ISRAEL JOURNAL OF MATHEMATICS 86 (1994), 331–348

THE POSITIVE-DIVERGENCE AND BLOWING-UP PROPERTIES

BY

KATALIN MARTON*

Mathematics Institute, Hungarian Academy of Sciences

AND

PAUL C. SHIELDS**

University of Toledo and Eötvös Loránd University

ABSTRACT

A property of ergodic finite-alphabet processes, called the blowing-up property, is shown to imply exponential rates of convergence for frequencies and entropy, which in turn imply a positive-divergence property. Furthermore, processes with the blowing-up property are finitely determined and the finitely determined property plus exponential rates of convergence for frequencies and for entropy implies blowing-up. It is also shown that finitary codings of i.i.d. processes have the blowing-up property.

1. Introduction

Our original motivation for the research reported in this paper was to study properties of an ergodic process Q that would guarantee that if P is ergodic and $P \neq Q$ then the limiting divergence rate $D(P\|Q)$ must be positive. (Definitions will be given later.) It was fairly easy to show that i.i.d. processes and aperiodic Markov chains have such a positive-divergence property. We were soon led, however, to the realization that some condition on Q is needed, for the second author was able to show that not every ergodic process has the positive-divergence property; in fact, even requiring that the process be a stationary coding of an i.i.d.

* Partially supported by Hungarian National Foundation for Scientific Research Grant OTKA 1906.

** Partially supported by NSF grant DMS-9024240.

Received June 29, 1992 and in revised form April 20, 1993

process is not enough. Counterexamples for the positive-divergence property and other divergence-rate questions are discussed in a separate paper, [9].

In this paper we show that if Q has exponential rates of convergence for frequencies and entropy then it has a stronger form of the positive-divergence property, namely, if $D(P\|Q)$ is close enough to 0 then P will be close to Q in both distribution and entropy (Theorem 1). As is well known, however, closeness in distribution and entropy forces closeness in the $\bar{d}$-metric, provided Q is a finitely determined process in the sense of Ornstein, [5]. Thus our Theorem 1 leads to the problem of characterizing those finitely determined processes that have exponential rates of convergence for frequencies and entropy. We were surprised to discover that a property, called the **blowing-up property**, which had earlier been shown to hold for i.i.d. processes, [3, Lemma 5.4] and [4], is in fact equivalent to the three conditions of having exponential rates of convergence for frequencies, of having exponential rates of convergence for entropy, and of having the finitely determined property (Theorem 2). Part of the proof of this result was supplied to us by Ornstein and Weiss.

Finally, we turn to the question of finding classes that have the blowing-up property, that is, extending the i.i.d. results of [3, 4]. We were able to show that a class of processes, which we call the finitary processes, have the blowing-up property (Theorem 3). Such processes are stationary codings of i.i.d. processes in which the window-width function is a finite-valued random variable. The finitary class is known to include the aperiodic Markov chains, [2, 11].

Notation, definitions, and precise statements of our results will be given in the next section; proofs will be given in Section 3.

ACKNOWLEDGEMENT: The authors wish to thank Imre Csiszár for fruitful conversations on the topics in this paper, and to Don Ornstein and Ben Weiss for their contribution of the proof that blowing-up implies finitely determined.

2. Definitions and statements of results

For our purposes a process is a shift-invariant Borel probability measure on the space A^∞ of sequences $x = \{x_n\}$, $x_i \in A$, drawn from a finite set A, which is called the alphabet. Thus, in particular, for us "process" means "stationary process;" we will, in fact, be mostly concerned with ergodic processes. In some cases, it will be convenient to think of a process as a shift-invariant measure on the space A^Z of doubly-infinite sequences.

Letters, such as P, Q, or R, will denote processes, while μ and ν will denote other probability measures. The shift $T = T_A$ is the transformation defined by $(Tx)_n = x_{n+1}$, $x \in A^\infty$. The sequence $a_m, a_{m+1}, \ldots, a_n$ will be denoted by a_m^n, the set of such a_m^n will be denoted by A_m^n, and A^n will denote A_1^n. The process P defines a measure P_k on A^k by the formula

$$P_k(a_1^k) = P\left(\{x\colon x_i = a_i,\ 1 \le i \le k\}\right).$$

If k is understood we use P instead of P_k. Note that a process is i.i.d. if and only if each P_k is the product measure defined by P_1.

Let μ and ν be probability measures on the finite set B. The divergence of μ with respect to ν is defined by

$$D(\mu\|\nu) = \sum_b \mu(b) \log \frac{\mu(b)}{\nu(b)},$$

where base 2 logarithms are used here and elsewhere in this paper. The divergence plays a role in coding theory; for example $D(P_k\|Q_k)$ is, essentially, the P_k-expected additional cost in using a k-block code that is optimal for Q_k, rather than the P_k-optimal code. For processes P and Q we define an upper and a lower divergence-rate as follows.

Definition 1: The **upper** and **lower divergence rates** of P with respect to Q are defined respectively by

$$D^*(P\|Q) = \limsup_n \frac{1}{n} D(P_n\|Q_n)$$

and

$$D_*(P\|Q) = \liminf_n \frac{1}{n} D(P_n\|Q_n). \qquad \blacksquare$$

For i.i.d. processes, the upper and lower rates are the same and are equal to $D(P_1\|Q_1)$. In this case the divergence-rate is an important concept in testing whether observed data comes from the null distribution P_1 or the alternative distribution Q_1. The divergence measures the exponential rate of convergence of the type II error, given a fixed type I error, that is,

$$D(P\|Q) = -\lim_{k\to\infty} \frac{1}{k} \log \min_{\substack{B \subset A^k \\ P_k(B) \ge 1-\epsilon}} Q_k(B).$$

The divergence-rate also plays an important role in large deviations theory. The upper and lower rates are equal for ergodic Markov chains where again the divergence-rate is important in hypothesis testing and large deviations theory. It is not hard to construct nonergodic processes for which the upper and lower rates are not equal, but such pathology can also happen for ergodic processes, [9].

