in. These include, in particular, penalized versions of the plug-in, analogous to the standard penalized maximum likelihood estimators used in statistics.

The analysis of the plug-in estimator presents some unexpected technical difficulties. One way to explain the source of these difficulties is by noting that there is a very close analogy, at least on the level of the mathematics, with the problem of maximum likelihood estimation [see also Section IV-B for another instance of this connection]. Beyond the superficial observation that they are both extremization problems over a space of probability distributions, a more accurate, albeit heuristic, illustration can be given as follows: Suppose we have a memoryless source with distribution P on some discrete alphabet, take the reproduction alphabet to be the same as the source alphabet, and look at the extreme case where no distortion is allowed. Then the plug-in estimator of the rate-distortion function (which now is simply the entropy) can be expressed as a trivial minimization over all possible coding distributions, i.e.,

$$H(P_{x_1^n}) = \min_Q [H(P_{x_1^n}) + H(P_{x_1^n} \| Q)] = -\frac{1}{n} \max_Q \left[\log Q^n(x_1^n) \right],$$

where $H(P\|Q)$ denotes the relative entropy, and Q^n is the n-fold product distribution of n independent random variables each distributed according to Q. Therefore, the computation of the plug-in estimate $H(P_{x_1^n})$ is exactly equivalent to the computation of the maximum likelihood estimate (MLE) of P over a class of distributions Q. Alternatively, in Csiszár's terminology, the minimization of the relative entropy above corresponds to the so-called "reversed I-projection" of $P_{x_1^n}$ onto the set of feasible distributions Q, which in this case consists of *all* distributions on the reproduction alphabet; see, e.g., [16] [13]. Formally, this projection is exactly the same as the computation of the MLE of P based on x_1^n.

In the general case of nonzero distortion $D > 0$, the plug-in estimator can similarly be expressed as, $R(P_{x_1^n}, D) = \min_Q R(P_{x_1^n}, Q, D)$, cf. (1) above. This (now highly nontrivial) minimization is mathematically very closely related to the problem of computing an I-projection as before. The tools we employ to analyze this minimization are based on the technique of epigraphical convergence [34] [4] (this is particularly clear in the proof of our main result, the lower bound in Theorem 5), and it is no coincidence that these same tools have also provided one of the most successful approaches to proving the consistency of MLEs. By the same token, this connection also explains why the consistency of the plug-in estimator involves subtleties similar to those cases where MLEs fail to be consistent [28].

In the way of motivation, we also mention that the asymptotic behavior of the plug-in estimator – and the technical intricacies involved in its analysis – also turn out to be important in extending some of Rissanen's celebrated ideas related to the Minimum Description Length (MDL) principle to the context of lossy data compression; this direction will be explored in subsequent work.

Throughout the paper we work with stationary and ergodic sources instead of memoryless sources, though we are still only interested in estimating the first-order rate-distortion function. One reason for this is that the full rate-distortion function can be estimated by looking at the process in sliding blocks of length m and then estimating the "marginal" rate-distortion function of these blocks for large m; see Section III-F. Another reason for allowing dependence in the data comes from simulation: For example, suppose we were interested in estimating the rate-distortion function of a distribution P that we cannot compute explicitly (as is the case for perhaps the majority of models used in image processing), but for which we have a Markov chain Monte Carlo (MCMC) sampling algorithm. The data generated by such an algorithm is not memoryless, yet we care only about the rate-distortion function of the marginal distribution. In Section IV-C we comment further on this issue, and also give consistency results for data produced by sources that may not be stationary.

II. Main Results

We begin with some notation and definitions that will remain in effect throughout the paper.

Suppose the random source $(X_n\,;\,n\geq 1)$ taking values in the source alphabet A is to be compressed in the reproduction alphabet $\hat{A}$, with respect to the single-letter distortion measures (ρ_n) arising from an arbitrary distortion function $\rho : A\times\hat{A}\mapsto[0,\infty)$. We assume that A and $\hat{A}$ are equipped with the σ-algebras $\mathcal{A}$ and $\hat{\mathcal{A}}$, respectively, that $(A,\mathcal{A})$ and $(\hat{A},\hat{\mathcal{A}})$ are Borel spaces, and that ρ is $\sigma(\mathcal{A}\times\hat{\mathcal{A}})$-measurable.[1] Suppose the source is stationary, and let P denote its marginal distribution on A. Then the (first-order) rate-distortion function $R_1(P,D)$ with respect to the distortion measure ρ is defined as,

$$R_1(P,D) := \inf_{(U,V)\sim W\in W(P,D)} I(U;V), \qquad D\geq 0,$$

where the infimum is over all $A\times\hat{A}$-valued random variables (U,V) with joint distribution W belonging to the set

$$W(P,D) := \left\{W : W^A = P,\ E_W[\rho(U,V)]\leq D\right\},$$

and where W^A denotes the marginal distribution of W on A, and similarly for $W^{\hat{A}}$; the infimum is taken to be $+\infty$ when $W(P,D)$ is empty. As usual, the mutual information $I(U;V)$ between two random variables U,V with joint distribution W, is defined as the relative entropy between W and the product of its two marginals, $W^A\times W^{\hat{A}}$. Here and throughout the paper, all familiar information-theoretic quantities are expressed in nats, and $\log$ denotes the natural logarithm. In particular, for any two probability measures μ,ν on the same space, the relative entropy $H(\mu\|\nu)$ is defined as $E_\mu[\log\frac{d\mu}{d\nu}]$ whenever the density $d\mu/d\nu$ exists, and it is taken to be $+\infty$ otherwise.

We write $D_{\rm c}(P)$ for the set of distortion values $D\geq 0$ for which $R_1(P,D)$ is continuous from the left, i.e.,

$$D_{\rm c}(P) := \{D\geq 0 : R_1(P,D) = \lim_{\lambda\uparrow 1} R_1(P,\lambda D)\}.$$

By convention, this set always includes 0 and any value of D for which $R_1(P,D)=\infty$. But since $R_1(P,D)$ is nonincreasing and convex in D [9] [11], $D_{\rm c}(P)$ actually includes *all* $D\geq 0$ with the only possible exception of the single value of D where $R_1(P,D)$ transitions from being finite to being infinite. Conditions guaranteeing that $D_{\rm c}(P)$ is indeed all of $[0,\infty)$ can be found in [11].

A. Estimation Problems and Plug-in Estimators

Given a finite-length data string $x_1^n := (x_1,x_2,\ldots,x_n)$ produced by a stationary source (X_n) as above with marginal distribution P, the **plug-in estimator** of the first-order rate-distortion function $R_1(P,D)$ is $R_1(P_{x_1^n},D)$, where $P_{x_1^n}$ is the *empirical distribution* induced by the sample x_1^n on A^n, namely,

$$P_{x_1^n}(C) := \frac{1}{n}\sum_{k=1}^n \mathbb{1}\{x_k\in C\} \qquad x_1^n\in A^n,\ C\in\mathcal{A}$$

and where $\mathbb{1}$ is the indicator function. Our first goal is to obtain conditions under which this estimator is strongly consistent. We call this the **nonparametric estimation problem**.

[1]Borel spaces include the Euclidean spaces $\mathbb{R}^d$ as well as all Polish spaces, and they allow us to avoid certain measure-theoretic pathologies while working with random sequences and conditional distributions [25]. Henceforth, all σ-algebras and the various product σ-algebras derived from them are understood from the context. We do not complete any of the σ-algebras, but we say that an event C holds *with probability 1* (w.p.1) if C contains a measurable subset C' that has probability 1.

We also consider the more general class of estimation problems mentioned in the Introduction. Suppose for a moment that our goal is to compress data produced by a *memoryless* source (X_n) with distribution P on A, and suppose also that we are restricted to using memoryless random codebooks with distributions Q belonging to some parametric family $\{Q_\theta \ : \ \theta \in \Theta\}$ where Θ indexes a subset of all probability distributions on $\hat{A}$. Using a random codebook with distribution Q to compress the data to within distortion D, yields (asymptotically) a rate of $R_1(P,Q,D)$ nats/symbol, where the rate-function $R_1(P,Q,D)$ is given by,

$$R_1(P,Q,D) = \inf_{W \in W(P,D)} H(W \| P \times Q).$$

See [36] [15] for details. From this it is immediate that the rate-distortion function of the source admits the decomposition given in (1). Having restricted attention to the class of codebook distributions $\{Q_\theta \ ; \ \theta \in \Theta\}$, then the best possible compression rate is:

$$R_1^\Theta(P,D) := \inf_{\theta \in \Theta} R_1(P,Q_\theta,D) \ \text{ nats/symbol.} \tag{3}$$

When θ indexes certain nice families, say Gaussian, the infimum $R_1^\Theta(P,D)$ can be analytically derived or easily computed, often for any distribution P, including an empirical distribution.

