Sphere-Covering, Measure Concentration, and Source Coding

Ioannis Kontoyiannis

Abstract— **Suppose A is a finite set, let P be a discrete distribution on A, and let M be an arbitrary "mass" function on A. We give a precise characterization of the most efficient way in which A^n can be almost-covered using spheres of a fixed radius. An almost-covering is a subset C_n of A^n, such that the union of the spheres centered at the points of C_n has probability close to one with respect to the product distribution P^n. Spheres are defined in terms of a single-letter distortion measure on A^n, and an efficient covering is one with small mass $M^n(C_n)$. In information-theoretic terms the sets C_n are rate-distortion codebooks, but instead of minimizing their size we seek to minimize their mass. With different choices for M and the distortion measure on A our results give various corollaries as special cases, including Shannon's classical rate-distortion theorem, a version of Stein's lemma (in hypothesis testing), and a new converse to some measure-concentration inequalities on discrete spaces. Under mild conditions, we generalize our results to abstract spaces and non-product measures.**

Keywords— **Sphere covering, measure-concentration, data compression, large deviations.**

I. Introduction

SUPPOSE A is a finite set and let P a discrete probability mass function on A (more general probability spaces are considered later). Assume that the distortion (or distance) $\rho(x, y)$ between x and y is measured by a fixed $\rho : A\times A \to [0, \infty)$, and for each $n \geq 1$ define a single-letter distortion measure (or coordinate-wise distance function) ρ_n by

$$\rho_n(x_1^n, y_1^n) = \frac{1}{n}\sum_{i=1}^{n} \rho(x_i, y_i), \tag{1}$$

for $x_1^n = (x_1, x_2, \ldots, x_n)$ and $y_1^n = (y_1, y_2, \ldots, y_n)$ in A^n.

Given a $D \geq 0$, we want to "almost" cover the product space A^n using a finite number of balls (or "spheres") $B(y_1^n, D)$, where

$$B(y_1^n, D) = \{x_1^n \in A^n \ :\ \rho_n(x_1^n, y_1^n) \leq D\} \tag{2}$$

is the (closed) ball of distortion-radius D centered at $y_1^n \in A^n$. For our purposes, an "almost covering" is a subset $C \subset A^n$, such that the union of the balls of radius D centered at the points of C have large P^n-probability, that is,

$$P^n\left([C]_D\right) \quad \text{is close to 1}, \tag{3}$$

where $[C]_D$ is the D-*blowup of* C defined as

$$[C]_D \stackrel{\triangle}{=} \{x_1^n \ :\ \rho_n(x_1^n, y_1^n) \leq D \ \text{ for some } y_1^n \in C\}.$$

I. Kontoyiannis is with the Division of Applied Mathematics, Brown University, Box F, 182 George Street, Providence, RI 02912, USA. Email: yiannis@dam.brown.edu [Permanent address: Department of Statistics, Purdue University, 1399 Mathematical Sciences Building, W. Lafayette, IN 47907-1399. Email: yiannis@stat.purdue.edu.]

Research supported in part by NSF grant #0073378-CCR.

More specifically, given a "mass function" $M : A \to (0, \infty)$, we are interested in covering A^n efficiently, namely, finding sets C that satisfy (3) and also have small mass

$$M^n(C) = \sum_{y_1^n \in C} M^n(y_1^n) = \sum_{y_1^n \in C} \prod_{i=1}^{n} M(y_i).$$

Our main question of interest is the following:

$$(*) \begin{cases} \textit{If the sets } \{C_n\} \textit{ asymptotically } D\textit{-cover } A^n, \\ \quad \textit{i.e., } P^n\left([C_n]_D\right) \to 1 \textit{as } n \to \infty, \\ \textit{how small can their masses } M^n(C_n) \textit{ be?} \end{cases}$$

This is partly motivated by the fact that several interesting questions can be easily restated in this form. Three such examples are presented below, and in the remainder of the paper $(*)$ is addressed and answered in detail. In particular, it is shown that $M^n(C_n)$ typically grows (or decays) exponentially in n, and an explicit lower bound, valid for all finite n, is given for the exponent $(1/n)\log M^n(C_n)$ of the mass of an arbitrary C_n. [Throughout the paper, 'log' denotes the natural logarithm.] Moreover, a sequence of sets C_n asymptotically achieving this lower bound is exhibited, showing that it is best possible. The outline of the proofs follows, to some extent, along similar lines as the proof of Shannon's rate-distortion theorem [16]. In particular, the "extremal" sets C_n achieving the lower bound are constructed probabilistically; each C_n consists of a collection of points y_1^n generated by taking independent and identically distributed (i.i.d.) samples from a suitable distribution on A^n.

Example 1. Measure Concentration on the Binary Cube: Take $A = \{0, 1\}$ so that A^n is the n-dimensional binary cube consisting of all binary strings of length n, and let P^n be a product probability distribution on A^n. Write $\rho_n(x_1^n, y_1^n)$ for the normalized Hamming distortion between x_1^n and y_1^n, so that $\rho_n(x_1^n, y_1^n)$ is the proportion of mismatches between the two strings; formally:

$$\rho_n(x_1^n, y_1^n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{x_i \neq y_i\}}, \quad x_1^n, y_1^n \in A^n. \tag{4}$$

Geometrically, if A^n is given the usual nearest-neighbor graph structure (two points are connected if and only if they differ in exactly one coordinate), then $\rho_n(x_1^n, y_1^n)$ is the graph distance between x_1^n and y_1^n, normalized by n.

A well-known measure-concentration inequality for subsets C_n of A^n states that, for any $D \geq 0$,

$$P^n([C_n]_D) \geq 1 - \frac{e^{-nD^2/2}}{P^n(C_n)}. \tag{5}$$

[See Proposition 2.1.1 in the comprehensive account by Talagrand [18], or Theorem 3.5 in the review paper by McDiarmid [13], and the references therein.] Roughly speaking, (5) says that "if C_n is not too small, $[C_n]_D$ is almost everything." In particular, it implies that for any sequence of sets $C_n \subset A^n$ and any $D \geq 0$,

$$\begin{aligned} &\textit{if} \qquad \liminf_{n\to\infty} \frac{1}{n} \log P^n(C_n) > -D^2/2, \\ &\textit{then} \qquad P^n([C_n]_D) \to 1. \end{aligned} \tag{6}$$

A natural question to ask is whether there is a converse to the above statement: If $P^n([C_n]_D) \to 1$, how small can the probabilities of the C_n be? Taking $M \equiv P$, this reduces to question $(*)$ above. In this context, $(*)$ can be thought of as the opposite of the usual isoperimetric problem. We are looking for sets with the "largest possible boundary"; sets C_n whose D-blowups (asymptotically) cover the entire space, but whose volumes $P^n(C_n)$ are as small as possible. A precise answer for this problem is given in Corollary 3 and the discussion following it, in the next section.

Example 2. Lossy Data Compression: Let A be a finite alphabet so that A^n consists of all possible messages of length n from A, and assume that messages are generated by a memoryless source with distribution P^n on A^n. A code for these messages consists of a codebook $C_n \subset A^n$ and an encoder $\phi_n : A^n \to C_n$. If we think of $\rho_n(x_1^n, y_1^n)$ as the distortion between a message x_1^n and its reproduction y_1^n, then for any given codebook C_n the best choice for the encoder is clearly the map ϕ_n taking each x_1^n to the y_1^n in C_n which minimizes the distortion $\rho_n(x_1^n, y_1^n)$. Hence, at least conceptually, finding good codes is the same as finding good codebooks. More specifically, if $D \geq 0$ is the maximum amount of distortion we are willing to tolerate, then a sequence of *good* codebooks $\{C_n\}$ is one with the following properties:

(a) The probability of encoding a message with distortion exceeding D is asymptotically negligible:

$$P^n([C_n]_D) \to 1.$$

(b) Good compression is achieved, that is, the sizes $|C_n|$ of the codebooks are small.

What is the best achievable compression performance? That is, if the codebooks $\{C_n\}$ satisfy (a), how small can their sizes be? Shannon's classical source coding theorem (cf. [16][2]) answers this question. In our notation, taking $M \equiv 1$ reduces the question to a special case of $(*)$, and in Corollary 2 in the next section we recover Shannon's theorem as a special case of Theorems 1 and 2.

Example 3. Hypothesis Testing: Let A be a finite set and P_1, P_2 be two probability distributions on A. Suppose that the null hypothesis that a sample $X_1^n = (X_1, X_2, \ldots, X_n)$ of n independent observations comes from P_1 is to be tested against the simple alternative hypothesis that X_1^n comes from P_2. A test between these two hypotheses can be thought of as a decision region $C_n \subset A^n$: If $X_1^n \in C_n$ we declare that $X_1^n \sim P_1^n$, otherwise we declare $X_1^n \sim P_2^n$. The two probabilities of error associated with this test are

$$\alpha_n = P_1^n(C_n^c) \quad \text{and} \quad \beta_n = P_2^n(C_n). \tag{7}$$

A good test has these two probabilities vanishing as fast as possible, and we may ask, if $\alpha_n \to 0$, how fast can β_n decay to zero? Taking ρ to be Hamming distortion, $D = 0$, $P = P_1$, and $M = P_2$, this reduces to our original question $(*)$. In Corollary 1 in the next section we answer this question by deducing a version of Stein's lemma from Theorems 1 and 2. It is worth noting that the connection between questions in hypothesis testing and information theory goes at least as far back as Strassen's 1964 paper [17] (see also Blahut's paper [3] in 1974, and Csiszár and Körner's book [6] for a detailed discussion).