The divergence $D(P_k\|Q_k)$ can be equal to 0 only if the two measures P_k and Q_k agree on all k-sequences. The divergence-rate for processes involves a limit and hence the situation is more complicated; in fact, the upper rate $D^*(P\|Q)$ can be 0 without P and Q being the same, even if P and Q are both ergodic, [9]. Let us say that an ergodic process Q has the **positive-divergence property** if $D_*(P\|Q) > 0$ for any ergodic process P which is not equal to Q.

Our first goal will be to show that a large class of ergodic processes Q do have the positive-divergence property. We will show that processes with exponential rates of convergence for frequencies have the positive-divergence property; in fact, for such processes the lower divergence-rate $D_*(P\|Q)$ has continuity properties at $P = Q$. Exponential rates of convergence for entropy yield comparable results for entropy. The rate of convergence concepts will now be defined after which we state the results.

The variational distance is defined by

$$|P_k - Q_k| = \sum_{a_1^k} |P_k(a_1^k) - Q_k(a_1^k)|,$$

and the entropy-rate is defined by

$$H(P) = \lim_n \frac{1}{n} H(P_n),$$

where $H(P_n) = -\sum_{a_1^n} P(a_1^n) \log P(a_1^n)$. The **empirical distribution of k-blocks** $\hat{\Delta}_{x_m^n,k}$ in x_m^n is the distribution on A^k defined by the (relative) frequencies of overlapping k-blocks, that is,

$$\hat{\Delta}_{x_m^n,k}(a_1^k) = \frac{|\{i \in [m, n-k+1] \colon x_i^{i+k-1} = a_1^k\}|}{n-k-m+2},$$

where $|\cdot|$ denotes cardinality.

The exponential rate concepts we will use are defined as follows.

Definition 2: An ergodic process Q has the **exponential rate of convergence property for frequencies** if given k and $\epsilon > 0$ there is a $\delta > 0$ and an N such that

$$Q\left(\{x_1^n : |\hat{\Delta}_{x_1^n,k} - Q_k| \geq \epsilon\}\right) \leq 2^{-n\delta},\ n \geq N. \quad \blacksquare$$

Definition 3: An ergodic process Q has the **exponential rate of convergence property for entropy** if given $\epsilon > 0$ there is a $\delta > 0$ and an N such that

$$Q\left(\{x_1^n : |-\log Q(x_1^n) - H(Q)| > \epsilon\}\right) \leq 2^{-n\delta},\ n \geq N. \quad \blacksquare$$

Our first theorem can now be stated.

THEOREM 1: *Let Q be an ergodic process.*

(i) *If Q has the exponential rate of convergence property for frequencies, then given $\epsilon > 0$ and a positive integer k there is a $\delta > 0$ such that if P is ergodic and $D_*(P\|Q) < \delta$ then $|P_k - Q_k| < \epsilon$.*

(ii) *If Q has the exponential rate of convergence property for entropy, then given $\epsilon > 0$, there is a $\delta > 0$ such that if P is an ergodic process and $D_*(P\|Q) < \delta$ then $|H(P) - H(Q)| < \epsilon$.*

Note, in particular, that processes with exponential rates of convergence for frequencies have the positive-divergence property.

It is well known that i.i.d. processes and ergodic Markov chains have the exponential rate of convergence property for entropy and for frequencies. Closeness in distribution and entropy is related to the concept of finitely determined process, a concept first introduced by Ornstein to characterize those processes that are stationary codings of i.i.d. processes, [5]. A process Q is **finitely determined** if given $\epsilon > 0$ there is a $\delta > 0$ and a positive integer k such that if P is any ergodic process such that $|P_k - Q_k| < \delta$ and $|H(P) - H(Q)| < \delta$ then $\bar{d}(P, Q) < \epsilon$. Here $\bar{d}(P, Q)$ denotes the $\bar{d}$-distance between the processes P and Q, defined in the next paragraph.

The distance between two n-sequences is the average Hamming distance, defined by

$$d_n(a_1^n, b_1^n) = \frac{1}{n}\sum_{i=1}^{n} d(a_i, b_i),$$

where $d(a,b)$ is 0 or 1, depending on whether $a = b$ or $a \neq b$. The $\bar{d}_n$-distance between two measures, μ and ν, on A^n is defined by

$$\bar{d}_n(\mu,\nu) = \inf_{\lambda \in J_n(\mu,\nu)} E_\lambda(d_n(x_1^n, y_1^n)),$$

where $J_n(\mu,\nu)$ denotes the set of all measures λ on $A^n \times A^n$ that have μ and ν as marginals, and E_λ denotes expectation with respect to λ. The $\bar{d}$-distance between two processes P and Q with alphabet A is defined by

$$\bar{d}(P,Q) = \lim_n \bar{d}_n(P_n, Q_n),$$

a limit that can be shown to exist, at least for stationary processes. The reader is referred to [5] for a discussion of the $\bar{d}$-metric.

