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, the problem of estimating the rate-distortion function of an unknown (not necessarily discrete-valued) source from empirical data is examined. The 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 demonstrating 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*—Consistency, entropy, estimation, maximum likelihood, plug-in estimator, rate-distortion function**

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, e.g., [6] [7] [8] [9] [10] [11]. Perhaps the simplest method is via the so-called **plug-in estimator**, where the entropy of P is estimated by $H(P_{x_1^n})$, 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: $H(P_{X_1^n}) \to H(P)$ in probability as $n \to \infty$. In fact, it is strongly consistent; the convergence holds with probability 1 (w.p.1) [12].

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. [13], 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. [14].

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 **plug-in estimator** $R(P_{X_1^n}, D)$, and determine conditions under which it is strongly consistent, that is, it converges to $R(P, D)$ w.p.1, as $n \to \infty$. We call this the **nonparametric estimation problem**, for reasons that will become clear below.

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. This paper is preceded by two technical reports [1] [2] and a longer version [3]. Preliminary results were presented at [4] [5].

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 w.p.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 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 [15] [16].

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 at least approximate $R(P_{x_1^n}, D)$, via, e.g., the Blahut-Arimoto algorithm [17] [18] [19]. 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 [20] [21].

The consistency problem can be framed in the following general setting. As has been observed by several authors, 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 to within distortion D [22] [23]. 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 $R^\Theta(P, D)$. Of course, when Θ indexes the set of all probability distributions on the reproduction alphabet, the parametric and nonparametric problems are identical, allowing 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 the strong consistency of the plug-in estimator in both the parametric and nonparametric estimation problems. It is shown that consistency holds in great generality for all distortion values D such that (s.t.) $R^\Theta(P, D)$ is continuous at D from the left. An example illustrating that consistency may actually fail at those points is given in Section III. 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. The proofs of all the results in this paper can be found in the longer version [3].

We consider extensions of these results in two directions. In Section IV-A we examine the problem of estimating the optimal reproduction distribution, i.e.. the distribution that achieves the infimum in equations (1) and (2). 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 that arise as modifications of the plug-in. These include penalized versions of the plug-in, analogous to the standard penalized maximum likelihood estimators in statistics.

The tools we employ to analyze convergence are based on the technique of epigraphical convergence [24] [25] (this is particularly clear in the proof of our main result, the lower bound in Theorem 5, in [3]), and it is no coincidence that these same tools have also provided one of the most successful approaches to proving the consistency of maximum likelihood estimators (MLEs).

Throughout the paper we work with stationary and ergodic (not only memoryless) sources,[1] though we are 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. Another reason for allowing dependence comes from simulation: Suppose, e.g., we wish to estimate the rate-distortion function of a distribution P that we cannot compute explicitly (as is the case of the majority of models used in image processing), but for which we have a Markov chain Monte Carlo (MCMC) sampling algorithm. The MCMC data is not memoryless, yet we only care about the first-order rate-distortion function.

II. Main Results

The following notation and definitions will remain in effect throughout the paper. Suppose the random source $(X_n) = (X_n : n \geq 1)$ 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.[2] 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,

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

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

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

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, $W^A \times W^{\hat{A}}$. 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 [27] [15], $D_{\rm c}(P)$ 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 [15].

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,

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

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**.

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

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

See [22] [23] 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 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.

[1] Since all of our results hold with probability one, we are effectively working with the much larger class of sources that are absolutely continuous w.r.t. some stationary and ergodic source. This is relevant for our comments about MCMC. Further results for nonstationary sources are in [3].

[2] 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 [26]. Henceforth, all σ-algebras are understood from the context; we do not complete any of them, 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.

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). 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,

$$D_c^\Theta(P) := \{D \geq 0 : R_1^\Theta(P,D) = \lim_{\lambda\uparrow 1} R_1^\Theta(P,\lambda D)\}.$$

Unlike $D_c(P)$, $D_c^\Theta(P)$ 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.,

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

[Throughout the paper we do not require limits to be finite, but say that $\lim_n a_n = \infty$ if a_n diverges to ∞; similarly for $-\infty$.] Of course in the special case where Θ indexes all probability distributions on $\hat{A}$, this reduces to the nonparametric problem, and (4) becomes $R_1(P_{X_1^n}, D) \overset{\text{w.p.1}}{\to} R_1(P,D)$. We separately treat the upper and lower bounds that combine to give (4).

The upper bound requires *no* further assumptions, although there can be certain pathological values of D for which it is not valid. In the nonparametric case, the only potential problem 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, for all $D \in D_c^\Theta(P)$:

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

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

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

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) that is relatively compact w.p.1 and s.t.