The rest of the paper is organized as follows. In Section II, Theorems 1 and 2 provide an answer to question $(*)$. In the remarks and corollaries following Theorem 2 we discuss and interpret this answer, and we present various applications along the lines of the three examples above. In Section III we consider the same problem in a much more general setting. We let A be an abstract space, and instead of product measures P^n we consider the n-dimensional marginals P_n of a stationary measure $\mathbb{P}$ on $A^{\mathbb{N}}$. In Theorems 3 and 4 we give analogs of Theorems 1 and 2, which hold essentially as long as the spaces (A^n, P_n) can be almost-covered by countably many ρ_n-balls. Since the results of Section II are essentially subsumed by Theorems 3 and 4, we only give the proofs of the more general statements, Theorems 3 and 4; they are proved in Section IV, and the Appendix contains the proofs of various technical steps needed along the way.

II. The Discrete Memoryless Case

Let A be a finite set and P be a discrete probability mass function on A. Fix a $\rho : A \times A \to [0, \infty)$, and for each $n \geq 1$ let ρ_n be the corresponding single-letter distortion measure on A^n defined as in (1). Also let $M : A \to (0, \infty)$ be an arbitrary positive mass function on A. We assume, without loss of generality, that $P(a) > 0$ for all $a \in A$, and also that for each $a \in A$ there exists a $b \in A$ with $\rho(a, b) = 0$ (otherwise we may consider $\rho'(x, y) = [\rho(x, y) - \min_{z\in A} \rho(x, z)]$ instead of ρ). Let $\{X_n\}$ denote a sequence of i.i.d. random variables with distribution P, and write $\mathbb{P} = P^{\mathbb{N}}$ for the product measure on $A^{\mathbb{N}}$ equipped with the usual σ-algebra generated by finite-dimensional cylinders. We write X_i^j for vectors of random variables $(X_i, X_{i+1}, \ldots, X_j)$, $1 \leq i \leq j \leq \infty$, and similarly $x_i^j = (x_i, x_{i+1}, \ldots, x_j) \in A^{j-i+1}$ for realizations of these random variables.

Next we define the rate function $R(D)$ that will provide the lower bound on the exponent of the mass of an arbitrary $C_n \subset A^n$. For $D \geq 0$ and Q a probability measure on A, let

$$I(P, Q, D) = \inf_{W \in \mathcal{M}(P,Q,D)} H(W \| P \times Q) \tag{8}$$

where $H(\mu\|\nu)$ denotes the relative entropy between the probability measures μ and ν, and $\mathcal{M}(P, Q, D)$ consists of all probability measures W on $A\times A$ such that W_X, the first marginal of W, is equal to P, W_Y, the second marginal, is Q, and $E_W[\rho(X,Y)] \leq D$; if $\mathcal{M}(P,Q,D)$ is empty, we let $I(P,Q,D)=\infty$. The rate function $R(D)$ is defined by

$$\begin{aligned} R(D) &= R(D;P,M) \\ &= \inf_Q \{I(P,Q,D) + E_Q[\log M(Y)]\}, \end{aligned} \qquad (9)$$

where the infimum is over all probability distributions Q on A. Recalling the definition of mutual information and combining the two infima in (8) and (9), $R(D)$ can equivalently be written in a more information-theoretic way as

$$\inf_{(X,Y):\ X\sim P,\ E\rho(X,Y)\leq D} \{I(X;Y) + E[\log M(Y)]\} \qquad (10)$$

where the infimum is taken over all jointly distributed random variables (X,Y) such that X has distribution P and $E\rho(X,Y) \leq D$. For any $x_1^n \in A^n$ and $C_n \subset A^n$, write

$$\rho_n(x_1^n, C_n) = \min_{y_1^n \in C_n} \rho_n(x_1^n, y_1^n).$$

In the following two Theorems we answer question $(*)$ stated in the Introduction. Theorem 1 contains a lower bound (valid for all n) on the mass of an arbitrary $C_n \subset A^n$, and Theorem 2 shows that this bound is asymptotically tight. In information-theoretic terms, Theorems 1 and 2 are generalized direct and converse coding theorems, for minimal-mass (rather than minimal-size) codebooks.

Theorem 1. Let $C_n \subset A^n$ be arbitrary and write $D = E_{P^n}[\rho_n(X_1^n, C_n)]$. Then

$$\frac{1}{n}\log M^n(C_n) \geq R(D).$$

Theorem 2. Assume that $\rho(x,y)=0$ if and only if $x=y$. For any $D\geq 0$ and any $\epsilon>0$ there is a sequence of sets $\{C_n\}$ such that:

$$\begin{aligned} &(i) \qquad \frac{1}{n}\log M^n(C_n) \leq R(D)+\epsilon \qquad \text{for all } n\geq 1 \\ &(ii) \qquad \rho_n(X_1^n, C_n) \leq D \qquad \text{eventually, } \mathbb{P}-a.s. \end{aligned}$$

Remark 1. From part (ii) of Theorem 2 we have that $\mathbb{I}_{[C_n]_D}(X_1^n) \to 1$ with probability one, so by Fatou's lemma, $P^n([C_n]_D) \to 1$. From this and (i) it is easy to deduce the following alternative version of Theorem 2 (see [11] for a proof): *For any $D\geq 0$ there is a sequence of sets $\{C_n^*\}$ such that:*

$$\begin{aligned} &(i') \qquad \limsup_{n\to\infty} \frac{1}{n}\log M^n(C_n^*) \leq R(D) \\ &(ii') \qquad P^n([C_n^*]_D) \to 1, \quad \text{and} \\ &(iii') \qquad \limsup_{n\to\infty} E_{P^n}[\rho_n(X_1^n, C_n^*)] \leq D \end{aligned}$$

Remark 2. The additional assumption on ρ in Theorem 2 is only made for the sake of simplicity, and it is not necessary for the validity of the result.

Theorems 3 and 4 in the following section give more general versions of Theorems 1 and 2, so their proofs are postponed until then. However, it is worth mentioning here that in the discrete case, Theorems 1 and 2 can be given much simpler proofs. In particular, Theorem 2 can be given an elementary proof by a direct application of Sanov's theorem (see [11]). Alternatively, Theorem 2 follows from Csiszár and Körner's type-covering lemma [6, p. 151].

Although the proof of Theorem 2 (and the more general version in Theorem 4) is somewhat technical, the idea behind the construction of the extremal sets C_n is simple: Suppose Q^* is a probability measure on A achieving the infimum in the definition of $R(D)$, so that

$$R(D) = I(P,Q^*,D) + E_{Q^*}[\log M(Y)] \triangleq I^* + L^*.$$

Write Q_n^* for the product measure $(Q^*)^n$, and let $\widehat{Q}_n$ be the measure obtained by conditioning Q_n^* to the set of points $y_1^n \in A^n$ whose empirical measures ("types") are uniformly close to Q^*. Then let C_n consist of approximately e^{nI^*} points y_1^n drawn i.i.d. from $\widehat{Q}_n$. Each point in the support of $\widehat{Q}_n$ has mass $M^n(y_1^n) \approx e^{nL^*}$ and C_n contains about e^{nI^*} of them, so $M^n(C_n)$ is close to $e^{nI^*}e^{nL^*} = e^{nR(D)}$. The main technical content of the proof is therefore to prove (ii), namely, that e^{nI^*} points indeed suffice to almost D-cover A^n.

The above construction also provides a nice interpretation for $R(D)$. If we had started with a different measure Q in place of Q^*, we would have ended up with sets C_n' of size $\approx \exp(nI(P,Q,D))$, consisting of points y_1^n of mass $M^n(y_1^n) \approx \exp(nE_Q(\log M(Y)))$, and the total mass of C_n' would be

$$M^n(C_n') \approx \exp\{n[I(P,Q,D) + E_Q(\log M(Y))]\}.$$

By optimizing over the choice of Q in (9) we are balancing the tradeoff between the size and the weight of the set C_n, between a few heavy points and many light ones.

It is also worth noting that the extremal sets C_n above were constructed by taking samples y_1^n from the measure $\widehat{Q}_n$. Unlike the usual proofs of the data compression theorem, here we cannot simply use the product measure Q_n^*. This is because we are not just interested in how many points y_1^n are needed to almost cover A^n, but also we need to control their masses $M^n(y_1^n)$. Since exponentially many y_1^n's are required to cover A^n, if they are generated from Q_n^* then there are bound to be some atypically heavy ones, and this drastically increases the total mass $M^n(C_n)$. Therefore, by restricting Q_n^* to be supported on the set of $y_1^n \in A^n$ whose empirical measures are uniformly close to Q^*, we are ensuring that the masses of the y_1^n will be essentially constant, and all approximately equal to e^{nL^*}.