Returning to our statement of results, we note that the following corollary to Theorem 1 is immediate.

COROLLARY 1: *If Q is a finitely determined process and Q has exponential rates of convergence for both frequencies and entropy then given $\epsilon > 0$ there is a $\delta > 0$ such that if P is ergodic and $D_*(P\|Q) < \delta$ then $\bar{d}(P,Q) < \epsilon$.*

We next introduce the blow-up idea. If $C \subseteq A^n$ then $[C]_\epsilon$ will denote the ϵ-neighborhood about C, that is

$$[C]_\epsilon = \{b_1^n \colon d_n(a_1^n, b_1^n) \le \epsilon, \text{ for some } a_1^n \in C\}.$$

The set $[C]_\epsilon$ will also be called the **ϵ-blow-up** of C.

This leads us to our key concept, the blowing-up property.

Definition 4: An ergodic process Q has the **blowing-up property** if given $\epsilon > 0$ there is a $\delta > 0$ and an N such that if $n \ge N$ and $C \subseteq A^n$ then

$$Q(C) \ge 2^{-n\delta} \implies Q([C]_\epsilon) \ge 1 - \epsilon. \quad \blacksquare \tag{1}$$

As noted in our introduction it has been shown that i.i.d. processes have the blowing-up property, [3, 4]. Our second principal result is the following characterization of processes with the blowing-up property.

THEOREM 2: *A process Q has the blowing-up property if and only if it is finitely determined and has the exponential rate of convergence property for both frequencies and entropy.*

As we will show in our proof of this theorem, the blowing-up property is essentially just a stronger version of a property called "extremality", a property that is known to be equivalent to the finitely determined property, [8, Lemma 1]. The proof that blowing-up implies extremality was supplied to us by Ornstein and Weiss.

The finitely determined processes are precisely the stationary codings of the i.i.d. processes. A process Q with alphabet A is a **stationary coding** of a process P with alphabet B if there is a measurable function F: $B^Z \mapsto A^Z$, called the encoder, such that $T_A F(x) = F(T_B x)$, $x \in B^Z$ and $Q = PT^{-1}$. Stationary codings of i.i.d. processes are known by many different names, such as B-processes, very weak Bernoulli processes, finitely determined processes, and almost block independent processes; each of which corresponds to a different characterization of such processes, [5, 10].

In the light of Theorem 2 it would be nice to have conditions on the encoder F which guarantee that the image of an i.i.d. process has the blowing-up property. As a step in this direction, we can show that finitary codings of i.i.d. processes have the blowing-up property. Finitary codings are defined as follows. Suppose Q has alphabet A, P has alphabet B, and Q is a stationary coding of P with encoder F. The function f: $B^Z \mapsto A$ defined by $f(x) = (F(x))_0$ is called the sliding-window encoder. The encoder F can be reconstructed from f by noting that $y = F(x)$ if and only if $y_n = f(T^n x)$. The coding is said to be **finitary** if there is a nonnegative integer-valued measurable function $w(x)$, called the window-width function, such that

$$x_{-w(x)}^{w(x)} = \tilde{x}_{-w(x)}^{w(x)} \Rightarrow f(x) = f(\tilde{x}), \text{ a.e.}$$

Definition 5: Q is a **finitary** process if it is a finitary coding of an i.i.d. process. ∎

Our principal result for finitary coding is the following.

THEOREM 3: *Finitary coding preserves the blowing-up property. In particular, a finitary process has the blowing-up property.*

Many processes are known to be finitary codings of i.i.d. processes, including the following.

1. The aperiodic Markov chains, [2, 11].
2. The m-dependent processes, [12].
3. The indecomposable finite-state processes.

The third item, the indecomposable finite-state processes, is included for they are just the finite codings of aperiodic Markov chains, that is, finitary codings in which the window width $w(x)$ is uniformly bounded. In summary, our results imply that all processes listed above have the blowing-up property, the exponential-rates property, and the positive-divergence property.

The blowing-up property and the finitary property are, of course, much stronger than the exponential-rates property. Thus, for example, the following processe have the exponential-rate properties and hence the positive-divergence property, but are not finitely determined, hence have neither the finitary property nor the blowing-up property.

1. The periodic, irreducible Markov chains.
2. The periodic, ergodic finite-state processes.
3. The rotation processes defined by irrational rotations of the circle together with partitions of the circle into intervals.

The following chain of implications connecting the finitary property (FP), the blowing-up property (BUP), the exponential rates properties (ERP), and the positive-divergence property (PDP), will be established in this paper.

$$\text{FP} \Rightarrow \text{BUP} \Rightarrow \text{ERP} \Rightarrow \text{PDP}.$$

The reverse implication ERP $\Rightarrow$ BUP does not hold (since periodic chains have ERP but not BUP). We suspect that the other reverse implications, BUP $\Rightarrow$ FP, and, PDP $\Rightarrow$ ERP, also fail to hold, although we have no proofs.

An example of a finitely determined process that does not have the positive-divergence property is constructed in [9]; in particular, there are finitely determined processes that do not have the blowing-up property. Thus blowing-up is a stronger property than finitely determined. A connection between blowing-up and other properties that are stronger than finitely determined, such as weak Bernoulli, seems unlikely, although this has not been explored.

3. Proofs

3.1 EXPONENTIAL RATES. The basic theorem connecting exponential rates of convergence for frequencies and entropy with the positive-divergence property is stated as follows.