Thus motivated, we now formally define the **parametric estimation problem**. Suppose (X_n) is a stationary source as above, and let $\{Q_\theta \ : \ \theta \in \Theta\}$ be a family of probability distributions on the reproduction alphabet $\hat{A}$ parameterized by an arbitrary parameter space Θ. The **plug-in estimator** for $R_1^\Theta(P,D)$ is $R_1^\Theta(P_{X_1^n},D)$, and we seek conditions for its strong consistency.

Note that $R_1^\Theta(P,D) = R_1(P,D)$ when $\{Q_\theta \ : \ \theta \in \Theta\}$ includes all probability distributions on $\hat{A}$, or if it simply includes the optimal reproduction distribution achieving the infimum in (1). Otherwise, $R_1^\Theta(P,D)$ may be strictly larger than $R_1(P,D)$. Therefore, the nonparametric problem is a special case of the parametric one, and we can consider the two situations in a common framework.

In the parametric scenario we write,

$$D_{\text{c}}^\Theta(P) := \left\{ D \geq 0 : R_1^\Theta(P,D) = \lim_{\lambda \uparrow 1} R_1^\Theta(P,\lambda D) \right\}.$$

Unlike $D_{\text{c}}(P)$, $D_{\text{c}}^\Theta(P)$ can exclude more than a single point.

B. Consistency

We investigate conditions under which the plug-in estimator $R_1^\Theta(P_{x_1^n},D)$ is strongly consistent, i.e.,[2]

$$R_1^\Theta(P_{X_1^n},D) \stackrel{\text{w.p.1}}{\to} R_1^\Theta(P,D). \tag{4}$$

Of course in the special case where Θ indexes all probability distributions on $\hat{A}$, this reduces to the nonparametric problem, and (4) becomes $R_1(P_{X_1^n},D) \stackrel{\text{w.p.1}}{\to} R_1(P,D)$. We separately treat the upper and lower bounds that combine to give (4).

The upper bound does not require *any* further regularity assumptions, although there can be certain pathological values of D for which it is not valid. In the nonparametric situation, the only potential problem point is the single value of D where $R_1(P,D)$ transitions from finite to infinite.

Theorem 4: If the source (X_n) is stationary and ergodic with $X_1 \sim P$, then

$$\limsup_{n \to \infty} R_1^\Theta(P_{X_1^n},D) \stackrel{\text{w.p.1}}{\leq} R_1^\Theta(P,D)$$

for all $D \in D_{\text{c}}^\Theta(P)$.

[2]Throughout the paper we do not require limits to be finite valued, but say that $\lim_n a_n = \infty$ if a_n diverges to ∞ (and similarly for $-\infty$).

As illustrated by a simple counterexample in Section III-D, the requirement that $D \in D_{\rm c}^{\Theta}(P)$ cannot be relaxed completely. The proof of the theorem, given in Section V, is a combination of the decomposition in (3) and the fact that $R_1(P_{X_1^n}, Q, D) \stackrel{\text{w.p.1}}{\to} R_1(P, Q, D)$ quite generally. Actually, from the proof we also obtain an upper bound on the $\liminf$,

$$\liminf_{n\to\infty} R_1^{\Theta}(P_{X_1^n}, D) \stackrel{\text{w.p.1}}{\leq} R_1^{\Theta}(P, D) \quad \text{for all } D \geq 0, \tag{5}$$

which provides some information even for those values of D where the upper bound in Theorem 4 may fail.

For the corresponding lower bound in (4), some mild additional assumptions are needed. We will always assume that Θ is a metric space, and also that the following two conditions are satisfied:

A1. The map $\theta \mapsto E_\theta[e^{\lambda\rho(x,Y)}]$ is continuous for each $x \in A$ and $\lambda \leq 0$, where E_θ denotes expectation w.r.t. Q_θ.

A2. For each $D \geq 0$, there exists a (possibly random) sequence (θ_n) with

$$\liminf_{n\to\infty} R_1(P_{X_1^n}, Q_{\theta_n}, D) \stackrel{\text{w.p.1}}{\leq} \liminf_{n\to\infty} R_1^{\Theta}(P_{X_1^n}, D), \tag{6}$$

and such that (θ_n) is relatively compact with probability 1.

Theorem 5: If Θ is separable, A1 and A2 hold, and (X_n) is stationary and ergodic with $X_1 \sim P$, then

$$\liminf_{n\to\infty} R_1^{\Theta}(P_{X_1^n}, D) \stackrel{\text{w.p.1}}{\geq} R_1^{\Theta}(P, D)$$

for all $D \geq 0$.

Although A1 and A2 may seem quite involved, they are fairly easy to verify in specific examples: For A1, we have the following sufficient conditions; as we prove in Section V, either one implies A1.

P1. Whenever $\theta_n \to \theta$, we also have that $Q_{\theta_n} \to Q_\theta$ setwise.[3]

N1. $(\hat{A}, \hat{\mathcal{A}})$ is a metric space with its Borel σ-algebra, $\rho(x,\cdot)$ is continuous for each $x \in A$ and $\theta_n \to \theta$ implies that $Q_{\theta_n} \to Q_\theta$ weakly.[4]

For A2, we first note that a sequence (θ_n) satisfying (6) *always* exists and that the inequality in (6) must always be an equality. The important requirement in A2 is that (θ_n) be relatively compact. In particular, A2 is trivially true if Θ is compact. More generally, the following two conditions make it easier to verify A2 in particular examples. In Section V we prove that either one implies A2 as long as the source is stationary and ergodic with marginal distribution P. For any subset K of the source alphabet A, we write $B(K,M)$ for the subset of $\hat{A}$ which is the union of all the distortion balls of radius $M \geq 0$ centered at points of K. Formally,

$$B(K,M) := \bigcup_{x\in K} \{y : \rho(x,y) \leq M\}, \quad K \subseteq A,\ M \geq 0.$$

P2. For each $D \geq 0$, there exists a $\Delta > 0$ and a $K \in \mathcal{A}$ such that $P(K) > D/(D+\Delta)$ and $\{\theta : Q_\theta(B(K, D+\Delta)) \geq \epsilon\}$ is relatively compact for each $\epsilon > 0$.

N2. $(\hat{A}, \hat{\mathcal{A}})$ is a metric space with its Borel σ-algebra, Θ is the set of all probability distributions on $\hat{A}$ with a metric that metrizes weak convergence of probability measures, and for each $\epsilon > 0$ and each $M > 0$ there exists a $K \in \mathcal{A}$ such that $P(K) > 1 - \epsilon$ and $B(K,M)$ is relatively compact.[5]

In Section III we describe concrete situations where these assumptions are valid.

[3]We say that $Q_m \to Q$ setwise if $E_{Q_m}(f) \to E_Q(f)$ for all bounded, measurable functions f, or equivalently, if $Q_m(C) \to Q(C)$ for all measurable sets C.

[4]We say that $Q_m \to Q$ weakly if $E_{Q_m}(f) \to E_Q(f)$ for all bounded, continuous functions f, or equivalently, if $Q_m(C) \to Q(C)$ for all measurable sets C with $Q(\partial C) = 0$.

[5]Θ can always be metrized in this way, and so that Θ will be separable (compact) if $\hat{A}$ is separable (compact) [6].

The proof of Theorem 5 has the following main ingredients. The separability of Θ and the continuity in A1 are used to ensure measurability and, in particular, for controlling exceptional sets. A1 is a local assumption that ensures $\inf_{\theta\in U} R_1(P_{X_1^n}, Q_\theta, D)$ is well behaved in small neighborhoods U. A2 is a global assumption that ensures the final analysis can be restricted to a small neighborhood.

Combining Theorems 4 and 5 gives conditions under which $R_1^\Theta(P_{X_1^n}, D) \overset{\text{w.p.1}}{\to} R_1^\Theta(P, D)$. In the non-parametric situation we have the following Corollary, which is a generalization of Corollary 3 in the Introduction; it follows immediately from the last two theorems.

Corollary 6: Suppose $(\hat{A}, \hat{\mathcal{A}})$ is a compact, separable metric space with its Borel σ-algebra and $\rho(x, \cdot)$ is continuous for each $x \in A$. If (X_n) is stationary and ergodic with $X_1 \sim P$, then $R_1(P_{X_1^n}, D) \overset{\text{w.p.1}}{\to} R_1(P, D)$ for all $D \in D_{\text{c}}(P)$. Furthermore, the compactness condition can be relaxed as in N2.

III. Examples

In all of the examples we assume that the source (X_n) is stationary and ergodic with $X_1 \sim P$.

A. Nonparametric Consistency: Discrete Alphabets

Let A and $\hat{A}$ be at most countable and let ρ be unbounded in the sense that for each fixed $x \in A$ and each fixed $M > 0$ there are only finitely many $y \in \hat{A}$ with $\rho(x, y) < M$. N1 and N2 are clearly satisfied in the nonparametric setting where Θ is the set of all probability distributions on $\hat{A}$, so $R_1(P_{X_1^n}, D) \overset{\text{w.p.1}}{\to} R_1(P, D)$ for all D except perhaps at the single value of D where $R_1(P, D)$ transitions from finite to infinite. If, in addition, for each x there exists a y with $\rho(x, y) = 0$, then $D_{\text{c}}(P) = [0, \infty)$ regardless of P [11], and the plug-in estimator is strongly consistent for all P and all D.