Next we derive corollaries from Theorems 1 and 2, along the lines of the examples in the Introduction. First, in the context of hypothesis testing, let P_1, P_2 be two probability distributions on A with all positive probabilities. Suppose that the null hypothesis that $X_1^n \sim P_1^n$ is to be tested against the alternative $X_1^n \sim P_2^n$. Given a test with an associated decision region $C_n \subset A^n$, its two probabilities of error α_n and β_n are defined as in (7). In the notation of this section, let ρ_n be Hamming distortion as in (4), $P = P_1$ and $M = P_2$. Observe that, here,

$$E_{P_1^n}[\rho_n(X_1^n, C_n)] \le E_{P_1^n}[\mathbb{I}_{C_n^c}(X_1^n)] = P_1^n(C_n^c),$$

and define, in the notation of (9), the error exponent

$$\varepsilon(\alpha) = -R(\alpha; P_1, P_2), \quad \alpha \in [0,1].$$

Noting that $\varepsilon(0) = H(P_1\|P_2)$, from Theorems 1 and 2 and Remark 1 we obtain the following version of Stein's lemma (see Lemma 6.1 in Bahadur's monograph [1], or Theorem 12.8.1 in [5]).

Corollary 1. Hypothesis Testing: Let $\alpha = \alpha_n = P_1^n(C_n^c)$ and $\beta = \beta_n = P_2^n(C_n)$ be the two error probabilities associated with an arbitrary sequence of tests $\{C_n\}$.

(a) For all $n \ge 1$, $\beta \ge e^{-n\varepsilon(\alpha)}$.

(b) If $\alpha_n \to 0$, then

$$\liminf_{n\to\infty} \frac{1}{n} \log \beta_n \ge -H(P_1\|P_2).$$

(c) There exists a sequence of decision regions C_n with associated tests whose error probabilities achieve $\alpha_n \to 0$ and $(1/n)\log\beta_n \to -H(P_1\|P_2)$, as $n \to \infty$.

Note that, although the decision regions C_n in (c) above achieve the best exponent in the error probability, they are not the overall optimal decision regions in the Neyman-Pearson sense [6].

In the case of data compression, we have random data X_1^n generated by some product distribution P^n. Given a single-letter distortion measure ρ_n and a maximum allowable distortion level $D \ge 0$, our objective is to find good codebooks C_n. As discussed in Example 2 above, good codebooks are those that asymptotically cover A^n, i.e., $P^n([C_n]_D) \to 1$, and whose sizes $|C_n|$ are relatively small. In our notation, if we take $M(\cdot) \equiv 1$, then $M^n(C_n) = |C_n|$ and the rate function $R(D)$ (from (9) or (10)) reduces to Shannon's *rate-distortion function*

$$\begin{aligned} R_S(D) &= \inf_Q \inf_{W\in\mathcal{M}(P,Q,D)} H(W\|P\times Q) \\ &= \inf_{(X,Y):\, X\sim P,\, E\rho(X,Y)\le D} I(X;Y). \end{aligned}$$

From Theorems 1 and 2 and Remark 1 we recover Shannon's source coding theorem (see [16][2]).

Corollary 2. Data Compression: For any $n \ge 1$, if the average distortion achieved by a codebook C_n is $D = E_{P^n}[\rho_n(X_1^n, C_n)]$, then

$$\frac{1}{n}\log|C_n| \ge R_S(D).$$

Moreover, for any $D \ge 0$, there is a sequence of codebooks $\{C_n\}$ such that $E_{P^n}[\rho_n(X_1^n, C_n)] \to D$, the codebooks C_n asymptotically cover A^n, $P^n([C_n]_D) \to 1$, and

$$\lim_{n\to\infty} \frac{1}{n}\log|C_n| = R_S(D).$$

Finally, in the context of measure-concentration, taking $M = P$ and writing $R_C(D)$ for the concentration exponent $R(D; P, P)$, we get:

Corollary 3. Converse Measure Concentration: Let $\{C_n\}$ be arbitrary sets.

(i) For any $n \ge 1$, if $D = E_{P^n}[\rho_n(X_1^n, C_n)]$, then $P^n(C_n) \ge e^{nR_C(D)}$.

(ii) If $P^n([C_n]_D) \to 1$, then

$$\liminf_{n\to\infty} \frac{1}{n}\log P^n(C_n) \ge R_C(D).$$

(iii) There is a sequence of sets $\{C_n\}$ that satisfy $P^n([C_n]_D) \to 1$ and $(1/n)\log P^n(C_n) \to R_C(D)$, as $n \to \infty$.

In particular, in the case of the binary cube, part (ii) of the corollary provides a precise converse to the measure-concentration statement in (6). Although the concentration exponent $R_C(D) = R(D; P, P)$ is not as explicit as the exponent $-D^2/2$ in (6), $R_C(D)$ is a well-behaved function and it is easy to evaluate it numerically. For example, Figure 1 shows the graph of $R_C(D)$ in the case of the binary cube, with P being the Bernoulli measure with $P(1) = 0.4$.

Fig. 1. Graph of the function $R_C(D) = R(D; P, P)$ for $0 \le D \le 1$, in the case of the binary cube $A^n = \{0,1\}^n$, with $P(1) = 0.4$.

In contrast with the measure concentration exponent $-D^2/2$ in (6), the quantity $R_C(D)$ actually depends on the distribution P. This is not a shortcoming of our method – it is part of the intrinsic structure of the problem.

Various easily checked properties of $R(D) = R(D; P, M)$ are stated without proof in Lemma 1 below – see [11] for a proof.

As mentioned in the Introduction, the question considered in Corollary 3 can be thought of as the opposite of the usual isoperimetric problem. Instead of large sets with small boundaries, we are looking for *small* sets with the *largest possible boundary.* It is therefore not surprising that the extremal sets in (6) and in Corollary 3 are very different. In the classical isoperimetric problem, the extremal sets typically look like Hamming balls around $0^n = (0, 0, \ldots, 0) \in A^n$, $B_n = \{x_1^n : \rho_n(x_1^n, 0^n) \le r/n\}$ (see the discussions in Section 2.3 of [18], p. 174 in [12], or the original paper by Harper [9]), while the extremal sets in our case are collections of vectors y_1^n drawn i.i.d. from the measure $\widehat{Q}_n$ on A^n.

Lemma 1.

(i) $R(D)$ is finite, nonincreasing, convex, and continuous for all $D \ge 0$.

(ii) If we let $R_{\min} = \min\{\log M(y) : y \in A\}$ and define $D_{\max} = D_{\max}(P)$ as

$$\min\{E_P[\rho(X, y)] : y \text{ such that } \log M(y) = R_{\min}\},$$

then

$$R(D) \text{ is } \begin{cases} = R_{\min} & \text{for } D \ge D_{\max} \\ > R_{\min} & \text{for } 0 \le D < D_{\max}. \end{cases}$$

III. The General Case

Let A be a Polish space (namely, a complete, separable metric space) equipped with its associated Borel σ-algebra $\mathcal{A}$, and let $\mathbb{P}$ be a probability measure on $(A^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}})$. Also let $(\hat{A}, \hat{\mathcal{A}})$ be a (possibly different) Polish space. Given a nonnegative measurable function $\rho : A \times \hat{A} \to [0, \infty)$, define $\rho_n : A^n \times \hat{A}^n \to [0, \infty)$ as in (1).

Let $\{X_n\}$ be a sequence of random variables distributed according to $\mathbb{P}$, and for each $n \ge 1$ write P_n for the n-dimensional marginal distribution of X_1^n. We say that $\mathbb{P}$ is a stationary measure if X_1^n has the same distribution as X_{1+k}^{n+k}, for any n, k. Let $M : \hat{A} \to (0, \infty)$ be a measurable "mass" function on $\hat{A}$, and for each $n \ge 1$ define

$$M^n(y_1^n) \triangleq \prod_{i=1}^n M(y_i) \quad y_1^n \in \hat{A}^n.$$

In order to avoid uninteresting technicalities we will assume throughout that M is bounded above and below, that is,

$$|\log M(y)| \le L_{\max} < \infty \quad \text{for all } y \in \hat{A}$$

for some constant $L_{\max}$. Next we define the natural analogs of the rate functions $I(P, Q, D)$ and $R(D)$. For $n \ge 1$, $D \ge 0$ and Q_n a probability measure on $(\hat{A}^n, \hat{\mathcal{A}}^n)$, let

$$I_n(P_n, Q_n, D) = \inf_{W_n \in \mathcal{M}_n(P_n, Q_n, D)} H(W_n \| P_n \times Q_n) \quad (11)$$

where $\mathcal{M}_n(P_n, Q_n, D)$ consists of all probability measures W_n on $(A^n \times \hat{A}^n, \mathcal{A}^n \times \hat{\mathcal{A}}^n)$ such that $W_{n,X}$, the first marginal of W_n, is equal to P_n, the second marginal $W_{n,Y}$ is Q_n, and $\int \rho_n \, dW_n \le D$; if $\mathcal{M}_n(P_n, Q_n, D)$ is empty, let $I_n(P_n, Q_n, D) = \infty$. Then $R_n(D) = R_n(D; P_n, M)$ is defined by

$$\inf_{Q_n} \{I_n(P_n, Q_n, D) + E_{Q_n}[\log M^n(Y_1^n)]\}, \quad (12)$$

where the infimum is over all probability measures Q_n on $(\hat{A}^n, \hat{\mathcal{A}}^n)$. Note that since $I_n(P_n, Q_n, D)$ is nonnegative and M is bounded away from zero, $R_n(D)$ is always well-defined. Recalling the definition of mutual information, $R_n(D)$ can also be written in a form analogous to (10) in the discrete case

$$R_n(D) = \inf_{(X_1^n, Y_1^n)} \{I(X_1^n; Y_1^n) + E[\log M^n(Y_1^n)]\} \quad (13)$$

where the infimum is taken over all jointly distributed (X_1^n, Y_1^n) such that $X_1^n \sim P_n$ and $E\rho_n(X_1^n, Y_1^n) \le D$. Finally, the rate function $R(D)$ is defined by

$$R(D) = \lim_{n\to\infty} \frac{1}{n} R_n(D)$$

whenever the limit exists. Next we state some simple properties of $R_n(D)$ and $R(D)$, proved in the Appendix.