THEOREM 1: *Let Q be an ergodic process.*

(i) *If Q has the exponential rate of convergence property for frequencies, then given $\epsilon > 0$ and a positive integer k there is a $\delta > 0$ such that if P is ergodic and $D_*(P\|Q) < \delta$ then $|P_k - Q_k| < \epsilon$.*

(ii) *If Q has the exponential rate of convergence property for entropy, then given $\epsilon > 0$, there is a $\delta > 0$ such that if P is an ergodic process and $D_*(P\|Q) < \delta$ then $|H(P) - H(Q)| < \epsilon$.*

Our proof will make use of a simple inequality, [7, Theorem 2.4.2], stated here as the following lemma.

LEMMA 1: *Let μ and ν be probability measures on the finite set A. Then*

$$\sum_a \mu(a) \left|\log \frac{\mu(a)}{\nu(a)}\right| \le D(\mu\|\nu) + 2\log 2.$$

Proof: Let A^- denote the set of $a \in A$ such that $\log(\mu(a)/\nu(a)) < 0$. The convexity of the logarithm function then gives

$$\begin{aligned}\sum_{a\in A^-} \mu(a) \left|\log \frac{\mu(a)}{\nu(a)}\right| &= \mu(A^-) \sum_{a\in A^-} \frac{\mu(a)}{\mu(A^-)} \log \frac{\nu(a)}{\mu(a)} \\ &\le \mu(A^-) \log \left[\frac{\nu(A^-)}{\mu(A^-)}\right] \\ &\le \mu(A^-) \log \frac{1}{\mu(A^-)} \le \log 2,\end{aligned}$$

from which the lemma easily follows. ∎

Proof of Theorem 1: Let α be a positive number to be specified later and suppose $D(P_n\|Q_n) < \alpha^2/2$. If n is large enough then the bound of Lemma 1 together with the Markov inequality implies that there is a set $B \subseteq A^n$ such that

(a) $P_n(B) > 1 - \alpha$.

(b) $2^{-\alpha n} P_n(x_1^n) \le Q_n(x_1^n) \le 2^{\alpha n} P_n(x_1^n)$, $x_1^n \in B$.

If k is fixed and n is large enough then the ergodic and entropy theorems applied to P tell us that there is a set $\tilde{B} \subseteq B$ such that

(c) $P_n(\tilde{B}) > 1 - 2\alpha$.

(d) $|\hat{\Delta}_{x_1^n,k} - P_k| < \alpha,\ x_1^n \in \tilde{B}$.

(e) $2^{-(H(P)+\alpha)n} \le P(x_1^n) \le 2^{-(H(P)-\alpha)n},\ x_1^n \in \tilde{B}$.

If Q has the exponential rate of convergence property for frequencies then

$$Q_n\left(\{x_1^n\colon |\hat{\Delta}_{x_1^n,k} - Q_k| \ge \alpha\}\right) \le 2^{-n\gamma},$$

for some $\gamma > 0$ and all n sufficiently large. Note, however, that conditions (b) and (c) imply that $Q_n(\tilde{B}) \ge (1-2\alpha)2^{-\alpha n}$. Thus, if α is small enough and n large enough then there will be an $x_1^n \in \tilde{B}$ such that $|\hat{\Delta}_{x_1^n,k} - Q_k| \le \alpha$. Since (d) also holds for this same x_1^n we conclude that $|P_k - Q_k| \le 2\alpha$. This establishes (i).

If Q has the exponential rate of convergence property for entropy, then, as in the preceding argument, we can assume that n is sufficiently large and α sufficiently small that there will be an $x_1^n \in \tilde{B}$ such that

$$2^{-(H(Q)+\alpha)n} \le Q(x_1^n) \le 2^{-(H(Q)-\alpha)n}.$$

This combined with (b) and (e) then yields $|H(P)-H(Q)| \le 4\alpha$ which establishes (ii), completing the proof of Theorem 1. ∎

3.2 Blowing-up and extremality. In this subsection we establish the connection between blowing-up, exponential rates of convergence, and finitely determined, which we summarize as the following theorem.

Theorem 2: *A process Q has the blowing-up property if and only if it is finitely determined and has the exponential rate of convergence property for frequencies and for entropy.*

To improve readability we break the proof into three parts, showing first that blowing-up implies exponential rates, then that blowing-up implies finitely determined, and finally that finitely determined plus exponential rates implies blowing-up.

3.2.1 Blowing-up implies exponential rates. Suppose Q has the blowing-up property. We first prove that Q has the exponential rate of convergence property for frequencies. The idea of the proof is that if the set of sequences with bad frequencies does not have exponentially small measure then it can be blown up by a small amount to get a set of large measure. If the amount of blow-up is small enough, however, then frequencies won't change much and hence we would have a set of large measure all of whose members have bad frequencies, contradicting the ergodic theorem.

The details of the frequency proof will now be given. Define $B(n,k,\epsilon) = \{x_1^n : |\hat{\Delta}_{x_1^n,k} - Q_k| \geq \epsilon\}$. Choose $\gamma = \epsilon/2k$ and note that $d_n(x_1^n, y_1^n) < \gamma$ implies that $|\hat{\Delta}_{x_1^n,k} - \hat{\Delta}_{y_1^n,k}| < \epsilon/2$. In particular, note that

$$(2) \qquad [B(n,k,\epsilon)]_\gamma \subseteq B(n,k,\epsilon/2).$$

Next use the blowing-up property to choose δ and N so that if $n \geq N$, $C \subset A^n$, and $Q(C) \geq 2^{-\delta n}$ then $Q([C]_\gamma) \geq 1-\epsilon$. Thus if $n \geq N$ and $Q_n(B(n,k,\epsilon)) \geq 2^{-\delta n}$ then we would have $Q_n([B(n,k,\epsilon)]_\gamma) \geq 1-\epsilon$ which combines with (2) to force $Q_n(B(n,k,\epsilon/2)) \geq 1-\epsilon$, which cannot be true for all large n since the ergodic theorem guarantees that $\lim_n Q_n(B(n,k,\epsilon/2)) = 0$ for each k.