This example also yields a different proof of the general consistency result mentioned in the Introduction, for the plug-in estimate of the entropy of a discrete-valued source: If we map $A = \hat{A}$ into the integers, let $\rho(x, y) = |x - y|$, and take $D = 0$, then we obtain the strong consistency of [2, Cor. 1].

B. Nonparametric Consistency: Continuous Alphabets

Again in the nonparametric setting, let $A = \hat{A} = \mathbb{R}^d$ be finite dimensional Euclidean space, and let $\rho(x, y) := f(\|x - y\|)$ for some function f of Euclidean distance where $f : [0, \infty) \to [0, \infty)$ is continuous and $f(t) \to \infty$ as $t \to \infty$. As in the previous example, N1 and N2 are clearly satisfied, so $R_1(P_{X_1^n}, D) \overset{\text{w.p.1}}{\to} R_1(P, D)$ for all D except perhaps at the single value of D where $R_1(P, D)$ transitions from finite to infinite. If furthermore $f(0) = 0$, then $D_{\text{c}}(P) = [0, \infty)$ regardless of P [11] and the plug-in estimator is strongly consistent for all P and all D.

This example includes the important special case of squared-error distortion: In the nonparametric problem, the plug-in estimator is always strongly consistent under squared-error distortion over finite dimensional Euclidean space, as stated in Corollary 2. This example also generalizes as follows. The alphabets A and $\hat{A}$ can be (perhaps different) subsets of $\mathbb{R}^d$, as long as $\hat{A}$ is closed. The use of Euclidean distance is not essential and we can take any $\rho \geq f$, so that ρ is not required to be translation invariant, as long as ρ is continuous over $\hat{A}$ for each fixed $x \in A$. This is enough for consistency except perhaps at a single value of D. To use the results in [11] to rule out any pathological values of D, that is, to show that $D_{\text{c}} = [0, \infty)$ we also need A to be closed, ρ to be continuous over A for each fixed y and $\inf_y \rho(x, y) = 0$ for each x.

C. Parametric Consistency for Gaussian Families

Let $A = \hat{A} = \mathbb{R}$, let ρ satisfy the assumptions of Example III-B, let $\Theta = \{(\mu, \sigma) \in \mathbb{R} \times [0, \infty)\}$ with the Euclidean metric, and for each $\theta = (\mu, \sigma)$ let Q_θ be Gaussian with mean μ and standard deviation σ [the case $\sigma = 0$ corresponds to the point mass at μ]. Conditions N1 and P2 are clearly satisfied, so $R_1^\Theta(P_{X_1^n}, D) \overset{\text{w.p.1}}{\to} R_1^\Theta(P, D)$ for all $D \in D_{\rm c}^\Theta(P)$. In the special case where $\rho(x,y) = (x-y)^2$ is squared-error distortion, then it is not too difficult [15] to show that

$$R_1^\Theta(P, D) = \max\left\{0, \frac{1}{2}\log\frac{\sigma_X^2}{D}\right\},$$

where σ_X^2 denotes the (possibly infinite) variance of P, so $D_{\rm c}^\Theta(P) = [0, \infty)$ and the convergence holds for all D. Furthermore, if the source P happens to also be Gaussian, then $R_1^\Theta(P, D) = R_1(P, D)$ and the plug-in estimator is also strongly consistent for the nonparametric problem.

D. Convergence Failure for $D \notin D_c(P)$

Let $A = \{0, 1\}$, $\hat{A} = \{0\}$, and $\rho(x, y) := |x - y|$. Since there is only one possible distribution on $\hat{A}$, it is easy to show that

$$R_1(P', D) = \begin{cases} 0 & \text{if } P'(1) \le D \\ \infty & \text{otherwise} \end{cases}$$

for any distribution P' on A. If $P(1) > 0$, the only possible trouble point for consistency is $D = P(1)$, which is not in $D_{\rm c}(P)$. It is easy to see that convergence (and therefore consistency) might fail at this point because $R_1(P_{X_1^n}, D)$ will jump back and forth between 0 and ∞ as $P_{X_1^n}(1)$ jumps above and below $D = P(1)$. The law of the iterated logarithm implies that this failure to converge happens with probability 1 when the source is memoryless. In general, when the source is stationary and ergodic, it turns out that convergence will fail with positive probability [24] [23] [20].

E. Consistency at a Point of Discontinuity in P

This slightly modified example from Csiszár [11] illustrates that $R_1(\cdot, D)$ can be discontinuous at P even though the plug-in estimator is consistent. Let $A = \hat{A} = \{1, 2, \dots\}$, let P' be any distribution on A with infinite entropy and with $P'(x) > 0$ for all x, and let $\rho(x, y) := P'(x)^{-1}\mathbb{1}\{x \ne y\} + |x - y|$. Note that $R_1(P', D) = \infty$ for all D.[6] This is a special case of Example III-A so the plug-in estimator is always strongly consistent regardless of P and D. Nevertheless, $R_1(\cdot, D)$ is discontinuous everywhere it is finite.

To see this, let the source P be any distribution on A with finite entropy $H(P)$. Note that $R_1(P, D) \le R_1(P, 0) = H(P) < \infty$. Define the mixture distribution $P_\epsilon := (1-\epsilon)P + \epsilon P'$. Then $P_\epsilon \to P$ in the topology of total variation[7] (and also any weaker topology) as $\epsilon \downarrow 0$, but $R_1(P_\epsilon, D) \not\to R_1(P, D)$ because $R_1(P_\epsilon, D) \ge \epsilon R_1(P', D/\epsilon) = \infty$ for all $\epsilon > 0$. See (13) below for a proof of this last inequality.[8]

The key property of ρ in this example is that there exists a P' with $R_1(P', D) = \infty$ for *all* D. If such a P' exists, then $R_1(\cdot, D)$ will be discontinuous in the topology of total variation at any point P where $R_1(P, D)$ is finite for exactly the same reason as above. Although this specific example is based on a rather pathological distortion measure, many unbounded distortion measures on continuous alphabets,

[6] $R_1(P', \cdot) \equiv \infty$, because any pair of random variables (U, V) with $U \sim P'$ and $E[\rho(U,V)] < \infty$ has $I(U;V) = \infty$. To see this, first note that $E[\rho(U,V)] < \infty$ implies that $\alpha(x) := \text{Prob}\{V = x | U = x\} \to 1$ as $x \to \infty$; simply use the definition of ρ and ignore the $|x - y|$ term. Computing the mutual information and using the log-sum inequality gives $I(U;V) \ge \kappa + \sum_x P'(x)\alpha(x)\log\bigl(\alpha(x)/Q(x)\bigr)$, where $V \sim Q$ and where κ is a finite constant that comes from all of the other terms in the definition of $I(U;V)$ combined together with the log-sum inequality. Since $\alpha(x) \to 1$ and since $\sum_x P'(x)\log\bigl(1/Q(x)\bigr) \ge H(P') = \infty$ for any probability distribution Q, we see that $I(U;V) = \infty$. We can ignore $\alpha(x)$ because the finiteness of the sum only depends on the behavior for large x, and for large enough x we have $\alpha(x) > 1/2$, say.

[7] The topology of total variation is metrized by the distance $d(P, P') := \sup_C |P(C) - P'(C)|$.

[8] An interesting special case of this example (based on the fact that $\sum_{x\ge 2}[x\log^\alpha x]^{-1}$ converges if and only if $\alpha > 1$) is $P'(x-1) \propto 1/(x\log^{1.5} x)$ (infinite entropy) and $P(x-1) \propto 1/(x\log^{2.5} x)$ (finite entropy), $x = 2, 3, \dots$, because the relative entropies $H(P'\|P)$ and $H(P\|P')$ are both finite, so $H(P_\epsilon\|P) \to 0$ and $H(P\|P_\epsilon) \to 0$ as $\epsilon \downarrow 0$. (From the convexity of relative entropy.) This counterexample thus shows that even closeness in relative entropy between two distributions (which is stronger than closeness in total variation) is not enough to guarantee the closeness of the rate-distortion functions of the corresponding distributions.

including squared-error distortion on $\mathbb{R}$, have such a P' and are thus discontinuous in the topology of total variation.[9]

F. Higher-Order Rate-Distortion Functions

Suppose that we want to estimate the mth-order rate-distortion function of a stationary and ergodic source (X_n) with mth order marginal distribution $X_1^m \sim P_m$, namely,

$$R_m(P_m, D) := \frac{1}{m} \inf_{(U,V)\sim W \in W_m(P_m,D)} I(U;V),$$

where the infimum is over all $A^m \times \hat{A}^m$-valued random variables, with joint distribution W in the set $W_m(P_m, D)$ of probability distributions on $A^m \times \hat{A}^m$ whose marginal distribution on A^m equals P_m, and which have $E[\rho_m(U,V)] \le D$ for

$$\rho_m(x_1^n, y_1^n) := \frac{1}{m}\sum_{k=1}^{m} \rho(x_k, y_k) \qquad x_1^m \in A^m,\ y_1^m \in \hat{A}^m.$$

All our results above immediately apply to this situation. We simply estimate the first-order rate-distortion function of the sliding-block process (Z_n) defined by $Z_k := (X_k, \dots, X_{k+m-1})$ with source alphabet A^m, reproduction alphabet $\hat{A}^m$ and distortion measure ρ_m, and then divide the estimate by m.