Lemma 2.

(i) For each $n \ge 1$, $R_n(D)$ is nonincreasing and convex in $D \ge 0$, and therefore also continuous at all D except possibly at the point

$$D_{\min}^{(n)} = \inf\{D \ge 0 : R_n(D) < +\infty\}.$$

(ii) If $R(D)$ exists for all $D \ge 0$ then it is nonincreasing and convex in $D \ge 0$, and therefore also continuous at all D except possibly at the point

$$D_{\min} = \inf\{D \ge 0 : R(D) < +\infty\}.$$

(iii) If $\mathbb{P}$ is a stationary measure, then

$$R(D) = \lim_{n\to\infty} \frac{1}{n} R_n(D) = \inf_{n \ge 1} \frac{1}{n} R_n(D) \quad \text{exists},$$

and $D_{\min} = \inf_n D_{\min}^{(n)}$.

(iv) The mutual information $I(X_1^n; Y_1^n)$ is concave in the marginal distribution P_n of X_1^n for a fixed conditional distribution of Y_1^n given X_1^n, and convex in the conditional distribution of Y_1^n given X_1^n for a fixed marginal distribution of X_1^n.

Next we state analogs of Theorems 1 and 2 in the general case. As before, we are interested in sets C_n that have large blowups but small masses; since M is bounded away from zero we may restrict our attention to finite sets C_n.

Theorem 3. Let $C_n \subset \hat{A}^n$ be an arbitrary finite set and write $D = E_{P_n}[\rho_n(X_1^n, C_n)]$. Then

$$\log M^n(C_n) \ge R_n(D). \quad (14)$$

If $\mathbb{P}$ is a stationary measure, then for all $n \ge 1$

$$\log M^n(C_n) \ge nR(D).$$

As will become apparent from its proof (in the following section), Theorem 3 remains true in great generality. The exact same proof works for arbitrary (non-product) positive mass functions M_n in place of M^n, and more general distortion measures ρ_n, not necessarily of the form in (1). Moreover, as long as $R_n(D)$ is well-defined, the assumption that M is bounded away from zero is unnecessary. In that case we can also consider countably infinite sets C_n, and (14) remains valid as long as $R_n(D)$ is continuous in D (see Lemma 2).

In the special case when $\mathbb{P}$ is a product measure it is not hard to check that $R_n(D) = nR(D)$ for all $n \geq 1$, so we can recover Theorem 1 from Theorem 3.

For Theorem 4 some additional assumptions are needed. We will assume that the function ρ is bounded, i.e., that there for some finite constant $\rho_{\max}$, $\rho(x,y) \leq \rho_{\max}$ for all $x \in A$, $y \in \hat{A}$. For $k \geq 1$, we say that $\mathbb{P}$ is stationary (respectively, ergodic) in k-blocks if the process $\{\widetilde{X}_n^{(k)}\ ;\ n \geq 0\} = \{X_{nk+1}^{(n+1)k}\ ;\ n \geq 0\}$ is stationary (resp. ergodic). If $\mathbb{P}$ is stationary then it is stationary in k-blocks for every k. But an ergodic measure $\mathbb{P}$ may not be ergodic in k-blocks. For the second part of the Theorem we will assume that $\mathbb{P}$ is ergodic in blocks, that is, that it is ergodic in k-blocks for all $k \geq 1$. Also, since $R(D) = \infty$ for D below $D_{\min}$, we restrict our attention to the case $D > D_{\min}$. Theorem 4 is proved in the next section.

Theorem 4. Assume that the functions ρ and $\log M$ are bounded, and that $\mathbb{P}$ is a stationary ergodic measure. For any $D > D_{\min}$ and any $\epsilon > 0$, there is a sequence of sets $\{C_n\}$ such that:

$$(i) \qquad \frac{1}{n}\log M^n(C_n) \leq R(D) + \epsilon \qquad \text{for all } n \geq 1$$

$$(ii) \qquad P_n([C_n]_D) \to 1 \qquad \text{as } n \to \infty.$$

If, moreover, $\mathbb{P}$ is ergodic in blocks, there are sets $\{C_n\}$ that satisfy (i) and

$$(iii) \qquad \rho_n(X_1^n, C_n) \leq D \qquad \text{eventually, } \mathbb{P}-\text{a.s.}$$

Remark 3. A corresponding version of the asymptotic form of Theorems 1 and 2 given in Remark 1 of the previous section can also be derived here, and it holds for every stationary ergodic $\mathbb{P}$.

Remark 4. The assumptions on the boundedness of ρ and $\log M$ are made for the purpose of technical convenience, and can probably be relaxed to appropriate moment conditions. Similarly, the assumption that M^n is a product measure can be relaxed to include sequences of measures M_n that have rapid mixing properties. Finally, the assumption that $\mathbb{P}$ is ergodic in blocks is not as severe as it may sound. For example, it is easy to see that any weakly mixing measure (in the ergodic-theoretic sense – see [14]) is ergodic in blocks.

IV. Proofs of Theorems 3 and 4

Proof of Theorem 3: Given an arbitrary C_n, let $\phi_n : A^n \to C_n$ be a function that maps each $x_1^n \in A^n$ to the closest y_1^n in C_n, i.e., $\rho_n(x_1^n, \phi(x_1^n)) = \rho_n(x_1^n, C_n)$. For $X_1^n \sim P_n$ define $Y_1^n = \phi_n(X_1^n)$, write Q_n for the (discrete) distribution of Y_1^n, and $W_n(dx_1^n, dy_1^n) = P_n(dx_1^n)\delta_{\phi_n(x_1^n)}(dy_1^n)$ for the joint distribution of (X_1^n, Y_1^n). Then $E_{W_n}[\rho_n(X_1^n, Y_1^n)] = D$, and by Jensen's inequality:

$$\begin{aligned}
\log M^n(C_n) &\geq \sum_{y_1^n \in C_n} Q_n(y_1^n) \log \frac{M^n(y_1^n)}{Q_n(y_1^n)} \\
&= \int dW_n(x_1^n, y_1^n) \log \frac{dW_n(x_1^n, y_1^n)}{d(P_n \times Q_n)} \\
&\quad + \sum_{y_1^n \in C_n} Q_n(y_1^n) \log M^n(y_1^n) \\
&= I(X_1^n; Y_1^n) + E_{Q_n}[\log M^n(Y_1^n)].
\end{aligned}$$

By the definition of $R_n(D)$, this is bounded below by $R_n(D)$. The second part follows immediately from the fact that $R_n(D) \geq nR(D)$, by Lemma 2 (ii). □

Before giving the proof of Theorem 4 we make some remarks on the methodology of the proof. The main technical step is established by an application of the Gärtner-Ellis theorem from large deviations. This is used to determine the asymptotics of the probability of distortion-balls. The same strategy has been applied by various authors in the recent literature in order to prove direct coding theorems; see, e.g., [19], [10] and the references therein, as well as the early work of Bucklew in [4]. The main difference here is that we are not only interested in the case of i.i.d. sources, and that the measures for which we need large deviations results are not product measures, making the application of the Gärtner-Ellis theorem more delicate. Finally we mention that in the random coding argument we employ, rather than generating a fixed number of codewords we generate infinitely many of them and look for the first "D-close match." This idea has already been used by [19] and [10], among others.

Proof of Theorem 4: The proof is given in 3 steps. First, for any $D > D_{\min}^{(1)}$ we construct sets C_n satisfying (i) and (iii) with $R_1(D)$ in place of $R(D)$. In the second step, assuming that $\mathbb{P}$ is ergodic in blocks, for each $D > D_{\min}$ we construct sets C_n satisfying (i) and (iii). In Step 3 we drop the assumption of the ergodicity in blocks, and for any $D > D_{\min}$ we construct sets C_n satisfying (i) and (ii).