Next we turn to the proof that blowing-up implies an exponential rate of convergence for entropy. One part of this is easy, for there cannot be too many sequences whose measure is too large, hence a small blow-up of such a set cannot possibly produce enough sequences to cover a large fraction of the measure. To make this precise, define

$$B^*(n,\epsilon) = \{x_1^n : Q_n(x_1^n) \geq 2^{-n(H-\epsilon)}\},$$

and note that $|B^*(n,\epsilon)| \leq 2^{n(H-\epsilon)}$. Thus we can choose $\alpha > 0$ so that

$$|[B^*(n,\epsilon)]_\alpha| \leq 2^{n(H-\epsilon/2)},$$

and hence $\lim_n Q_n([B^*(n,\epsilon)]_\alpha) = 0$. In particular, if

$$Q(C) \geq 2^{-\delta n} \Longrightarrow Q([C]_\alpha) \geq 1-\alpha,\ n \geq N,$$

then we must have $Q(B^*(n,\epsilon)) < 2^{-\delta n}$, $n \geq N$.

An exponential bound for the measure of the set

$$B_*(n,\epsilon) = \{x_1^n : Q_n(x_1^n) \leq 2^{-n(H+\epsilon)}\}$$

of sequences of too-small probability is a bit trickier to obtain and will make use of the exponential rate of convergence of frequencies. The idea is that when n is sufficiently large then, except for a set of exponentially small probability, most of x_1^n will be covered by k-blocks whose measure is about 2^{-kH}. This gives an exponential bound on the number of such x_1^n, which in turn means that it is exponentially very unlikely that such an x_1^n can have probability much smaller than 2^{-nH}. The details of this argument follow.

First use the Shannon–McMillan theorem to choose k so large that

$$Q_k(B_*(k, \epsilon/4)) < \alpha,$$

where α will be specified in a moment. For $n \geq N$ put

$$T_n = \{x_1^n \colon \hat{\Delta}_{x_1^n,k}(B_*(k, \epsilon/4)) < 2\alpha\}.$$

A standard argument, for example, see [6, p. 912], then shows that there is an $\alpha > 0$ and an N such that

$$|T_n| \leq 2^{n(H+\epsilon/2)}, \; n \geq N. \tag{3}$$

Thus we have $Q_n(T_n \cap B_*(n, \epsilon)) \leq 2^{-n\epsilon/2}$. But we have already shown that there is an exponential rate of convergence for frequencies, so we can choose $\delta > 0$ so that $Q_n(T_n) \geq 1 - 2^{-\delta n}$. Thus

$$Q_n(B_*(n, \epsilon)) \leq 2^{-\delta n} + 2^{-n\epsilon/2},$$

which gives the desired exponential bound.

3.2.2 Blowing-up implies finitely determined. The key idea here is that the blowing-up property is essentially just a stronger version of one form of the "extremality" property, a property known to be equivalent to the finitely determined property, [8, Lemma 1]. A process Q has the **extremality property** if given $\epsilon > 0$ there is an N and a $\delta > 0$ such that if $n \geq N$ and $\mathcal{C}$ is any partition of A^n such that $Q_n(C) \geq 2^{-\delta n}$, $C \in \mathcal{C}$ then,

$$\bar{d}_n(Q_n, Q_n(\cdot|C)) \leq \epsilon,$$

except for a subcollection of $\mathcal{C}$ of total Q_n measure at most ϵ, where $Q_n(\cdot|C)$ is the conditional measure on $C \subset A^n$, defined by $Q_n(B|C) = Q_n(B \cap C)/Q_n(C)$. Extremality allows the possibility that the conditional measure $Q_n(\cdot|C)$ may not be close to the unconditioned measure Q_n in $\bar{d}_n$, for some members $C \in \mathcal{C}$, so long as such "bad" sets have small total probability. The following lemma, whose proof was supplied to us by Ornstein and Weiss, asserts that for processes with the blowing-up property, *all* the measures $Q_n(\cdot|C)$ will be close to Q_n in $\bar{d}_n$. Thus blowing-up implies extremality, which in turn implies finitely determined.

LEMMA 2: *If Q has the blowing-up property then given $\epsilon > 0$ there is an N and a $\delta > 0$ such that if $n \geq N$ and $C \subset A^n$ satisfies $Q_n(C) \geq 2^{-\delta n}$ then* $\bar{d}_n(Q_n, Q_n(\cdot|C)) \leq \epsilon$.

Proof: Given $\epsilon > 0$ let us choose $\delta > 0$ and N such that if $n \geq N$ and $C \subset A^n$, then

$$Q_n(C) \geq 2^{-n\delta} \Longrightarrow Q_n([C]_\epsilon) \geq 1 - \epsilon. \tag{4}$$

Fix $n \geq N$ and a set $C \subset A^n$ such that $Q_n(C) \geq 2^{-n\delta/2}$ and note that $Q_n([C]_\epsilon) \geq 1 - \epsilon$. We will show that

$$\bar{d}_n(Q_n, Q_n(\cdot|C)) \leq 2\epsilon + 2^{-n\delta/2}. \tag{5}$$

This, of course, will establish the lemma and hence the fact that blowing-up implies finitely determined.