IV. Further Results

A. Estimation of the Optimal Reproduction Distribution

So far, we concentrated on conditions under which the plug-in estimator is consistent; these guarantee an (asymptotically) accurate estimate of the best compression rate $R_1^\Theta(P,D) = \inf_{\theta\in\Theta} R_1(P, Q_\theta, D)$ that can be achieved by codes restricted to some class of distributions $\{Q_\theta\ ;\ \theta \in \Theta\}$. Now suppose this infimum is achieved by some θ^*, corresponding to the optimal reproduction distribution Q_{θ^*}. Here we use a simple modification of the plug-in estimator in order to obtain estimates $\theta_n = \theta_n(x_1^n)$ for the optimal reproduction parameter θ^* based on the data x_1^n. Specifically, since we have conditions under which

$$\inf_{\theta\in\Theta} R_1(P_{x_1^n}, Q_\theta, D) \approx \inf_{\theta\in\Theta} R_1(P, Q_\theta, D), \tag{7}$$

we naturally consider the sequence of estimators which achieve the infima on the left-hand-side of (7) for each $n \ge 1$; that is, we simply replace the $\inf$ by an $\arg\inf$. Since these arg-infima may not exist or may not be unique, we actually consider any sequence of **approximate minimizers** (θ_n) that have $R_1(P_{X_1^n}, Q_{\theta_n}, D) \approx R_1^\Theta(P_{X_1^n}, D)$ in the sense that (9) below holds. Similarly, minimizers θ^* of the right-hand-side of (7) may not exist or be unique, either. We thus consider the (possibly empty) set Θ^* containing all the minimizers of $R_1(P, Q_\theta, D)$ and address the problem of whether the estimators θ_n converge to Θ^*, meaning that θ_n is eventually in any neighborhood of Θ^*.

Our proofs are in part based on a recent result from [24] [23].

Theorem 7: [24] [23] If the source (X_n) is stationary and ergodic with $X_1 \sim P$, then

$$\liminf_{n\to\infty} R_1(P_{X_1^n}, Q, D) \overset{\text{w.p.1}}{=} R_1(P, Q, D)$$

for all $D \ge 0$ and

$$\lim_{n\to\infty} R_1(P_{X_1^n}, Q, D) \overset{\text{w.p.1}}{=} R_1(P, Q, D) \tag{8}$$

[9] For squared-error distortion, let P' be any distribution over discrete points $\{x_1, x_2, \dots\} \subset \mathbb{R}$ where $x_k \ge x_{k-1} + 2^{1/P'(x_k)}$ and where $H(P') = \infty$. This is essentially the same as Csiszár's example above because any pair of random variables (U,V) with $E[\rho(U,V)] < \infty$ must have $\text{Prob}\{V$ closer to x_k than any other $x_j | U = x_k\} \to 1$ as $k \to \infty$.

for all D in the set

$$D_{\mathrm{c}}(P,Q) := \left\{ D \geq 0 : R_1(P,Q,D) = \lim_{\lambda \uparrow 1} R_1(P,Q,\lambda D) \right\}.$$

Similar to $D_{\mathrm{c}}(P)$, $D_{\mathrm{c}}(P,Q)$ always contains 0 and any point where $R_1(P,Q,D) = \infty$. Since the function $R_1(P,Q,D)$ is convex and nonincreasing in D [24] [23], $D_{\mathrm{c}}(P,Q)$ is the entire interval $[0,\infty)$, except perhaps the single point where $R_1(P,Q,D)$ transitions from finite to infinite.

Somewhat loosely speaking, the main point of this paper is to give conditions under which an infimum over Q can be moved inside the limit in the above theorem. It turns out that our method of proof works equally well for moving an arg-infimum inside the limit. The next theorem, proved in Section V, is a strong consistency result giving conditions under which the approximate minimizers (θ_n) converge to the optimal parameters $\{\theta^*\}$ corresponding to the optimal reproduction distributions $\{Q_{\theta^*}\}$.

Theorem 8: Suppose the source (X_n) is stationary and ergodic with $X_1 \sim P$, the parameter set Θ is separable, and A1 and A2 hold. Then for all $D \in D_{\mathrm{c}}^{\Theta}(P)$, the set

$$\Theta^* := \arg\inf_{\theta \in \Theta} R_1^{\Theta}(P, Q_\theta, D)$$

is not empty and any (typically random) sequence (θ_n) of approximate minimizers, i.e., satisfying,

$$\limsup_{n\to\infty} R_1(P_{X_1^n}, Q_{\theta_n}, D) \leq \limsup_{n\to\infty} R_1^{\Theta}(P_{X_1^n}, D), \tag{9}$$

has all of its limit points in Θ^* with probability 1. Furthermore, if $R_1^{\Theta}(P,D) < \infty$ and either P2 or N2 holds, then any sequence of approximate minimizers (θ_n) is relatively compact with probability 1. Hence, $\theta_n \to \Theta^*$ with probability 1.

B. More General Estimators

The upper and lower bounds of Theorems 4 and 5 can be combined to extend our results to a variety of estimators besides the ones considered already. For example, instead of the simple plug-in estimator,

$$R_1^{\Theta}(P_{x_1^n}, D) = \inf_{\theta\in\Theta} R_1(P_{x_1^n}, Q_\theta, D)$$

we may wish to consider MDL-style penalized estimators, of the form,

$$\inf_{\theta\in\Theta} \left\{ R_1(P_{x_1^n}, Q_\theta, D) + F_n(\theta) \right\}, \tag{10}$$

for appropriate (nonnegative) penalty functions $F_n(\theta)$. The penalty functions express our preference for certain (typically less complex) subsets of Θ over others. This issue is, of course, particularly important when estimating the optimal reproduction distribution as discussed in the previous section. Note that in the case when no distortion is allowed, these estimators reduce to the classical ones used in lossless data compression and in MDL-based model selection [13]. Indeed, if $A = \hat{A}$ are discrete sets, ρ is Hamming distance and $D = 0$, then the estimator in (10) becomes,

$$-\frac{1}{n} \sup_{\theta\in\Theta} \left\{ \log Q_\theta^n(x_1^n) - nF_n(\theta) \right\},$$

which is precisely the general form of a penalized maximum likelihood estimator. [As usual, Q^n denotes the n-fold product distribution on $\hat{A}^n$ corresponding to the marginal distribution Q.]

More generally, suppose we have a sequence of functions $(\varphi_n(x_1^n, \theta, D))$ with the properties that,

$$\varphi_n(x_1^n, \theta, D) \geq R_1(P_{x_1^n}, Q_\theta, D) \tag{11a}$$

$$\limsup_{n\to\infty} \varphi_n(X_1^n, \theta, D) \overset{\text{w.p.1}}{=} \limsup_{n\to\infty} R_1(P_{X_1^n}, Q_\theta, D) \tag{11b}$$

for all n, x_1^n, θ and D. For each such sequence of functions (φ_n), we define a new estimator for $R_1^\Theta(P,D)$ by,

$$\varphi_n^\Theta(x_1^n, D) := \inf_{\theta\in\Theta} \varphi_n(x_1^n, \theta, D).$$

Condition (11a) implies that any lower bound for the plug-in estimator also holds here. Also, by considering a single θ' for which

$$\limsup_n R_1(P_{X_1^n}, Q_{\theta'}, D) \overset{\text{w.p.1}}{\leq} R_1^\Theta(P,D) + \epsilon,$$

we see that (11b) similarly implies a corresponding upper bound. We thus obtain:

Corollary 9: Theorems 4, 5 and 8 remain valid if $R_1^\Theta(P_{X_1^n}, D)$ is replaced by $\varphi_n^\Theta(X_1^n, D)$ for any sequence of functions (φ_n) satisfying (11a) and (11b).

For example, the penalized plug-in estimators above satisfy the conditions of the corollary, as long as the penalty functions F_n satisfy, for each θ, $F_n(\theta) \to 0$ as $n \to \infty$.