A. Step 1:

Let $\mathbb{P}$ and $D > D_{\min}^{(1)}$ be fixed, and let an arbitrary $\epsilon > 0$ be given. By Lemma 2 we can choose a $D' \in (D_{\min}, D)$ such that $R_1(D') \leq R_1(D) + \epsilon/8$ and a probability measure Q^* on $(\hat{A}, \hat{\mathcal{A}})$ such that

$$\begin{aligned}
I^* + L^* &\stackrel{\triangle}{=} I_1(P_1, Q^*, D') + E_{Q^*}[\log M(Y)] \\
&\leq R_1(D) + \epsilon/8 \ \leq\ R_1(D) + \epsilon/4. \qquad (15)
\end{aligned}$$

Also we can pick a $W^* \in \mathcal{M}_1(P_1, Q^*, D')$ such that

$$H(W^* \| P_1 \times Q^*) \leq I^* + \epsilon/4. \qquad (16)$$

For $n \geq 1$, write Q_n^* for the product measure $(Q^*)^n$, and define

$$\mathcal{H}_n = \left\{ y_1^n \in \hat{A}^n\ :\ \frac{1}{n}\sum_{i=1}^n \log M(y_i) \leq L^* + \epsilon/4 \right\}.$$

Let $\widetilde{Q}_n$ be the measure Q_n^* conditioned on $\mathcal{H}_n$, $\widetilde{Q}_n(F) = Q_n^*(F \cap \mathcal{H}_n)/Q_n^*(\mathcal{H}_n)$, for $F \in \hat{\mathcal{A}}^n$. For each $n \geq 1$, let $\{Y(i) = (Y_1(i), Y_2(i), \ldots, Y_n(i))\ ;\ i \geq 1\}$ be i.i.d. random vectors $Y(i) \sim \widetilde{Q}_n$, and define

$$C_n = \{Y(i)\ :\ 1 \leq i \leq e^{n(I^*+\epsilon/2)}\}.$$

By the definition of $\mathcal{H}_n$, any $y_1^n \in \mathcal{G}_n$ has $M^n(y_1^n) \leq e^{n(L^*+\epsilon/4)}$, so by (15)

$$M^n(C_n) \leq e^{n(I^*+\epsilon/2)} e^{n(L^*+\epsilon/4)} \leq e^{n(R_1(D)+\epsilon)}$$

and (i) of the Theorem is satisfied with $R_1(D)$ in place of $R(D)$. Let X_1^n be a random vector with distribution P_n, and let i_n be the index of the first $Y(i)$ that matches X_1^n within ρ_n-distortion D. To verify (iii) we will show that

$$i_n \leq e^{n(I^*+\epsilon/2)} \quad \text{eventually, } \mathbb{P}\times\mathbb{Q} - \text{a.s.}$$

where $\mathbb{Q} = \prod_{n\geq 1}(\widetilde{Q}_n)^{\mathbb{N}}$, and this will follow from the following two statements:

$$\limsup_{n\to\infty} \frac{1}{n} \log \left[i_n\, \widetilde{Q}_n(B(X_1^n, D)) \right] \leq 0 \quad \mathbb{P}\times\mathbb{Q} - \text{a.s.} \quad (17)$$

$$\liminf_{n\to\infty} \frac{1}{n} \log \widetilde{Q}_n(B(X_1^n, D)) \geq -(I^* + \epsilon/4) \quad \mathbb{P} - \text{a.s.} \quad (18)$$

The proof of (17) follows easily from the observation that, conditional on X_1^n, the distribution of i_n is geometric with parameter $p = \widetilde{Q}_n(B(X_1^n, D))$; see, e.g., the derivation of (31) in [10].

To prove (18), first note that by the law of large numbers $Q_n^*(\mathcal{H}_n) \to 1$, as $n \to \infty$, so (18) is equivalent to

$$\liminf_{n\to\infty} \frac{1}{n} \log Q_n^* \left(B(X_1^n, D) \cap \mathcal{H}_n \right) \geq -(I^* + \epsilon/4), \quad (19)$$

$\mathbb{P}$−a.s. Let $Y_1, Y_2, \ldots$ be i.i.d. random variables with common distribution Q^*. For any realization x_1^∞ of $\mathbb{P}$, define the random vectors ξ_i and Z_n by

$$\xi_i = (\rho(x_i, Y_i), \log M(Y_i)), \qquad i \geq 1$$
$$Z_n = \frac{1}{n}\sum_{i=1}^n \xi_i, \qquad n \geq 1.$$

Also let $\Lambda_n(\lambda)$ be the log-moment generating function of Z_n,

$$\Lambda_n(\lambda) = \log E\left[e^{(\lambda, Z_n)} \right], \qquad \lambda = (\lambda_1, \lambda_2) \in \mathbb{R}^2,$$

where $(\cdot,\cdot)$ denotes the usual inner product in $\mathbb{R}^2$. Then for $\mathbb{P}$-almost any x_1^∞, by the ergodic theorem,

$$\begin{aligned} \frac{1}{n}\Lambda_n(n\lambda) &= \frac{1}{n}\log E\left[e^{\sum_{i=1}^n (\lambda, \xi_i)} \right] \\ &= \frac{1}{n}\sum_{i=1}^n \log E_{Q^*}\left[e^{\lambda_1\rho(x_i,Y)+\lambda_2 \log M(Y)} \right] \\ &\to E_{P_1}\left\{ \log E_{Q^*}\left[e^{\lambda_1\rho(X,Y)+\lambda_2 \log M(Y)} \right] \right\} \quad (20) \end{aligned}$$

where X and Y above are independent random variables with distributions P_1 and Q^*, respectively. Next we will need the following lemma. Its proof is a simple application of the dominated convergence theorem, using Jensen's inequality and the boundedness of ρ and $\log M$.

Lemma 3. For $k \geq 1$ and probability measures μ and ν on $(A^k, \mathcal{A}^k)$ and $(\hat{A}^k, \hat{\mathcal{A}}^k)$, respectively, define $\Lambda_{\mu,\nu}(\lambda)$ by

$$\int \log \left\{ \int [\exp(\lambda_1 \rho_k(x_1^k, y_1^k) + \lambda_2 \frac{1}{k} \log M^k(y_1^k))] d\nu(y_1^k) \right\} d\mu(x_1^k),$$

for $\lambda = (\lambda_1, \lambda_2) \in \mathbb{R}^2$. Then $\Lambda_{\mu,\nu}$ is convex, finite, and differentiable for all $\lambda \in \mathbb{R}^2$.

From Lemma 3 we have that the limiting expression in (20), which equals Λ_{P_1,Q^*}, is finite and differentiable everywhere. Therefore we can apply the Gärtner-Ellis theorem (Theorem 2.3.6 in [7]) to the sequence of random vectors Z_n, along $\mathbb{P}$-almost any x_1^∞, to get that

$$\liminf_{n\to\infty} \frac{1}{n} \log Q_n^* \left(B(x_1^n, D) \cap \mathcal{H}_n \right)$$

is equal to

$$\liminf_{n\to\infty} \frac{1}{n} \log \Pr(Z_n \in F) \geq -\inf_{z\in F} \Lambda^*(z) \quad \mathbb{P} - \text{a.s.} \quad (21)$$

where $F = \{z = (z_1, z_2) \in \mathbb{R}^2\ :\ z_1 < D,\ z_2 < L^* + \epsilon/4\}$ and

$$\Lambda^*_{P_1,Q^*}(z) = \sup_{\lambda\in\mathbb{R}^2} [(\lambda, z) - \Lambda_{P_1,Q^*}(\lambda)]$$

is the Fenchel-Legendre transform of $\Lambda_{P_1,Q^*}(\lambda)$. Recall our choice of W^* in (16). Then for any bounded measurable function $\phi : \hat{A} \to \mathbb{R}$ and any fixed $x \in A$,

$$H(W^*(\cdot|x) \| Q^*(\cdot)) \geq \int \phi(y) dW^*(y|x) - \log \int e^{\phi(y)} dQ^*(y)$$

(see, e.g., Lemma 6.2.13 in [7]). Fixing $x \in A$ and $\lambda \in \mathbb{R}^2$ for a moment, take $\phi(y) = \lambda_1 \rho(x, y) + \lambda_2 \log M(y)$, and integrate both sides $dP_1(x)$ to get that

$$H(W^* \| P_1 \times Q^*)$$

is bounded below by

$$\lambda_1 E_{W^*}(\rho) + \lambda_2 E_{Q^*}[\log M(Y)] - \Lambda_{P_1,Q^*}(\lambda).$$

Taking the supremum over all $\lambda \in \mathbb{R}^2$ and recalling (16) this becomes

$$I^* + \epsilon/4 \geq H(W^* \| P_1 \times Q^*) \geq \Lambda^*_{P_1,Q^*}(D^*, L^*)$$

where $D^* = \int \rho\, dW^* \leq D' < D$, so

$$I^* + \epsilon/4 \geq \inf_{z\in F} \Lambda^*_{P_1,Q^*}(z).$$

Combining this with the bound (21) yields (19) as required, and completes the proof of this step.

B. Step 2:

Assume $\mathbb{P}$ is ergodic in blocks, and let $\mathbb{P}$ and $D > D_{\min}$ be fixed and an arbitrary $\epsilon > 0$ be given. By Lemma 2 we can pick $k \geq 1$ large enough so that $D_{\min}^{(k)} < D$ and $(1/k)R_k(D) \leq R(D) + \epsilon/8$. This step consists of essentially repeating the argument of Step 1 along blocks of length k. Choose a $D' \in (D_{\min}^{(k)}, D)$ such that

$$\frac{1}{k}R_k(D') \leq \frac{1}{k}R_k(D) + \epsilon/16, \tag{22}$$

and a probability measure Q_k^* on $(\hat{A}^k, \hat{\mathcal{A}}^k)$ achieving

$$\begin{aligned} I_k^* + L_k^* &\stackrel{\triangle}{=} \frac{1}{k}I_k(P_k, Q_k^*, D') + \frac{1}{k}E_{Q_k^*}[\log M^k(Y_1^k)] \\ &\leq \frac{1}{k}R_k(D'), \end{aligned} \tag{23}$$

so that

$$I_k^* + L_k^* \leq R(D) + \epsilon/4. \tag{24}$$

Also pick a $W_k^* \in \mathcal{M}_k(P_k, Q_k^*, D')$ such that

$$\frac{1}{k}H(W_k^* \| P_k \times Q_k^*) \leq I_k^* + \epsilon/4. \tag{25}$$

For any $n \geq 1$ write $n = mk + r$ for integers $m \geq 0$ and $0 \leq r < k$, and define

$$\mathcal{H}_{n,k} = \left\{ y_1^n \in \hat{A}^n \ : \ \frac{1}{n}\sum_{i=1}^n \log M(y_i) \leq L_k^* + \epsilon/4 \right\}.$$