Let (W, μ) and (Z, ν) be two nonatomic probability spaces and choose measurable partitions, $\{W(a_1^n): a_1^n \in A^n\}$ of (W, μ) and $\{Z(a_1^n), a_1^n \in C\}$ of (Z, ν), such that

$$\begin{aligned} \mu(W(a_1^n)) =& Q_n(a_1^n),\ a_1^n \in A^n, \\ \nu(Z(a_1^n)) =& Q_n(a_1^n|C),\ a_1^n \in C. \end{aligned}$$

The **names** of $w \in W$ and $z \in Z$ are the respective sequences, w_1^n and z_1^n, such that $w \in W(w_1^n)$ and $z \in Z(z_1^n)$. Our goal is to show that there is a measure preserving transformation $\varphi: W \mapsto Z$ such that $d_n(w_1^n, \varphi(w)_1^n) \leq \epsilon$, except for a set of measure at most $\epsilon + 2^{-n\delta/2}$; this will establish (5).

Let us say that an invertible measure preserving mapping φ from $\tilde{W} \subset W$ onto $\tilde{Z} \subset Z$ is an **ϵ-matching on** $\tilde{W}$ if $d_n(w_1^n, \varphi(w)_1^n) \leq \epsilon$, except for a subset of $\tilde{W}$ of measure at most $\epsilon\mu(\tilde{W})$. We shall use the fact that $Q_n([C]_\epsilon) \geq 1 - \epsilon$ to construct an ϵ-matching φ on a subset W^1 of W of positive measure, then show that either the complement of W^1 has measure less than $2^{-n\delta/2}$ or we can extend the ϵ-matching to a larger set.

The key to the construction is the fact that if α is small enough then we can cut off an α-fraction of each $W(a_1^n)$ for which $a_1^n \in [C]_\epsilon$, and assign it to a subset of a $Z(b_1^n)$ set of the same measure for which $d_n(a_1^n, b_1^n) \leq \epsilon$. Since $Q_n([C]_\epsilon) \geq 1 - \epsilon$ we can therefore achieve an ϵ-matching on an α-fraction of W. If the set of sequences in C that are not fully covered has measure at least $2^{-n\delta}$ then it blows up to a large set and we can repeat the construction on the unassigned part, removing another positive fraction and assigning it to produce an ϵ-matching on a larger fraction of the space.

Now let us fill in the details of the above argument. We will first show that if α is a small enough positive number then there are two collections of disjoint, measurable sets

$$\{W^1(a_1^n): a_1^n \in A^n,\}, \ \{Z^1(a_1^n): a_1^n \in A^n\},$$

such that the following hold.

(a) $W^1(a_1^n)$ is a subset of $W(a_1^n)$ of measure $\alpha\mu(W(a_1^n))$.

(b) $Z^1(a_1^n)$ has measure equal to $\mu(W^1(a_1^n))$ and is entirely contained in some $Z(b_1^n)$.

(c) If $a_1^n \in [C]_\epsilon$ then $Z^1(a_1^n) \subseteq Z(b_1^n)$, where $d_n(b_1^n, a_1^n) \leq \epsilon$.

(d) There is at least one $b_1^n \in C$ such that $Z(b_1^n)$ is the union of the $Z^1(a_1^n)$ that meet it.

Since the spaces (W, μ) and (Z, ν) are nonatomic while the sets A^n and C are finite, we can certainly, for each $a_1^n \in [C]_\epsilon$, cut off an α-fraction $W^1(a_1^n)$ of $W(a_1^n)$ and assign it in any way we like to a subset $Z^1(a_1^n)$ of some $Z(b_1^n)$ for which $d_n(b_1^n, a_1^n) \leq \epsilon$, provided only that α is small enough. Thus we can force (a), (b), and (c) to hold for $a_1^n \in [C]_\epsilon$. Then, making α smaller if necessary, we can cut off α-fractions of the remaining $W(a_1^n)$ and assign them however we want so that (a), (b), and (c) will hold for all a_1^n. Condition (d) can be forced to hold merely by scaling each set by the same same suitably chosen factor $\lambda \geq 1$.

The mapping φ is defined on the union $W^1 = \cup W^1(a_1^n)$ by mapping each $W^1(a_1^n)$ in a measure preserving way onto the corresponding $Z^1(a_1^n)$. Note that φ is an ϵ-matching on W^1, by property (c) and the fact that $Q_n([C]_\epsilon) \geq 1 - \epsilon$. This completes the first stage of the construction.

Next define C_1 to be the set of all $b_1^n \in C$ that were not fully covered at the first stage, that is, $Z(b_1^n)$ is not the union of the $Z^1(a_1^n)$ that are contained in it. Property (d) guarantees that C_1 is strictly smaller than C. If $Q_n(C_1) \geq 2^{-\delta n}$ then $Q_n([C_1]_\epsilon) \geq 1 - \epsilon$, by (4), and we replicate the preceding construction to obtain a subset W^2 of $W - W^1$ of positive measure α_1, and an ϵ-matching φ of W^2 onto a subset of $Z - Z^1$ such that the subset $C_2 \subset C_1$ of sequences that are still not fully covered by the range of φ has cardinality smaller than C_1. Thus we can keep going, as long as the set of sequences that are not fully covered has measure at least $2^{-\delta n}$. Since we can force the cardinality of the non-fully covered sequences to shrink, the process will stop after some finite number of steps, say k, at which point we will have an ϵ-matching on a subset $W^1 \cup W^2 \cup \cdots \cup W^k$ such that the set C_k of $b_1^n \in C_{k-1}$ that are not fully covered satisfies