Another example is the sequence of estimators based on the "lossy likelihoods" of [21], namely,

$$\varphi_n(x_1^n, \theta, D) = -\frac{1}{n} \log Q_\theta^n \left(B_n(x_1^n, D)\right)$$

where $B_n(x_1^n, D)$ denotes the distortion-ball of radius D centered at x_1^n,

$$B_n(x_1^n, D) := \left\{ y_1^n \in \hat{A}^n : \frac{1}{n}\sum_{k=1}^n \rho(x_k, y_k) \leq D \right\},$$

cf. [14]. Again, both conditions (11a) and (11b) are valid in this case [24] [23].

C. Nonstationary Sources

As mentioned in the introduction, part of our motivation comes from considering the problem of estimating the rate-distortion function of distributions P which cannot be computed analytically, but which can be easily simulated by MCMC algorithms, as is very often the case in image processing, for example. Of course, MCMC samples are typically not stationary. However, the distribution of the entire sequence of MCMC samples is dominated by (i.e., is absolutely continuous with respect to) a stationary and ergodic distribution, namely, the distribution of the same Markov chain started from its stationary distribution, which is of course the target distribution P. Therefore, all of our results remain valid: Results that hold with probability 1 in the stationary case necessarily hold with probability 1 in the nonstationary case. The only minor technicality is that the initial distribution of the MCMC chain needs to be absolutely continuous with respect to P.

More generally (for non-Markov sources), the requirements of stationarity and ergodicity are more restrictive than necessary. An inspection of the proofs (both here and in the proof of Theorem 7 in [24] [23]), reveals that we only need the source to have the following law-of-large-numbers property:

LLN. There exists a random variable X taking values in the source alphabet A, such that,

$$\tfrac{1}{n}\textstyle\sum_{k=1}^n f(X_k) \overset{\text{w.p.1}}{\to} E[f(X)],$$

for every nonnegative measurable function f.

Theorem 10: Theorems 4, 5 and 8, Corollary 9 and the alternative conditions for A2 remain valid if, instead of being stationary and ergodic with $X_1 \sim P$, the source merely satisfies the LLN property for some random variable $X \sim P$. If the distortion measure ρ is bounded, then the LLN property need only hold for bounded, measurable f.

Every stationary and ergodic source satisfies this LLN property as does any source whose distribution is dominated by the distribution of a stationary and ergodic source. This LLN property is somewhat different from the requirement that the source be asymptotically mean stationary (a.m.s.) with an ergodic mean stationary distribution [17]. The latter is a stronger assumption in the sense that f can depend on the entire future of the process, i.e., $n^{-1}\sum_{k=1}^{n} f(X_k, X_{k+1}, \ldots) \stackrel{\text{w.p.1}}{\to} E[f(X^\infty)]$, where X^∞ is now a random variable on the infinite sequence space. It is a weaker assumption in that this convergence need only hold for bounded f. The final statement of Theorem 10 implies that our consistency results hold for a.m.s. sources (with ergodic mean stationary distributions) as long as the distortion measure is bounded.

V. Proofs

We frequently use the alternative representation [24] [23]

$$R_1(P,Q,D) = \sup_{\lambda \le 0}\Big[\lambda D - E_{X\sim P}\big[\log E_{Y\sim Q} e^{\lambda\rho(X,Y)}\big]\Big] \tag{12}$$

which is valid for all choices of P, Q and D.

This representation makes it easy to prove that

$$R_1^\Theta(\epsilon P' + (1-\epsilon)P, D) \ge \epsilon R_1^\Theta(P', D/\epsilon) \tag{13}$$

for $\epsilon \in (0,1)$, which is used above in Example III-E. Indeed,

$$\begin{aligned}
&R_1(\epsilon P' + (1-\epsilon)P, Q_\theta, D)\\
&\quad= \sup_{\lambda\le 0}\Big[\lambda D - \epsilon E_{X\sim P'}\big[\log E_{Y\sim Q_\theta} e^{\lambda\rho(X,Y)}\big] - (1-\epsilon)E_{X\sim P}\big[\log E_{Y\sim Q_\theta} e^{\lambda\rho(X,Y)}\big]\Big]\\
&\quad\ge \sup_{\lambda\le 0}\Big[\lambda D - \epsilon E_{X\sim P'}\big[\log E_{Y\sim Q_\theta}[e^{\lambda\rho(X,Y)}]\big]\Big]\\
&\quad= \epsilon \sup_{\lambda\le 0}\Big[\lambda D/\epsilon - E_{X\sim P'}\big[\log E_{Y\sim Q_\theta}[e^{\lambda\rho(X,Y)}]\big]\Big]\\
&\quad= \epsilon R_1(P', Q_\theta, D/\epsilon).
\end{aligned}$$

Taking the infimum over $\theta \in \Theta$ on both sides gives (13).

A. Measurability

Here we discuss the various measurability assumptions that are used throughout the paper. Note that we do not always establish the measurability of an event if it contains another measurable event that has probability 1.

Since ρ is product measurable, $x \mapsto E_\theta[e^{\lambda\rho(x,Y)}]$ is measurable. This implies that $x_1^n \mapsto \lambda D - E_{P_{x_1^n}}\Big\{\log E_\theta[e^{\lambda\rho(X,Y)}]\Big\}$ is measurable. Since this is concave in λ [23], we can evaluate the supremum over all $\lambda \le 0$ in (12) by considering only countably many $\lambda \le 0$, which means that $x_1^n \mapsto R_1(P_{x_1^n}, Q_\theta, D)$ is measurable.

If Θ is a separable metric space and $f : \Theta \times A^n \to \bar{\mathbb{R}}$ is measurable for fixed $\theta \in \Theta$ and continuous for fixed $x_1^n \in A^n$, then $x_1^n \mapsto \sup_{\theta\in U} f(\theta, x_1^n)$ is measurable for any subset $U \subseteq \Theta$. This is because $\sup_{\theta\in U} f = \sup_{\theta \in U'} f$ for any (at most) countable dense subset $U' \subseteq U$, and the latter is measurable because U' is (at most) countable. Since Θ is separable, such a U' always exists, and since $f(\cdot, x_1^n)$ is continuous, U' can be chosen independently of x_1^n. An identical argument holds for $\inf_{\theta\in U} f$. We make use of this frequently in the lower bound, where the necessary continuity comes from A1.

B. Proof of Theorem 4

The upper bound in Theorem 4 is deduced from Theorem 7 as follows. If $D = 0$ or $R_1^\Theta(P, D) = \infty$, then choose $D' = D$, otherwise, choose $D' < D$ such that $R_1^\Theta(P, D') \le R_1^\Theta(P, D) + \epsilon/2$. We can always do this since $D \in D_c^\Theta(P)$. Now pick $\theta \in \Theta$ with $R_1(P, Q_\theta, D') \le R_1^\Theta(P, D') + \epsilon/2$. This ensures that $D \in D_c(P, Q_\theta)$ and Theorem 7 gives

$$\limsup_{n\to\infty} R_1^\Theta(P_{X_1^n}, D) \le \limsup_{n\to\infty} R_1(P_{X_1^n}, Q_\theta, D) \overset{\text{w.p.1}}{=} R_1(P, Q_\theta, D) \le R_1(P, Q_\theta, D') \le R_1^\Theta(P, D) + \epsilon$$

completing the proof. Notice that if we switch the $\limsup$ to a $\liminf$, we can remove any restrictions on D since there are no restrictions in this case in Theorem 7. This gives (5).

C. Proof of Theorem 5

Here we prove the lower bound of Theorem 5. Let τ denote the metric on Θ and let $O(\theta, \epsilon) := \{\theta' : \tau(\theta', \theta) < \epsilon\}$ denote the open ball of radius ϵ centered at θ. The main goal is to prove that

$$\lim_{\epsilon\downarrow 0} \liminf_{n\to\infty} \inf_{\theta'\in O(\theta,\epsilon)} R_1(P_{X_1^n}, Q_{\theta'}, D) \overset{\text{w.p.1}}{\ge} R_1(P, Q_\theta, D) \tag{14}$$

for all $\theta \in \Theta$ simultaneously, that is, the exceptional set can be chosen independently of θ. To see how this gives the lower bound, first choose a sequence (θ_n) according to A2 and a subsequence (n_k) along which the $\liminf$ on the left side of (6) is actually a limit. Let θ^* be a limit point of the subsequence (θ_{n_k}). Note that such a θ^* exists with probability 1 by assumption A2 and that it depends on X_1^∞. We have,

$$\begin{aligned}\liminf_{n\to\infty} R_1^\Theta(P_{X_1^n}, D) &\ge \liminf_{n\to\infty} R_1(P_{X_1^n}, Q_{\theta_n}, D) = \lim_{k\to\infty} R_1(P_{X_1^{n_k}}, Q_{\theta_{n_k}}, D) \\ &\overset{\text{w.p.1}}{\ge} \liminf_{n\to\infty} \inf_{\theta'\in O(\theta^*,\epsilon)} R_1(P_{X_1^n}, Q_{\theta'}, D)\end{aligned} \tag{15}$$

for each $\epsilon > 0$. The first inequality is from (6) and the last is valid because infinitely many elements of (θ_{n_k}) are in $O(\theta^*, \epsilon)$ for any $\epsilon > 0$. Letting $\epsilon \downarrow 0$ in (15) and using (14) gives

$$\liminf_{n\to\infty} R_1^\Theta(P_{X_1^n}, D) \overset{\text{w.p.1}}{\ge} R_1(P, Q_{\theta^*}, D) \ge R_1^\Theta(P, D)$$

as desired. Note that with (5) this also implies that that θ^* achieves the infimum in the definition of $R_1^\Theta(P, D)$.