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

Theorem 5: If Θ is separable, A1 and A2 hold, and (X_n) is stationary and ergodic with $X_1 \sim P$, then for all $D \geq 0$:

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

Although A1 and A2 may seem quite involved, they are fairly easy to verify in specific examples: For A1, we show in [3] that either of the following two conditions imply A1.

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

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

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

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 [3] we prove that either one implies A2 as long as the source is stationary and ergodic with marginal distribution P, and in Section III we describe concrete situations where these assumptions are valid. 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,

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

P2. For each $D \geq 0$, there exists a $\Delta > 0$ and a $K \in \mathcal{A}$ s.t. $P(K) > D/(D+\Delta)$ 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 $K \in \mathcal{A}$ s.t. $P(K) > 1 - \epsilon$ and $B(K,M)$ is relatively compact. [Note that Θ can always be metrized in this way, and so that Θ will be separable (compact) if $\hat{A}$ is separable (compact) [28].]

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 nonparametric 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 $D \in D_c(P)$. Furthermore, the compactness condition can be relaxed as in N2.

III. Examples

Suppose the source (X_n) is stationary and ergodic with $X_1 \sim P$.

Example A. Nonparametric Consistency with 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 $D_c(P) = [0,\infty)$ regardless of P [15], and the plug-in estimator is strongly consistent for all P and all D.

This example also yields a new proof of the general consistency result mentioned in the Introduction, for the plug-in entropy estimator: If we take $A = \hat{A} = \{0,1,\ldots\}$, let $\rho(x,y) = |x-y|$, and take $D = 0$, then we obtain the strong consistency of [12, Cor. 1].

Example B. Nonparametric Consistency with 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 $D_{\mathrm{c}}(P)=[0,\infty)$ regardless of P [15] and the plug-in estimator is strongly consistent for all P and all D. This 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 extends to more general distortion measures on subsets $A, \hat{A}$ of $\mathbb{R}^d$; see [3].

Example C. Parametric Consistency for Gaussian Families

Let $A=\hat{A}=\mathbb{R}$, let ρ satisfy the assumptions of Example 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 $D\in D_{\mathrm{c}}^\Theta(P)$. In the special case where $\rho(x,y)=(x-y)^2$ is squared-error distortion, then it is not too difficult [23] to show that

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

where σ_X^2 denotes the (possibly infinite) variance of P, so $D_{\mathrm{c}}^\Theta(P)=[0,\infty)$ and convergence holds for all D. Furthermore, if the source P is 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.

Example 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, for any distribution P' on A, we have $R_1(P',D)=0$ for $D\geq P'(1)$ and $R_1(P',D)=\infty$ otherwise. If $P(1)>0$, the only possible trouble point for consistency is $D=P(1)$, which is not in $D_{\mathrm{c}}(P)$. 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 w.p.1 when the source is memoryless, and for a stationary ergodic source convergence will fail with positive probability [29].

Example E. Consistency at a Point of Discontinuity in P

This slightly modified example from [15] 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 $\rho(x,y):=P'(x)^{-1}\mathbb{1}\{x\neq y\}+|x-y|$. Note that $R_1(P',D)=\infty$ for all D (see [3] for a detailed explanation). This is a special case of Example 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 $R_1(P,D)\leq R_1(P,0)=H(P)<\infty$. Define the mixture distribution $P_\epsilon:=(1-\epsilon)P+\epsilon P'$. Then $P_\epsilon\to P$ in the topology of total variation as $\epsilon\downarrow 0$, but $R_1(P_\epsilon,D)\not\to R_1(P,D)$ because $R_1(P_\epsilon,D)\geq\epsilon R_1(P',D/\epsilon)=\infty$ for all $\epsilon>0$; see [3] for a proof of this last inequality.

Extensions of this example in [3] show that even closeness in relative entropy between two distributions is not enough to guarantee the closeness of the corresponding rate-distortion functions. Moreover, similar examples can be constructed for real-valued sources w.r.t. squared-error distortion.

Example 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,

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

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 $E[\rho_m(U,V)]\leq D$ for $\rho_m(x_1^n,y_1^n):=\frac{1}{m}\sum_{k=1}^m\rho(x_k,y_k)$. All the above results apply; we simply estimate the first-order rate-distortion function of the "blocked" process $(Z_k:=(X_k,\dots,X_{k+m-1}))$, w.r.t. the reproduction alphabet $\hat{A}^m$ and the distortion measure ρ_m, and 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 $n\geq 1$; 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 [29].