$$Q_n(C_k) < 2^{-\delta n}.$$

Since we started with $Q_n(C) \geq 2^{-\delta n/2}$, we therefore must have

$$Q_n\left(W^1 \cup W^2 \cup \cdots \cup W^k\right) \geq 1 - 2^{-\delta n/2}.$$

The mapping φ can be extended in an arbitrary measure preserving way to the complement of $\cup_i W^i$. Thus we will have $d_n(\varphi(w)_1^n, w_1^n)) \leq \epsilon$, except on a set of measure at most $\epsilon + 2^{-\delta n/2}$, which completes the proof of the desired result (5). ∎

Remark 1: As the above proof shows, blowing-up actually implies the apparently stronger property that given $\epsilon > 0$ there is a $\delta > 0$ and an N such that if $n \geq N$ and $C \subseteq A^n$ then

$$Q(C) \geq 2^{-n\delta} \implies \bar{d}_n(Q, Q([C]_\epsilon) < \epsilon.$$

3.2.3 Finitely determined plus exponential rates implies blowing-up. We first note a simple connection between the blow-up of a set and the $\bar{d}_n$-distance concept. Let $C \subset A^n$ and let $d_n(x_1^n, C)$ denote the distance from x_1^n to C, that is, the minimum of $d_n(x_1^n, y_1^n)$, $y_1^n \in C$. For any probability measure μ on A^n it is clear that

$$E_\mu(d_n(x_1^n, C)) \leq \bar{d}_n(\mu, \mu(\cdot|C)),$$

so that if $\bar{d}_n(\mu, \mu(\cdot|C)) < \epsilon^2$, then, by the Markov inequality, the set of x_1^n such that $d_n(x_1^n, C) \geq \epsilon$ has μ measure less than ϵ. This proves the following

(6) $$\bar{d}_n(\mu, \mu(\cdot|C)) < \epsilon^2 \Rightarrow \mu([C]_\epsilon) > 1 - \epsilon.$$

Assume that Q is finitely determined and has the exponential rate of convergence property for both frequencies and entropy. The finitely determined property implies that conditioning on subsets of A^n whose members have good enough k-block distribution and good enough entropy will produce measures close in $\bar{d}_n$ to the original measure. This result, which is the reason why finitely determined processes have the "extremality" property, is stated here as the following lemma.

LEMMA 3: *Let Q be finitely determined. Given $\epsilon > 0$ there is a $\delta > 0$ and positive integers k and $N > k$ such that if $n \geq N$ and $C \subset A^n$ is such that*

(a) $|\hat{\Delta}_{x_1^n,k} - Q_k| \leq \delta$, $x_1^n \in C$.

(b) $|\frac{1}{n}H(Q_n(\cdot|C) - H(Q)| \leq \delta$,

then $\bar{d}_n(Q_n, Q_n(\cdot|C)) < \epsilon$.

Let δ, k and N be chosen using the lemma. For $n \geq N$ let

$$(7) \qquad \mathcal{B}_n = \{x_1^n\colon |\hat{\Delta}_{x_1^n,k} - Q_k| \geq \delta\} \cup \{x_1^n\colon |\log Q_n(x_1^n) + nH| \geq n\delta/2\},$$

that is, the set of n-sequences that have bad frequencies of k-blocks or too-large or too-small probability. Use the exponential rate property to choose $\alpha \in (0, \delta)$ so that $Q_n(\mathcal{B}_n) \leq 2^{-\alpha n}$, $n \geq N$.

Suppose $C \subset A^n$ is such that $Q_n(C) \geq 2^{-\gamma n}$, where $\gamma = \alpha/2$, and put $\tilde{C} = C - C \cap \mathcal{B}_n$. If n is large enough then we have

(i) $Q_n(\tilde{C}) > 2^{-\gamma n/2}$.

(ii) $|\hat{\Delta}_{x_1^n,k} - Q_k| \leq \delta$, $x_1^n \in \tilde{C}$.

(iii) $2^{-n(H+\delta/2)} \leq Q_n(x_1^n) \leq 2^{-n(H-\delta/2)}$, $x_1^n \in \tilde{C}$.

A simple calculation using (i) and (iii) gives the entropy bound

$$\left|\frac{1}{n}H(Q_n(\cdot|C) - H(Q)\right| < \delta.$$

Thus we can apply the lemma and conclude that $\bar{d}_n(Q_n, Q_n(\cdot|C)) < \epsilon$. We then apply the connection between $\bar{d}$-closeness and blowing-up, namely (6), to complete the proof that Q has the blowing-up property.

Remark 2: Note that for any ergodic process and $\delta > 0$ there is an n such that $Q_n(\mathcal{B}_n) < \delta$, where $\mathcal{B}_n$ is the set defined in (2). Thus if $Q_n(C) \geq 2\delta$ then (ii) and (iii) will hold along with (i') $Q_n(\tilde{C}) > \delta$. This shows that if Q is finitely determined then Q must have the weak blowing-up property, a property applied to some information theory questions in [1]. A process Q has the **weak blowing-up property** if given $\epsilon > 0$ and $\delta > 0$, there is an N such that if $n \geq N$ and $C \subseteq A^n$ then

$$(8) \qquad Q(C) \geq \delta \implies Q([C]_\epsilon) \geq 1 - \epsilon.$$

It is not hard to show that weak blowing-up is an isomorphism invariant and that there are processes with entropy 0 that have the weak blowing-up property.