We need only prove (14). For any $\lambda \le 0$, $\theta \in \Theta$ and $\epsilon > 0$, the pointwise ergodic theorem gives

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n \sup_{\theta'\in O(\theta,\epsilon)} \log E_{\theta'}[e^{\lambda\rho(X_k,Y)}] \overset{\text{w.p.1}}{=} E_P\left[\sup_{\theta'\in O(\theta,\epsilon)} \log E_{\theta'}[e^{\lambda\rho(X,Y)}]\right]. \tag{16}$$

(See Section V-A for measurability.) Fix an at most countable, dense subset $\tilde{\Theta} \subseteq \Theta$. We can choose the exceptional sets in (16) independently of $\theta \in \tilde{\Theta}$ and $\epsilon > 0$ rational. For any $\theta \in \Theta$ and $\epsilon > 0$ we can choose a $\tilde{\theta} \in \tilde{\Theta}$ and a rational $\tilde{\epsilon} > \epsilon$ such that $O(\theta, \epsilon) \subseteq O(\tilde{\theta}, \tilde{\epsilon}) \subseteq O(\theta, 2\epsilon)$. Since the exceptional sets in (16) do not depend on $\tilde{\theta}$ and $\tilde{\epsilon}$, we have that

$$\begin{aligned}\limsup_{n\to\infty} \sup_{\theta'\in O(\theta,\epsilon)} \frac{1}{n} \sum_{k=1}^n \log E_{\theta'}[e^{\lambda\rho(X_k,Y)}] &\le \limsup_{n\to\infty} \frac{1}{n} \sum_{k=1}^n \sup_{\theta'\in O(\theta,\epsilon)} \log E_{\theta'}[e^{\lambda\rho(X_k,Y)}] \\ &\le \limsup_{n\to\infty} \frac{1}{n} \sum_{k=1}^n \sup_{\theta'\in O(\tilde{\theta},\tilde{\epsilon})} \log E_{\theta'}[e^{\lambda\rho(X_k,Y)}] \overset{\text{w.p.1}}{=} E_P\left[\sup_{\theta'\in O(\tilde{\theta},\tilde{\epsilon})} \log E_{\theta'}[e^{\lambda\rho(X,Y)}]\right] \\ &\le E_P\left[\sup_{\theta'\in O(\theta,2\epsilon)} \log E_{\theta'}[e^{\lambda\rho(X,Y)}]\right]\end{aligned} \tag{17}$$

simultaneously for all $\theta \in \Theta$ and $\epsilon > 0$, that is, the exceptional set can be chosen independently of θ and ϵ.

The monotone convergence theorem and the continuity in A1 give

$$\lim_{\epsilon \downarrow 0} E_P \left[\sup_{\theta' \in O(\theta, 2\epsilon)} \log E_{\theta'}[e^{\lambda \rho(X,Y)}] \right] = E_P \left[\lim_{\epsilon \downarrow 0} \sup_{\theta' \in O(\theta, 2\epsilon)} \log E_{\theta'}[e^{\lambda \rho(X,Y)}] \right] = E_P \left[\log E_\theta [e^{\lambda \rho(X,Y)}] \right].$$

Combining this with (17) and letting $\epsilon \downarrow 0$ gives

$$\lim_{\epsilon \downarrow 0} \limsup_{n \to \infty} \sup_{\theta' \in O(\theta, \epsilon)} \frac{1}{n} \sum_{k=1}^{n} \log E_{\theta'}[e^{\lambda \rho(X_k, Y)}] \overset{\text{w.p.1}}{\leq} E_P \left[\log E_\theta [e^{\lambda \rho(X,Y)}] \right] \tag{18}$$

simultaneously for all $\theta \in \Theta$.

Both sides of (18) are nondecreasing with λ. Furthermore, the right side of (18) is continuous from above for $\lambda < 0$. (To see this, use the dominated convergence theorem to move the limit through E_θ and the monotone convergence theorem to move the limit through E_P.) These two facts imply that we can also choose the exceptional sets independently of $\lambda \leq 0$ (by first applying (18) for λ rational and then squeezing). Applying (18) to the representation in (12) gives, for each $\lambda \leq 0$,

$$\begin{aligned} \lim_{\epsilon \downarrow 0} \liminf_{n \to \infty} \inf_{\theta' \in O(\theta, \epsilon)} R_1(P_{X_1^n}, Q_{\theta'}, D) &\geq \lim_{\epsilon \downarrow 0} \liminf_{n \to \infty} \inf_{\theta' \in O(\theta, \epsilon)} \left[\lambda D - \frac{1}{n} \sum_{k=1}^{n} \log E_{\theta'}[e^{\lambda \rho(X_k, Y)}] \right] \\ &= \lambda D - \lim_{\epsilon \downarrow 0} \limsup_{n \to \infty} \sup_{\theta' \in O(\theta, \epsilon)} \frac{1}{n} \sum_{k=1}^{n} \log E_{\theta'}[e^{\lambda \rho(X_k, Y)}] \overset{\text{w.p.1}}{\geq} \lambda D - E_P \left[\log E_\theta [e^{\lambda \rho(X,Y)}] \right] \end{aligned}$$

simultaneously for all $\theta \in \Theta$ and $\lambda \leq 0$. Optimizing over $\lambda \leq 0$ on the right gives (14).

D. Alternative Assumptions

Here we discuss the various alternative assumptions that imply A1 and A2. P1 implies A1 because $y \mapsto e^{\lambda \rho(x,y)}$ is bounded and measurable for each $x \in A$ and $\lambda \leq 0$. N1 implies A1 because $y \mapsto e^{\lambda \rho(x,y)}$ is bounded and continuous for each $x \in A$ and $\lambda \leq 0$.

1) P2 Implies A2: Here we prove that P2 implies A2 when (X_n) is stationary and ergodic with $X_1 \sim P$. Fix D, Δ and K according to P2, so that $T_\epsilon := \{\theta : Q_\theta(B(K, D + \Delta)) \geq \epsilon\}$ is relatively compact for each $\epsilon > 0$. We will first show that

$$\lim_{\epsilon \downarrow 0} \liminf_{n \to \infty} \inf_{\theta \in T_\epsilon^c} R_1(P_{X_1^n}, Q_\theta, D) \overset{\text{w.p.1}}{=} \infty \tag{19}$$

where T_ϵ^c denotes the complement of T_ϵ. If T_ϵ^c is empty for some $\epsilon > 0$, then (19) follows from the convention that $\inf \emptyset = \infty$. We can thus focus on the case where T_ϵ^c is not empty for all $\epsilon > 0$.

Define $\lambda_\epsilon := (\log \epsilon)/(D + \Delta)$. Since

$$\rho(x, y) \geq (D + \Delta) \mathbb{1}\{x \in K, y \in B(K, D + \Delta)^c\}$$

we have for any $\theta \in T_\epsilon^c$

$$\log E_\theta [e^{\lambda_\epsilon \rho(x, Y)}] \leq \mathbb{1}\{x \in K\} \log \left[\epsilon + e^{\lambda_\epsilon (D + \Delta)} \right] = \mathbb{1}\{x \in K\} \log(2\epsilon).$$

This and the representation in (12) imply that

$$\begin{aligned} \inf_{\theta \in T_\epsilon^c} R_1(P_{X_1^n}, Q_\theta, D) &\geq \inf_{\theta \in T_\epsilon^c} \left[\lambda_\epsilon D - \frac{1}{n} \sum_{k=1}^{n} \log E_\theta [e^{\lambda_\epsilon \rho(X_k, Y)}] \right] \\ &\geq \frac{D}{D + \Delta} \log \epsilon - \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}\{X_k \in K\} \log(2\epsilon). \end{aligned}$$

Taking limits, the pointwise ergodic theorem gives

$$\liminf_{n\to\infty} \inf_{\theta\in T_\epsilon^c} R_1(P_{X_1^n}, Q_\theta, D) \overset{\text{w.p.1}}{\geq} \frac{D}{D+\Delta}\log\epsilon - P(K)\log(2\epsilon). \tag{20}$$

Letting $\epsilon \downarrow 0$ (ϵ rational) and noting that $P(K) > D/(D+\Delta)$ by assumption gives (19).