Write $Q_{n,k}^*$ for the measure

$$\left[\prod_{i=1}^m Q_k^*\right] \times [Q_k^*]_r,$$

where $[Q_k^*]_r$ denotes the restriction of Q_k^* to $(\hat{A}^r, \hat{\mathcal{A}}^r)$, and let $\widetilde{Q}_{n,k}$ be the measure $Q_{n,k}^*$ conditioned on $\mathcal{H}_{n,k}$. For each $n \geq 1$, let $\{Y(i) = (Y_1(i), Y_2(i), \ldots, Y_n(i)) \ ; \ i \geq 1\}$ be i.i.d. random vectors $Y(i) \sim \widetilde{Q}_n$, and let C_n consist of the first $e^{n(I_k^* + \epsilon/2)}$ of them. As before, by the definitions of $\mathcal{H}_{n,k}$ and C_n, and using (24), it easily follows that

$$\frac{1}{n}\log M^n(C_n) \leq R(D) + \epsilon$$

so (i) of the Theorem is satisfied. Let $Y_1, Y_2, \ldots, Y_n$ be distributed according to $Q_{n,k}^*$, and note that the random vectors $Y_{ik+1}^{(i+1)k}$ are i.i.d. with distribution Q_k^* (for $i = 0, 1, \ldots, m-1$). Therefore, as $n \to \infty$, by the law of large numbers we have that with probability 1,

$$\begin{aligned} &\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \log M(Y_i) \leq \\ &\left(\frac{m}{n}\right)\frac{1}{m}\sum_{i=0}^{m-1} \log M^k(Y_{ik+1}^{(i+1)k}) + \frac{kL_{\max}}{n} \to L_k^*. \end{aligned} \tag{26}$$

Following the same steps as before, to verify (iii) it suffices to show that

$$\liminf_{n\to\infty} \frac{1}{n}\log \widetilde{Q}_{n,k}(B(X_1^n, D)) \geq -(I_k^* + \epsilon/4) \quad \mathbb{P}-\text{a.s.}$$

and, in view of (26), this reduces to

$$\liminf_{n\to\infty} \frac{1}{n}\log Q_{n,k}^*\left(B(X_1^n, D) \cap \mathcal{H}_{n,k}\right) \geq -(I_k^* + \epsilon/4), \tag{27}$$

$\mathbb{P}-$a.s. For an arbitrary realization x_1^∞ from $\mathbb{P}$ and with Y_1^n as above, consider blocks of length k. For $i = 0, 1, \ldots, m-1$, we write

$$\widetilde{Y}_i^{(k)} = Y_{ik+1}^{(i+1)k} \quad \text{and} \quad \widetilde{x}_i^{(k)} = x_{ik+1}^{(i+1)k}$$

so that the probability $Q_{n,k}^*\left(B(X_1^n, D) \cap \mathcal{H}_{n,k}\right)$ can be written as

$$\begin{aligned} Q_{n,k}^* \Bigg\{ &\left(\frac{mk}{n}\right)\frac{1}{m}\sum_{i=0}^{m-1} \rho_k(\widetilde{Y}_i^{(k)}, \widetilde{x}_i^{(k)}) \\ &+ \frac{r}{n}\rho_r(Y_{n-r+1}^n, x_{n-r+1}^n) \leq D \\ \text{and} \quad &\left(\frac{mk}{n}\right)\frac{1}{m}\sum_{i=0}^{m-1}\frac{1}{k}\log M^k(\widetilde{Y}_i^{(k)}) \\ &+\frac{1}{n}\log M^r(Y_{n-r+1}^n) \leq L_k^* + \epsilon/4 \Bigg\}. \end{aligned}$$

Since we assume $\rho(x,y) \leq \rho_{\max}$ and $|\log M(y)| \leq L_{\max}$ for all $x \in A$, $y \in \hat{A}$, then for all n large enough (uniformly in x_1^∞) the above probability is bounded below by

$$\begin{aligned} (Q_k^*)^m \ \Bigg\{ &\frac{1}{m}\sum_{i=0}^{m-1} \rho_k(\widetilde{Y}_i^{(k)}, \widetilde{x}_i^{(k)}) \leq D' + \epsilon/8 \\ \text{and} \ &\frac{1}{m}\sum_{i=0}^{m-1}\frac{1}{k}\log M^k(\widetilde{Y}_i^{(k)}) \leq L_k^* + \epsilon/8 \Bigg\}. \end{aligned}$$

Now we are in the same situation as in the previous step, with the i.i.d. random variables $\widetilde{Y}_i^{(k)}$ in place of the Y_i, the ergodic process $\{\widetilde{X}_i^{(k)}\}$ in place of $\{X_i\}$, and $D' + \epsilon/8$ in place of D. Repeating the same argument as in Step 1 and invoking Lemma 3 and the Gärtner-Ellis theorem,

$$\begin{aligned} &\liminf_{n\to\infty} \frac{1}{n}\log Q_{n,k}^*\left(B(X_1^n, D) \cap \mathcal{H}_{n,k}\right) \geq \\ &\quad - \inf_{z_1 < D' + \epsilon/8,\ z_2 < L_k^* + \epsilon/8} \Lambda_k^*(z_1, z_2) \quad \mathbb{P}-\text{a.s.} \end{aligned} \tag{28}$$

where, in the notation of Lemma 3, $\Lambda_k^*(z)$ is the Fenchel-Legendre transform of $\Lambda_{P_k, Q_k^*}(\lambda)$. Recall our choice of W_k^* in (25) and write $D_k^* = \int \rho_k \, dW_k^* \leq D'$. Then by Lemma 6.2.13 from [7] together with (25) we get

$$I_k^* + \epsilon/4 \geq \frac{1}{k}H(W_k^* \| P_k \times Q_k^*) \geq \Lambda_k^*(D^*, L_k^*),$$

and this together with (28) proves (27), concluding this step.

C. Step 3:

In this part we invoke the ergodic decomposition theorem to remove the assumption that $\mathbb{P}$ is ergodic in blocks. Although similar to Berger's proof of the abstract coding theorem (see pp. 278-281 in [2]), the argument below is significantly more delicate. [In particular we need to avoid appealing to Perez's "generalized AEP" which subsequently turned out to be incorrect at that level of generality.]

As in Step 2, let $\mathbb{P}$ and $D > D_{\min}$ be fixed, and let an $\epsilon > 0$ be given. Pick $k \geq 1$ large enough so that $D_{\min}^{(k)} < D$ and $\frac{1}{k} R_k(D) \leq R(D) + \epsilon/8$, and pick $D' \in (D_{\min}^{(k)}, D)$ such that (22) holds. Also choose Q_k^* and W_k^* as in Step 2 so that (23), (24) and (25) all hold.

Let $\Omega = (A^k)^{\mathbb{N}}$, $\mathcal{F} = (\mathcal{A}^k)^{\mathbb{N}}$, and note that there is a natural 1-1 correspondence between sets in $F \in \mathcal{A}^{\mathbb{N}}$ and sets in $\widetilde{F} \in (\mathcal{A}^k)^{\mathbb{N}}$: Writing $\widetilde{x}_i = x_{ik+1}^{(i+1)k}$,

$$\widetilde{F} = \{\widetilde{x}_1^\infty \ : \ x_1^\infty \in F\}. \tag{29}$$

Let μ be the stationary measure on $(\Omega, \mathcal{F})$ describing the distribution of the "blocked" process $\{\widetilde{X}_i = X_{ik+1}^{(i+1)k} \ ; \ i \geq 0\}$, where, since k is fixed throughout the rest of the proof, we have dropped the superscript in $\widetilde{X}_i^{(k)}$. Although μ may not be ergodic, from the ergodic decomposition theorem we get the following information (see pp. 278-279 in [2]).

Lemma 4. There is an integer k' dividing k, and probability measures μ_i, $i = 0, 1, \ldots, k'-1$ on $(\Omega, \mathcal{F})$ with the following properties:

(i) $\mu = (1/k') \sum_{i=0}^{k'-1} \mu_i$.

(ii) Each μ_i is stationary and ergodic.

(iii) For each i, let $\mathbb{P}^{(i)}$ denote the measure on $(A^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}})$ induced by μ_i:

$$\mathbb{P}^{(i)}(F) = \mu_i(\widetilde{F}), \quad F \in \mathcal{A}^{\mathbb{N}}$$

[recall the notation of (29)]. Then $\mathbb{P} = (1/k') \sum_{i=0}^{k'-1} \mathbb{P}^{(i)}$, and each $\mathbb{P}^{(i)}$ is stationary in k'-blocks and ergodic in k'-blocks.

(iv) For each $0 \leq i \leq k'$ and $j \geq 0$, the distribution that $\mathbb{P}^{(i)}$ induces on the process $\{X_{j+n} \ ; \ n \geq 1\}$ is $\mathbb{P}^{(i+j \bmod k')}$.