3.3 FINITARY CODING AND BLOWING-UP. We turn now to the proof of our final result.

THEOREM 3: *Finitary coding preserves the blowing up property. In particular, a finitary process has the blowing-up property.*

Proof: Suppose P has the blowing-up property and Q is a finitary coding of P with encoder $F\colon B^Z \mapsto A^Z$, sliding-window encoder $f\colon B^Z \mapsto A$, and window-width function $w(x)$, $x \in B^Z$. The finitary assumption means that $w(x)$ has finite nonnegative integer values and that $f(x)$ depends only on the values $x_{-w(x)}^{w(x)}$. Thus, in particular, there is a k so large that $P(w(x) > k) < \epsilon^2$. For a $\delta > 0$ to be specified in the next paragraph and an integer n let $C \subset A^n$ satisfy $Q_n(C) \geq 2^{-n\delta}$ and put $D = F^{-1}C$. Let $\tilde{D}$ be the projection of D onto B_{-k+1}^{n+k} and note that $P(\tilde{D}) \geq 2^{-n\delta}$.

Since we have assumed that P has the blowing-up property we can suppose that δ is small enough and n is large enough so that if $\gamma = \epsilon/(2k+1)$ then $P([\tilde{D}]_\gamma) \geq 1 - \epsilon$. Next let G be the set of all x_{-k}^k such that $w(x) \leq k$, and define

$$T = \{x_{-k+1}^{n+k}\colon \hat{\Delta}_{x_{-k+1}^{n+k}, 2k+1}(G) \geq 1 - \epsilon\}.$$

The Markov inequality together with $P(w(x) > k) < \epsilon^2$ then implies that $P(T) \geq 1 - \epsilon$.

Now consider a sequence $x_{-k+1}^{n+k} \in [\tilde{D}]_\gamma \cap T$. In this sequence $(1-\epsilon)n$ of the sliding $(2k+1)$-blocks belong to G, and, moreover, there is a sequence $\bar{x}_{-k+1}^{n+k} \in \tilde{D}$ such that

$$d_{n+2k}(x_{-k+1}^{n+k}, \bar{x}_{-k+1}^{n+k}) < \gamma = \frac{\epsilon}{2k+1}.$$

Thus fewer than $(2k+1)\gamma n \leq \epsilon n$ of the $(2k+1)$-blocks in x_{-k+1}^{n+k} can differ from the corresponding block in $\bar{x}_{-k+1}^{n+k}$. In particular, there are at least $(1-2\epsilon)n$ $(2k+1)$-blocks in x_{-k+1}^{n+k} that belong to G and, at the same time coincide with the corresponding block in $\bar{x}_{-k+1}^{n+k}$.

Now choose y and $\bar{y} \in D$ such that $y_{-k+1}^{n+k} = x_{-k+1}^{n+k}$, and $\bar{y}_{-k+1}^{n+k} = \bar{x}_{-k+1}^{n+k}$, and put $z = F(y)$, $\bar{z} = F(\bar{y})$. The sequence $\bar{z}_1^n$ belongs to C and we have $d_n(z_1^n, \bar{z}_1^n) \leq 2\epsilon$, so that $z_1^n \in [C]_{2\epsilon}$. This proves that

$$Q_n([C]_{2\epsilon}) \geq P([D]_\gamma \cap T) \geq 1 - 2\epsilon,$$

and completes the proof of Theorem 3. ∎

References

[1] R. Ahlswede, P. Gács and J. Körner, *Bounds on conditional probabilities with applications in multi-user communications*, Z. Wahr. Ver. Geb. **34** (1976), 157–177.

[2] M. Akcoglu, M. Rahe, and A. del Junco, *Finitary codes between Markov processes*, Z. Wahr. Ver. Geb. **47** (1979), 305–314.

[3] I. Csiszár and J. Körner, *Information Theory. Coding Theorems for Discrete Memoryless Systems*, Akadémiai Kiadó, Budapest, 1981.

[4] K. Marton, *A simple proof of the blowing-up lemma*, IEEE Trans. Inform. Th. **IT-42** (1986), 445–447.

[5] D. Ornstein, *Ergodic Theory, Randomness, and Dynamical Systems*, Yale University Press, New Haven, 1974.

[6] D. Ornstein and B. Weiss, *How sampling reveals a process*, Annals of Probab. **18** (1990), 905–930.

[7] M. Pinsker, *Information and information stability of random variables and processes* (In Russian), Vol. 7 of the series *Problemy Peredači Informacii*, AN SSSR, Moscow, 1960. English translation: Holden-Day, San Francisco, 1964.

[8] D. Rudolph, *If a two-point extension of a Bernoulli shift has an ergodic square, then it is Bernoulli*, Israel J. Math. **30** (1978), 159–180.

[9] P. Shields, *Two divergence-rate counterexamples*, J. Theoretical Probability **6** (1993), 521–545.

[10] P. Shields, *Stationary coding of processes*, IEEE Trans. Inform. Th. **IT-25** (1979), 283–291.

[11] M. Smorodinsky and M. Keane, *Finitary isomorphism of irreducible Markov shifts*, Israel J. Math. **34** (1979), 281–286.

[12] M. Smorodinsky, *Finitary isomorphism of m-dependent processes*, Contemporary Math. **135** (1992), 373–376.