Now we will show that (19) implies A2. Fix a realization x_1^∞ of X_1^∞ for which (19) holds. Let (n_k) be a subsequence for which

$$L := \liminf_{n\to\infty} R_1^\Theta(P_{x_1^n}, D) = \lim_{k\to\infty} R_1^\Theta(P_{x_1^{n_k}}, D).$$

If $L = \infty$, we can simply take $\theta_n = \theta$ for any constant θ and all n. If $L < \infty$, choose θ_{n_k} so that

$$\lim_{k\to\infty} R_1(P_{x_1^{n_k}}, Q_{\theta_{n_k}}, D) = L.$$

Then (19) implies that there exists an $\epsilon > 0$ for which θ_{n_k} must be in T_ϵ for all k large enough. Since T_ϵ has compact closure, the subsequence (θ_{n_k}) is relatively compact and it can always be embedded in a relatively compact sequence (θ_n). Since x_1^∞ is (with probability 1) arbitrary, the proof is complete.

2) N2 Implies A2: Here we prove that N2 implies A2 when (X_n) is stationary and ergodic with $X_1 \sim P$. For each $\epsilon > 0$ and each $M > 0$, let $K(\epsilon, M)$ be the set in N2. The pointwise ergodic theorem gives,

$$\lim_{n\to\infty} P_{X_1^n}(K(\epsilon, M)) \overset{\text{w.p.1}}{=} P(K(\epsilon, M)). \tag{21}$$

Fix a realization x_1^∞ of X_1^∞ for which (21) holds for all rational ϵ and M. Let (n_k) be a subsequence for which

$$L := \liminf_{n\to\infty} R_1^\Theta(P_{x_1^n}, D) = \lim_{k\to\infty} R_1^\Theta(P_{x_1^{n_k}}, D).$$

If $L = \infty$, we can simply take $\theta_n = \theta$ for any constant θ and all n. If $L < \infty$, for k large enough both sides are finite and we can choose $W_k \in W(P_{x_1^{n_k}}, D)$ so that

$$H(W_k \| W_k^A \times W_k^{\hat{A}}) \leq R_1^\Theta(P_{x_1^{n_k}}, D) + 1/k.$$

Let $Q_{\theta_{n_k}} = W_k^{\hat{A}}$ and note that

$$R_1(P_{x_1^{n_k}}, Q_{\theta_{n_k}}, D) \leq R_1^\Theta(P_{x_1^{n_k}}, D) + 1/k.$$

We will show that θ_{n_k} is relatively compact by showing that the sequence $(Q_k := Q_{\theta_{n_k}})$ is tight.[10] This will complete the proof just like in the previous section.

Fix $\epsilon > 0$ rational and $M > 2D/\epsilon$ rational. Let $K = K(\epsilon/2, M)$. We have

$$D \geq E_{(U,V)\sim W_k}[\rho(U,V)] \geq MW_k(K \times B(K,M)^c) \geq 2DW_k(K \times B(K,M)^c)/\epsilon.$$

This implies that $W_k(K \times B(K,M)^c) \leq \epsilon/2$ and we can bound

$$\begin{aligned} Q_k(B(K,M)) &= W_k^{\hat{A}}(B(K,M)) \geq W_k(K \times B(K,M)) = P_{x_1^{n_k}}(K) - W_k(K \times B(K,M)^c) \\ &\geq P_{x_1^{n_k}}(K) - \epsilon/2. \end{aligned}$$

Taking limits and applying (21) gives

$$\liminf_{n\to\infty} Q_k(B(K,M)) \geq P(K) - \epsilon/2 > 1 - \epsilon.$$

Since $B(K,M)$ has compact closure and since ϵ was arbitrary, the sequence (Q_k) is tight.

[10] A sequence of probability measures (Q_k) on $(\hat{A}, \hat{\mathcal{A}})$ is said to be *tight* if $\sup_F \liminf_{k\to\infty} Q_k(F) = 1$, where the supremum is over all compact (measurable) $F \subseteq \hat{A}$. If (Q_k) is tight, then Prohorov's Theorem states that it is relatively compact in the topology of weak convergence of probability measures [25].

E. Proof of Theorem 8

Here we prove the convergence-of-minimizers result given in Theorem 8. The proof of Theorem 5 in Section V-C shows that Θ^* is not empty. The assumptions ensure that both the lower and upper bounds for consistency of the plug-in estimator hold, so that $R_1^\Theta(P_{X_1^n}, D) \overset{\text{w.p.1}}{\to} R_1^\Theta(P, D)$. This shows that any sequence (θ_n) satisfying (9) also satisfies (6) with probability 1, and that the $\limsup$ and the $\liminf$ agree. Let θ^* be any limit point of this sequence (if one exists). Following the steps at the beginning of the proof of Theorem 5 in Section V-C, we see that $\theta^* \in \Theta^*$.

Now further suppose that $R_1^\Theta(P, D)$ is finite so that

$$R_1(P_{X_1^n}, Q_{\theta_n}, D) \overset{\text{w.p.1}}{\to} R_1^\Theta(P, D) < M < \infty. \tag{22}$$

We want to show that the sequence (θ_n) is relatively compact with probability 1. If P2 holds, then (19) immediately implies that there exists an $\epsilon > 0$ such that $\theta_n \in T_\epsilon$ eventually, with probability 1. Since T_ϵ is relatively compact, so is (θ_n).

Alternatively, suppose N2 holds. To show that (θ_n) is relatively compact with probability 1, we need only show that (Q_{θ_n}) is tight w.p.1. Fix a realization x_1^∞ where the convergence in (22) holds, where (21) holds for all rational ϵ and M, and where $R_1^\Theta(P_{x_1^n}, D) \to R_1^\Theta(P, D)$. For n large enough, the left side of (22) is finite, so $W(P_{x_1^n}, D)$ is not empty and we can choose a sequence (W_n) with $W_n \in W(P_{x_1^n}, D)$ so that

$$H(W_n \| P_{x_1^n} \times Q_{\theta_n}) \to R_1^\Theta(P, D).$$

Let $Q_n := W_n^{\hat{A}}$. An inspection of the above proof that N2 implies A2 shows that the sequence (Q_n) is tight. We will show that $H(Q_n \| Q_{\theta_n}) \to 0$, implying that (Q_{θ_n}) is also tight (because, for example, relative entropy bounds total variation distance). Indeed,

$$\underbrace{H(W_n \| P_{x_1^n} \times Q_{\theta_n})}_{a_n} = H(W_n \| P_{x_1^n} \times W_n^{\hat{A}}) + H(W_n^{\hat{A}} \| Q_{\theta_n}) \geq \underbrace{R_1^\Theta(P_{x_1^n}, D)}_{b_n} + \underbrace{H(Q_n \| Q_{\theta_n})}_{c_n}.$$

Since a_n and b_n both converge to $R_1^\Theta(P, D)$, which is finite, $c_n \to 0$, as claimed.

F. Proof of Theorem 10

Here we prove the result of Theorem 10, based on the law-of-large-numbers property. Inspecting all of the proofs in this paper reveals that the assumption of a stationary and ergodic source is only used to invoke the pointwise ergodic theorem. Furthermore, the pointwise ergodic theorem is not needed in full generality, only the LLN property is used. The relevant equations are (16), (20) and (21). Note that if ρ is bounded, then it is enough to have the LLN property hold for bounded f.

Equation (8) from Theorem 7, which we used in the proof of the upper bound, also assumes a stationary and ergodic source. The proof of a more general result than Theorem 7 is in [24] [23], but that result makes extensive use of the stationarity assumption. A careful reading reveals that only the LLN property is needed for (8). For completeness, we will give a proof, referring only to [23] for results that do not depend on the nature of the source. Specifically, what we need to prove for the upper bound is that

$$\limsup_{n\to\infty} R_1(P_{X_1^n}, Q, D) \overset{\text{w.p.1}}{\leq} R_1(P, Q, D) \tag{23}$$

for all $D \in D_{\text{c}}(P, Q)$.

If the source satisfies the LLN property for a random variable X with distribution P, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \log E_{Y\sim Q}[e^{\lambda\rho(X_k,Y)}] \overset{\text{w.p.1}}{=} E_{X\sim P}\left[\log E_{Y\sim Q}[e^{\lambda\rho(X,Y)}]\right] := \Lambda(\lambda). \tag{24}$$

Furthermore, since both sides are monotone in λ, the exceptional sets can be chosen independently of λ. The LLN property also implies that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} E_{Y\sim Q}[\rho(X_k, Y)] \stackrel{\text{w.p.1}}{=} E_{X\sim P}\left[E_{Y\sim Q}[\rho(X,Y)]\right] := D_{\text{ave}}. \tag{25}$$

Note that if ρ is bounded, then the LLN property need only hold for bounded f in both (24) and (25).