For each $i = 0, 1, \ldots, k'-1$, let $\mu_{i,1}$ denote the first-order marginal of μ_i and write $R(D|i) = R_1(D; \mu_{i,1}, \widetilde{M})$ for the first-order rate function of the measure μ_i, with respect to the distortion measure ρ_k, and with mass function $\widetilde{M} = M^k$. Since W_k^* chosen as above has its A^k-marginal equal to P_k we can write it as $W_k^* = V_k^* \circ P_k$ where $V_k^*(\cdot | X_1^n)$ denote the regular conditional probability distributions. Write $P_k^{(i)}$ for the k-dimensional marginals of the measures $\mathbb{P}^{(i)}$, and define probability measures $W_k^{(i)}$ on $(A^n \times \hat{A}^n, \mathcal{A}^n \times \hat{\mathcal{A}}^n)$ by $W_k^{(i)} = V_k^* \circ P_k^{(i)}$. Let $D_i = \int \rho_k \, dW_k^{(i)}$ so that by Lemma 4 (iii),

$$\frac{1}{k'} \sum_{i=0}^{k'-1} D_i = \int \rho_k \, dW_k^* \leq D'. \tag{30}$$

Similarly, writing $Q_k^{(i)}$ for the $\hat{A}^k$-marginal of $W_k^{(i)}$ and applying Lemma 4 (iii),

$$\frac{1}{k'} \sum_{i=0}^{k'-1} \int \log M^k(y_1^k) \, dQ_k^{(i)}(y_1^k) \ = \ \int \log M^k(y_1^k) \, dQ_k^*(y_1^k) \tag{31}$$

and using the concavity of mutual information (Lemma 2),

$$\frac{1}{k'} \sum_{i=0}^{k'-1} H(W_k^{(i)} \| P_k^{(i)} \times Q_k^{(i)}) \leq H(W_k^* \| P_k \times Q_k^*). \tag{32}$$

For $N \geq 1$ large enough we can use the result of Step 2 to get N-dimensional sets B_i that almost-cover $(\hat{A}^k)^N$ with respect to μ_i. Specifically, consider N large enough so that

$$\frac{\max\{\rho_{\max}, \ L_{\max}, \ 1\}}{kN} < \min\{\epsilon/8, \ (D - D')/2\}. \tag{33}$$

For any such N, by the result of Step 2 we can choose sets $B_i \subset (\hat{A}^k)^N$ such that, for each i,

$$\mu_i\left([B_i]_{D_i}\right) \ \geq \ 1 - \epsilon_N, \quad \epsilon_N \to 0 \text{ as } N \to \infty, \tag{34}$$

$$\text{and} \quad \widetilde{M}^N(B_i) \ \leq \ \exp\{N(R(D_i|i) + \epsilon/8)\}. \tag{35}$$

Now choose and fix an arbitrary $y^* \in \hat{A}$, and for $n = k'(Nk+1)$ define new sets $B_i^* \subset \hat{A}^n$ by

$$B_i^* = \prod_{j=0}^{k'-1} \left[B_{i+j \bmod k'} \times \{y^*\}\right],$$

where $\prod$ denotes the cartesian product. Then, by (33), for any x_1^n, $\rho_n(x_1^n, B_i^*)$ is strictly less than

$$\frac{D - D'}{2} + \frac{1}{k'} \sum_{j=0}^{k'-1} \rho_{kN}\left(x_{j(kN+1)+1}^{j(kN+1)+kN}, B_{i+j \bmod k'}\right).$$

Note that each one of the blocks $x_{j(kN+1)+1}^{j(kN+1)+kN}$ above belongs to a different ergodic mode of the blocked process $\{\widetilde{X}_i\}$, explaining the role of the letters y^* in the construction of the new codebooks B_i^*. Now, by a simple union bound,

$$\begin{aligned} &\mathbb{P}^{(i)}([B_i^*]_D) \\ &\stackrel{(a)}{\geq} \quad 1 - \sum_{j=0}^{k'-1} \left[1 - \mathbb{P}^{(i+j \bmod k')}\left([B_{i+j \bmod k'}]_{D_i}\right)\right] \\ &\stackrel{(b)}{=} \quad 1 - \sum_{i=0}^{k'-1} \left[1 - \mu_i\left([B_i]_{D_i}\right)\ \right] \\ &\stackrel{(c)}{\geq} \quad 1 - k'\epsilon_N, \end{aligned} \tag{36}$$

where we used (30) in (a), Lemma 4 (iv) in (b), and (34) in (c). Also, using the definition of B_i^* and the bounds (33) and (35), $(1/n)\log M^n(B_i^*)$ is bounded above by

$$\frac{\log M(Y^*)}{kN+1} + \frac{1}{k'}\sum_{j=0}^{k'-1}\left[\frac{1}{kN}\log \widetilde{M}^N(B_{i+j \bmod k'})\right]$$
$$\leq \epsilon/8 + \frac{1}{k'}\sum_{j=0}^{k'-1}\left[\frac{1}{k}(R(D_j|j)+\epsilon/8)\right],$$

but from the definition of $R(D|j)$ and (32) and (31) this is

$$\leq \epsilon/4 + \frac{1}{k'}\sum_{j=0}^{k'-1}\left[\frac{1}{k}H(W_k^{(j)}\|P_k^{(j)}\times Q_k^{(j)}) + \frac{1}{k}\int \log M^k(y_1^k)\, dQ_k^{(j)}(y_1^k)\right].$$

Therefore,

$$\frac{1}{n}\log M^n(B_i^*) \leq I_k^* + L_k^* + \epsilon/2 \leq R(D) + 3\epsilon/4, \tag{37}$$

where the last inequality follows from (24). So in (36) and (37) we have shown that, for *all* $i = 0, 1, \ldots, k'-1$,

$$\mathbb{P}^{(i)}([B_i^*]_D) \geq 1 - k'\epsilon_N \quad \text{and} \tag{38}$$
$$\frac{1}{n}\log M^n(B_i^*) \leq R(D) + 3\epsilon/4. \tag{39}$$

Finally we define sets $C_n \subset \hat{A}^n$ by

$$C_n = \cup_{i=0}^{k'-1} B_i^*.$$

From the last two bounds above and (33), the sets C_n have

$$\frac{1}{n}\log M^n(C_n) \leq \frac{\log k'}{n} + R(D) + 3\epsilon/4 \leq R(D) + \epsilon,$$

and by Lemma 4 (*iii*), $P_n([C_n]_D)$ equals

$$\frac{1}{k'}\sum_{i=0}^{k'-1}\mathbb{P}^{(i)}([C_n]_D) \geq \frac{1}{k'}\sum_{i=0}^{k'-1}\mathbb{P}^{(i)}([B_i^*]_D) \geq 1 - \epsilon_n'$$

where $\epsilon_n' = k'\epsilon_N$ when $n = k'(Nk+1)$.

In short, we have shown that for any $D > D_{\min}$ and any $\epsilon > 0$, there exist (fixed) integers k, k' and N_0 such that: There is a sequence of sets C_n, for $n = k'(Nk+1)$, $N \geq N_0$, satisfying:

$$(1/n)\log M^n(C_n) \leq R(D) + \epsilon \ \text{ for all } n,$$
$$\text{and} \quad P_n([C_n]_D) \to 1 \ \text{ as } n \to \infty.$$

Since this is an asymptotic result, it is not hard to see that the restriction on n being of the form $n = k'(Nk+1)$ can be easily dropped to produce a sequence of sets $\{C_n \,;\, n \geq 1\}$ satisfying (*i*) and (*ii*) of Theorem 4. To see this, note that for intermediate values of the form $n' = k'(Nk+1) + s$ with $1 \leq s \leq kk'-1$ we can generate an efficient codebook $C_{n'}$ simply by adding an arbitrary block of length s, say $(y^*, y^*, \ldots, y^*) \in \hat{A}^s$, to the end of each codeword in C_n. Since the distortion measure ρ is bounded, the additional distortion achieved by the new codebook will be at most of order $1/n$, and this is asymptotically negligible. Similarly, since the number of codewords remains unchanged and the mass function M is bounded, the mass of each individual codeword will increase by no more than a constant factor in the exponent, and therefore the mass of the codebook codebook will increase by an amount that is at most of order $1/n$. □

Acknowledgment

The author wishes to thank Imre Csiszár and the anonymous referees for their useful comments, and also Amir Dembo for his comments on an earlier draft of this paper.

Appendix

Proof of Lemma 2: First recall that part (*iv*) is a well-known information theoretic fact; see, e.g., Corollary 5.5.5 in [8].

Since the sets $\mathcal{M}_n(P_n, Q_n, D)$ are increasing in D, $R_n(D)$ is nonincreasing in D. Next we claim that relative entropy is jointly convex in its two arguments. Let μ, ν be two probability measures over a Polish space $(S, \mathcal{S})$. In the case when μ and ν both consist of only a finite number of atoms, the joint convexity of $H(\mu\|\nu)$ is well-known (see, e.g., Theorem 2.7.2 in [5]). In general, $H(\mu\|\nu)$ can be written as

$$H(\mu\|\nu) = \sup_{\{E_i\}}\sum_i \mu(E_i)\log\frac{\mu(E_i)}{\nu(E_i)}$$

where the supremum is over all finite measurable partitions of S (see Theorem 2.4.1 in [15]). Therefore $H(\mu\|\nu)$ is the pointwise supremum of convex functions, hence itself convex. Combining the two infima, $R_n(D)$ can equivalently be written as the infimum of

$$H(W_n\|W_{n,X}\times W_{n,Y}) + E_{W_{n,Y}}[\log M^n(Y_1^n)] \tag{40}$$

over all $W_n \in \mathcal{M}_n(P_n, D)$, where

$$\mathcal{M}_n(P_n, D) = \cup_{Q_n}\mathcal{M}_n(P_n, Q_n, D).$$

Using this together with the joint convexity of relative entropy shows that $R_n(D)$ is convex. Since it is also nonincreasing and bounded away from $-\infty$, $R_n(D)$ is also continuous at all D except possibly at $D_{\min}^{(n)}$. This proves (*i*).