Theorem 7: [29] If the source (X_n) is stationary and ergodic with $X_1\sim P$, then for all $D\geq 0$ we have

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

and for all $D\in D_{\mathrm{c}}(P,Q):=\{D\geq 0: R_1(P,Q,D)=\lim_{\lambda\uparrow 1}R_1(P,Q,\lambda D)\}$ we have

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

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 $R_1(P,Q,D)$ is convex and nonincreasing in D [29], $D_{\mathrm{c}}(P,Q)$ is the entire interval $[0,\infty)$, except perhaps the single point where $R_1(P,Q,D)$ transitions from finite to infinite.

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 methods work equally well for moving an arg-infimum inside the limit. The next theorem, proved in [3], 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^* w.p.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 w.p.1. Hence $\theta_n \to \Theta^*$ w.p.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 other estimators. 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, e.g.,

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

for appropriate (nonnegative) penalty functions $F_n(\theta)$. The penalty functions express our preference for certain (typically less complex) subsets of Θ. This is particularly important when estimating the optimal reproduction distribution as discussed above. 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 [30]. Indeed, if $A = \hat{A}$ are discrete sets, ρ is Hamming distance and $D = 0$, then the estimator in (10) becomes,

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

which precisely corresponds to penalized maximum likelihood estimators. [Q^n denotes the n-fold product distribution of Q.]

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

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

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

For each such sequence (φ_n), we define a new estimator for $R_1^\Theta(P,D)$:

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

Condition (11a) implies that any lower bound for the plug-in estimator also holds here. Also, considering a single θ' for which $\limsup_n R_1(P_{X_1^n}, Q_{\theta'}, D) \leq R_1^\Theta(P,D)+\epsilon$, w.p.1, shows that (11b) similarly implies a corresponding upper bound:

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 (φ_n) satisfying (11a)(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 [4], namely,

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

where $B_n(x_1^n, D)$ denotes the distortion-ball of radius D around x_1^n, $B_n(x_1^n, D) := \{y_1^n \in \hat{A}^n : \rho_n(x_1^n, y_1^n) \leq D\}$; cf. [31]. Again, conditions (11a) and (11b) are valid in this case [29].

Acknowledgment

We are grateful to M. Madiman for many useful comments.

References

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

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

[3] M. T. Harrison and I. Kontoyiannis. (2007, Feb.) Estimation of the rate-distortion function. [Online]. Available: http://arxiv.org/abs/cs/0702018

[4] 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.

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

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

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

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

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

[18] 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.

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

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

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

[22] 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.

[23] A. Dembo and I. Kontoyiannis, "Source coding, large deviations, and approximate pattern matching," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1590–1615, June 2002.

[24] 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.

[25] 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.

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

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

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

[29] M. Harrison, "The generalized asymptotic equipartition property: Necessary and sufficient conditions," *IEEE Trans. Inform. Theory*, (in press).

[30] 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.

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

PLACE
PHOTO
HERE

Matthew T. Harrison received the Ph.D. degree in applied mathematics from Brown University, Providence, RI, in 2005.

He has had postdoctoral appointments at the Mathematical Sciences Research Institute and at Brown University. Currently, he is a Visiting Assistant Professor in the Department of Statistics, Carnegie Mellon University, Pittsburgh, PA. His research interests include lossy source coding, statistical methods for neuroscience and statistical approaches to computer vision.

PLACE
PHOTO
HERE

Ioannis Kontoyiannis was born in Athens, Greece, in 1972. He received the B.Sc. degree in mathematics in 1992 from Imperial College (University of London), and in 1993 he obtained a distinction in Part III of the Cambridge University Pure Mathematics Tripos. In 1997 he received the M.S. degree in statistics, and in 1998 the Ph.D. degree in electrical engineering, both from Stanford University. Between June and December 1995 he worked at IBM Research, on a NASA-IBM satellite image processing and compression project. From June 1998 to August 2001 we was an Assistant Professor with the Department of Statistics at Purdue University (and also, by courtesy, with the Department of Mathematics, and the School of Electrical and Computer Engineering). From August 2000 until July 2005 he has was an Assistant, then Associate Professor, with the Division of Applied Mathematics and with the Department of Computer Science at Brown University. Since March 2005 he has been an Associate Professor in the Department of Informatics, at the Athens University of Economics and Business.

In 2002 he was awarded the Manning endowed assistant professorship, and in 2005 he was awarded an honorary Master of Arts degree Ad Eundem, both by Brown University. In 2004 he was awarded a Sloan Foundation Research Fellowship. Currently he serves as an associate editor for the IEEE Transactions on Information Theory. His research interests include data compression, applied probability, information theory, statistics, simulation, and mathematical biology.