Define $\Lambda^*(D) := \sup_{\lambda\le 0}\left[\lambda D - \Lambda(\lambda)\right]$ and $D_{\min} := \inf\{D \ge 0 : \Lambda^*(D) < \infty\}$, with the convention that the infimum of the empty set equals $+\infty$. In [23] it is shown that $D_{\min} \le D_{\text{ave}}$, that Λ^* is convex, nonincreasing and continuous from the right, and that

$$\Lambda^*(D) = \begin{cases} \infty & \text{if } D < D_{\min} \\ \text{strictly convex} & \text{if } D_{\min} < D < D_{\text{ave}} \\ 0 & \text{if } D \ge D_{\text{ave}} \end{cases}$$

where some of these cases may be empty. Notice that Λ^* is continuous except perhaps at $D_{\min}$, where it will not be continuous from the left if $\Lambda^*(D_{\min}) < \infty$.

Fix a realization x_1^∞ of X_1^∞ for which (24) holds for all λ and for which (25) holds. Define the random variables

$$Z_n := \frac{1}{n}\sum_{k=1}^{n} \rho(x_k, Y_k)$$

for $n \ge 1$, where the sequence (Y_k) consists of independent and identically distributed (i.i.d.) random variables with common distribution Q. Then (24) implies that

$$\lim_{n\to\infty} \frac{1}{n} \log E[e^{\lambda n Z_n}] = \Lambda(\lambda). \tag{26}$$

We will first show that

$$\lim_{n\to\infty} -\frac{1}{n} \log \operatorname{Prob}\{Z_n \le D\} = \Lambda^*(D) = R(P,Q,D) \tag{27}$$

for all $D \ge 0$ except the special case when both $D = D_{\min}$ and $\Lambda^*(D_{\min}) < \infty$. The second equality in (27) is always valid [24] [23]. If $D < D_{\min}$, or $D = D_{\min}$ and $\Lambda^*(D_{\min}) = \infty$, or $D_{\min} < D \le D_{\text{ave}}$, the first equality in (27) follows from [23, Lemma 11], which is a slight modification of the Gärtner-Ellis Theorem in the theory of large deviations. The aforementioned properties of Λ^* and the convergence in (26) are what we need to use [23, Lemma 11]. If $D > D_{\text{ave}}$, then $\Lambda^*(D) = 0$ and we need only show that $\liminf_n \operatorname{Prob}\{Z_n \le D\} > 0$. But this follows from Chebychev's inequality and (25) because

$$\operatorname{Prob}\{Z_n \le D\} = 1 - \operatorname{Prob}\{Z_n > D\} \ge 1 - E[Z_n]/D \to 1 - D_{\text{ave}}/D > 0.$$

This proves (27), except for the special case when $D = D_{\min}$ and $\Lambda^*(D_{\min}) < \infty$ – which exactly corresponds to $D \notin D_{\text{c}}(P,Q)$.

Finally, (27) gives (23) because [24] [23]

$$R_1(P_{x_1^n}, Q, D) \le -\frac{1}{n} \log \operatorname{Prob}\{Z_n \le D\}$$

and because x_1^∞ is (with probability 1) arbitrary.

Acknowledgments

The authors wish to thank M. Madiman for many useful comments.

REFERENCES

[1] R. Ahlswede, "Extremal properties of rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. 36, no. 1, pp. 166–171, 1990.

[2] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Structures Algorithms*, vol. 19, no. 3-4, pp. 163–193, 2001.

[3] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inform. Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[4] H. Attouch and R. J.-B. Wets, "Epigraphical analysis," in *Analyse Non Linèaire*, ser. Annales de l'Institut Henri Poincaré, H. Attouch, J.-P. Aubin, F. Clarke, and I. Ekeland, Eds. Paris: Gauthier-Villars, 1989, pp. 73–100.

[5] T. Benjamin, "Rate distortion functions for discrete sources with continuous reproductions," Master's Thesis, Cornell University, 1973.

[6] P. Billingsley, *Convergence of Probability Measures*, 2nd ed. New York: John Wiley & Sons Inc., 1999.

[7] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[8] H. Cai, S. Kulkarni, and S. Verdú, "Universal entropy estimation via block sorting," *IEEE Trans. Inform. Theory*, vol. 50, no. 7, pp. 1551–1561, 2004.

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

[10] T. Cover and R. Wesel, "A gambling estimate of the rate-distortion function for images," in *Proc. Data Compression Conf. – DCC 94*, IEEE. Los Alamitos, California: IEEE Computer Society Press, 1994.

[11] I. Csiszár, "On an extremum problem of information theory," *Studia Scientiarum Mathematicarum Hungarica*, vol. 9, pp. 57–71, 1974.

[12] I. Csiszár, "On the computation of rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. 20, pp. 122–124, 1974.

[13] I. Csiszár and P. Shields, "Information theory and statistics: A tutorial," *Foundations and Trends in Communications and Information Theory*, vol. 1, pp. 1–111, 2004.

[14] A. Dembo and I. Kontoyiannis, "The asymptotics of waiting times between stationary processes, allowing distortion," *Ann. Appl. Probab.*, vol. 9, pp. 413–429, 1999.

[15] ——, "Source coding, large deviations, and approximate pattern matching," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1590–1615, June 2002.

[16] R. Dykstra and J. Lemke, "Duality of I projections and maximum likelihood estimation for log-linear models under cone constraints," *J. Amer. Statist. Assoc.*, vol. 83, no. 402, pp. 546–554, 1988.

[17] R. Gray and J. Kieffer, "Asymptotically mean stationary measures," *Ann. Probab.*, vol. 8, no. 5, pp. 962–973, 1980.

[18] M. Harrison, "The convergence of lossy maximum likelihood estimators," Brown University, Div. of Applied Mathematics, Providence, RI, APPTS #03-5, Jul. 2003.

[19] ——, "Epi-convergence of lossy likelihoods," Brown University, Div. of Applied Mathematics, Providence, RI, APPTS #03-4, Apr. 2003.

[20] ——, "The first order asymptotics of waiting times between stationary processes under nonstandard conditions," Brown University, Division of Applied Mathematics, Providence, RI, APPTS #03-3, Apr. 2003.

[21] M. Harrison and I. Kontoyiannis, "Maximum likelihood estimation for lossy data compression," in *Proc. 40th Ann. Allerton Conf. Comm. Contr. Comp.*, Allerton, IL, Oct. 2002, pp. 596–604.

[22] ——, "On estimating the rate-distortion function," in *Proc. 2006 IEEE ISIT*, Seattle, WA, Jul. 2006.

[23] M. T. Harrison. (2007, Nov.) The generalized asymptotic equipartition property: Necessary and sufficient conditions. [Online]. Available: http://arxiv.org/abs/0711.2666

[24] M. T. Harrison, "The generalized asymptotic equipartition property: Necessary and sufficient conditions," *IEEE Trans. Inform. Theory*, (to appear).

[25] O. Kallenberg, *Foundations of Modern Probability*, 2nd ed. New York: Springer, 2002.

[26] I. Kontoyiannis, P. Algoet, Y. Suhov, and A. Wyner, "Nonparametric entropy estimation for stationary processes and random fields, with applications to English text," *IEEE Trans. Inform. Theory*, vol. 44, no. 3, pp. 1319–1327, 1998.

[27] O. Koval and Y. Rytsar, "About estimation of the rate distortion function of the generalized Gaussian distribution under mean square error criteria," in *Proc. XVI Open Scientific and Technical of Young Scientists and Specialists of Institute of Physics and Mechanics*, Los Alamitos, California, May 2001, pp. 201–204, ySC-2001.

[28] L. Le Cam, "Maximum likelihood: An introduction," *International Statistical Review*, vol. 58, no. 2, pp. 153–171, 1990.

[29] M. Levene and G. Loizou, "Computing the entropy of user navigation in the web," *International Journal of Information Technology and Decision Making*, vol. 2, no. 3, pp. 459–476, 2000.

[30] D. Loewenstern and P. N. Yianilos, "Significantly lower entropy estimates for natural DNA sequences," *Journal of Computational Biology*, vol. 6, no. 1, pp. 125–142, 1999.

[31] W. Nemenman, W. Bialek, and R. de Ruyter van Steveninck, "Entropy and information in neural spike trains: progress on the sampling problem," *Physical Review E*, p. 056111, 2004.

[32] L. Paninski, "Estimation of entropy and mutual information," *Neural Comput.*, vol. 15, pp. 1191–1253, 2003.

[33] K. Rose, "A mapping approach to rate-distortion computation and analysis," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 1939–1952, 1994.

[34] G. Salinetti, "Consistency of statistical estimators: the epigraphical view," in *Stochastic Optimization: Algorithms and Applications*, S. Uryasev and P. M. Pardalos, Eds. Dordrecht: Kluwer Academic Publishers, 2001, pp. 365–383.

[35] T. Schürmann and P. Grassberger, "Entropy estimation of symbol sequences," *Chaos*, vol. 6, no. 3, pp. 414–427, 1996.

[36] E.-H. Yang and J. Kieffer, "On the performance of data compression algorithms based upon string matching," *IEEE Trans. Inform. Theory*, vol. 44, no. 1, pp. 47–65, 1998.