For part (*ii*) notice that if $R(D)$ exists for all D then it must also be nonincreasing and convex in $D \geq 0$ since $R_n(D)$ is; therefore, it must also be continuous except possibly at $D_{\min}$.

For part (*iii*), let $m, n \geq 1$ arbitrary, and let $W_m \in \mathcal{M}_m(P_m, D)$ and $W_n \in \mathcal{M}_n(P_n, D)$. Define a probability measure W_{m+n} on $(A^{m+n}\times\hat{A}^{m+n}\mathcal{A}^{m+n}\times\hat{\mathcal{A}}^{m+n})$ by

$$W_{m+n}(dx_1^{m+n}, dy_1^{m+n}) = W_m(dy_1^m|x_1^m)W_n(dy_{m+1}^{m+n}|x_{m+1}^{m+n})P(dx_1^{m+n}).$$

Notice that $W_{m+n} \in \mathcal{M}_{m+n}(P_{m+n}, D)$, and that, if (X_1^{m+n}, Y_1^{m+n}) are random vectors distributed according to W_{m+n}, then Y_1^m and Y_{m+1}^{m+n} are conditionally independent given X_1^{m+n}. Therefore, $R_{m+n}(D)$ is

$$\begin{aligned}
&\stackrel{(a)}{\leq} H(W_{m+n}\|W_{m+n,X}\times W_{m+n,Y}) \\
&\qquad + E_{W_{m+n,Y}}[\log M^{m+n}(Y_1^{m+n})] \\
&= I(X_1^{m+n};Y_1^{m+n}) + E_{W_{m+n,Y}}[\log M^{m+n}(Y_1^{m+n})] \\
&\stackrel{(b)}{\leq} I(X_1^m;Y_1^m) + I(X_{m+1}^{m+n};Y_{m+1}^{m+n}) \\
&\qquad + E_{W_{m,Y}}[\log M^m(Y_1^m)] + E_{W_{n,Y}}[\log M^n(Y_1^n)]
\end{aligned}$$

where (a) follows from (40) and (b) follows from the conditional independence of Y_1^m and Y_{m+1}^{m+n} given X_1^{m+n} (see, e.g., Lemma 9.4.2 in [8]). So we have shown that $R_{m+n}(D)$ is bounded above by

$$\begin{aligned}
&H(W_m\|W_{m,X}\times W_{m,Y}) + E_{W_{m,Y}}[\log M^m(Y_1^m)] \\
&\quad + H(W_n\|W_{n,X}\times W_{n,Y}) + E_{W_{n,Y}}[\log M^n(Y_1^n)],
\end{aligned}$$

and taking the infimum over all $W_m \in \mathcal{M}_m(P_m, D)$ and $W_n \in \mathcal{M}_n(P_n, D)$ yields

$$R_{m+n}(D) \leq R_m(D) + R_n(D). \tag{41}$$

[Note that in the above argument we implicitly assumed that we could find some $W_m \in \mathcal{M}_m(P_m, D)$ and a $W_n \in \mathcal{M}_n(P_n, D)$; if this was not the case, then either $R_m(D)$ or $R_n(D)$ would be equal to $+\infty$, and (41) would still trivially hold.] Therefore the sequence $\{R_n(D)\}$ is subadditive.

Next we claim that if $R_n(D) < \infty$ for some D, then $R_N(D) < \infty$ for all $N \geq n$. To see this first note that, by the boundedness of M we need only worry about the mutual information term in the definition of $R_n(D)$ in (13). Assuming $R_n(D) < \infty$ implies that there exist (X_1^n, Y_1^n) with $I(X_1^n;Y_1^n) < \infty$ and $E[\rho_n(X_1^n, Y_1^n)] \leq D$. In fact, by the convexity of mutual information in the conditional distributions (part (iv) of this Lemma) we can restrict ourselves to stationary vectors (X_1^n, Y_1^n). Based on (X_1^n, Y_1^n) we define (X_1^{n+1}, Y_1^{n+1}) as follows: Let X_1^{n+1} have the source distribution, and, given X_1^{n+1}, define two conditionally independent random vectors Y_1^n and $\tilde{Y}_2^{n+1}$ so that Y_1^n has the same distribution as before, and $\tilde{Y}_2^{n+1}$ has the same distribution given X_2^{n+1} as Y_1^n given X_1^n. Let $Y_{n+1} = \tilde{Y}_{n+1}$. Then by the chain rule for mutual information we have that $I(X_1^{n+1};Y_1^{n+1}) = I(X_1^n;Y_1^n) + I(X_1^{n+1};\tilde{Y}_{n+1}|Y_1^n) \leq I(X_1^n;Y_1^n) + I(X_1^{n+1};\tilde{Y}_2^{n+1}) \leq 2I(X_1^n;Y_1^n)$. Therefore $I(X_1^{n+1};Y_1^{n+1}) < \infty$, and by stationarity $E[\rho_{n+1}(X_1^{n+1}, Y_1^{n+1})] \leq D$. This implies that $D_{\min}^{(n)}$ is nonincreasing in n, so it follows that $D_{\min}$, whenever defined is equal to $\inf_n D_{\min}^{(n)}$ as claimed. Finally, subadditivity and the fact that $D_{\min}^{(n)}$ is nonincreasing in n imply that $\lim_n (1/n)R_n(D) = \inf_n (1/n)R_n(D)$ for all $D \geq 0$. □

References

[1] R.R. Bahadur, *Some Limit Theorems in Statistics*, SIAM, Philadelphia, PA, 1971.

[2] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice-Hall Inc., Englewood Cliffs, NJ, 1971.

[3] R.E. Blahut, "Hypothesis testing and information theory," *IEEE Trans. Inform. Theory*, vol. 20, no. 4, pp. 405–417, 1974.

[4] J.A. Bucklew, "A large deviation theory proof of the abstract alphabet source coding theorem," *IEEE Trans. Inform. Theory*, vol. 34, no. 5, pp. 1081–1083, 1988.

[5] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, J. Wiley, New York, 1991.

[6] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.

[7] A. Dembo and O. Zeitouni, *Large Deviations Techniques And Applications*, Springer-Verlag, New York, second edition, 1998.

[8] R.M. Gray, *Entropy and Information Theory*, Springer-Verlag, New York, 1990.

[9] L.H. Harper, "Optimal numberings and isoperimetric problems on graphs," *J. Combinatorial Theory*, vol. 1, pp. 385–393, 1966.

[10] I. Kontoyiannis, "Pointwise redundancy in lossy data compression and universal lossy data compression," *IEEE Trans. Inform. Theory*, vol. 46, no. 1, pp. 136–152, January 2000.

[11] I. Kontoyiannis, "Efficient sphere-covering and converse measure concentration via generalized coding theorems," Tech. Rep. 99-24, Department of Statistics, Purdue University, October 1999, [Available at www.stat.purdue.edu/people/yiannis].

[12] C. McDiarmid, "On the method of bounded differences," in *Surveys in combinatorics (Norwich, 1989)*. London Math. Soc. Lecture Note Ser., 141, 1989, pp. 148–188, Cambridge Univ. Press, Cambridge.

[13] C. McDiarmid, "Concentration," in *Probabilistic methods for algorithmic discrete mathematics*. Algorithms Combin., 16, 1998, pp. 195–248, Springer, Berlin.

[14] K. Petersen, *Ergodic Theory*, Cambridge University Press, Cambridge, 1983.

[15] M.S. Pinsker, *Information and Information Stability of Random Variables and Processes*, Holden-Day, San Fransisco, 1964.

[16] C.E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec.*, vol. part 4, pp. 142–163, 1959, Reprinted in D. Slepian (ed.), *Key Papers in the Development of Information Theory*, IEEE Press, 1974.

[17] V. Strassen, "Asymptotische Abschätzungen in Shannons Informationstheorie," in *Trans. Third Prague Conf. Information Theory, Statist. Decision Functions, Random Processes (Liblice, 1962)*, pp. 689–723. Publ. House Czech. Acad. Sci., Prague, 1964.

[18] M. Talagrand, "Concentration of measure and isoperimetric inequalities in product spaces," *Inst. Hautes Études Sci. Publ. Math.*, vol. No. 81, pp. 73–205, 1995.

[19] E.-h. Yang and Z. Zhang, "On the redundancy of lossy source coding with abstract alphabets," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1092-1110, 1999.

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 in electrical engineering, both from Stanford University. Between June and December 1995 he worked at IBM Research, on a satellite image processing and compression project, funded by NASA and IBM. He has been 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) since 1998. During the 2000-01 academic year he is visiting the Applied Mathematics Division of Brown University. His research interests include data compression, applied probability, statistical genetics, nonparametric statistics, entropy theory of stationary processes and random fields, and ergodic theory.