Reprinted from ADVANCES IN MATHEMATICS
All Rights Reserved by Academic Press, New York and London

Vol. 115, No. 1, September 15, 1995
Printed in Belgium

Convergence under $\times_q$ of $\times_p$ Invariant Measures on the Circle

AIMEE JOHNSON AND DANIEL J. RUDOLPH*

Mathematics Department, Tufts University, Medford, Massachusetts 02155; and Mathematics Department, University of Maryland, College Park, Maryland 20742

Consider the semigroup of maps $\{T_n\}_{n\in\mathbb{N}}$ where $T_n(x) = nx$ mod 1. Suppose p and q are *multiplicatively independent* integers in that they are not both powers of the same integer. Further suppose that μ is a Borel probability measure, invariant, ergodic, and of positive entropy for T_p. We show that the sequence of measures $T_q^i(\mu)$ converges weak* to Lebesgue measure on a subsequence of values i of *uniform full density*. © 1995 Academic Press, Inc.

1. INTRODUCTION

Consider the abelian semigroup of maps $\{T_n\}_{n\in\mathbb{N}}$ where $T_n(x) = nx$ mod 1 on the circle, which we will write as $[0, 1)$. Symbolically, the map T_n is the shift on the n-adic expansion of points in $[0, 1)$. There has been much interest in the dynamics of subsemigroups of this semigroup, in particular doubly generated ones of the form $\{T_p^i T_q^j\}_{i,j\in\mathbb{N}}$ where p and q are not both powers of the same integer (we refer to such as *multiplicatively independent* integers.) Furstenberg [F] has shown that the only infinite closed subsets of $[0, 1)$ invariant for such a subsemigroup is all of $[0, 1)$. In [R] Rudolph showed that the only Borel probability measures invariant for such a subsemigroup where p and q are actually relatively prime either have zero entropy for all elements, or must be Lebesgue measure. Johnson [J] extended this result showing that the necessary and sufficient condition forcing a positive entropy ergodic measure to be Lebesgue is multiplicative independence of p and q.

Another direction of interest is to be seen in the work of Schmidt [S] and Pearce and Keane [PK]. They have shown that for certain types of T_p invariant measures μ, μ-a.e. x is in fact normal to the base q, i.e. is generic for Lebesgue measure under T_q. The measures they consider are i.i.d. or mixing Markov on the symbolic level, hence definitely of positive

* This work supported in part by NSF Grant DMS-88-02593.

0001-8708/95 $12.00

entropy and ergodic. This has recently been strengthened by Feldman and Smordinsky [FS] to the case of μ a *nondegenerate p-Bernoulli* measure.

Our results here will concern a weaker class of measures, those which simply are ergodic and of positive entropy for T_p. Our conclusions will also be weaker, being convergence in probability rather than pointwise. Let us be more precise.

DEFINITION 1.1. Consider the collection of intervals $I \subseteq \mathbb{N}$, $I = \{i, i+1, ..., j\}$. Order these according to their length $l(I) = j - i + 1$ with $I_1 < I_2$ iff $l(I_1) < l(I_2)$.

Fix the multiplicatively independent pair p, q and let $T = T_p$ and $S = T_q$. For any Borel measure μ on $[0, 1)$, set

$$\mu_I = \frac{1}{l(I)} \sum_{d \in I} S^d \mu.$$

Let λ denote Lebesgue measure. Our principal result is the following:

THEOREM 1.2. *If μ is a Borel probability measure, invariant and ergodic for T with entropy $h_\mu(T) > 0$ then*

$$\lim_{l(I) \to \infty} \mu_I = \lambda.$$

DEFINITION 1.3. A subset $A \subseteq \mathbb{N}$ has *uniform full density* in $\mathbb{N}$ if for every small α there exists an M so that if $l(I) > M$ then $\#(A \cap I)/l(I) \geqslant 1 - \alpha$. (We prefer this more evocative name to the more standard Banach density 1.)

We say a sequence of measures μ_i converges to μ in uniform full density if for every $\varepsilon > 0$ the set $A(\varepsilon) = \{i: w_*(\mu_i, \mu) < \varepsilon\}$ has uniform full density. Here by w_* we mean some choice for a metric giving the weak* topology on measures.

COROLLARY 1.4. *For μ as in Theorem 1.2, the measures $S^i(\mu)$ converge to λ in uniform full density.*

Proof. If the corollary is false we can find an $\varepsilon > 0$ and a sequence of intervals I_k with $l(I_k) \to \infty$ so that

$$\frac{\#(A(\varepsilon)^c \cap I_k)}{l(I_k)} > \varepsilon.$$

That is to say, a fixed fraction of the $S^i(\mu)$'s in each I_k **stay a fixed fraction** away from λ. Using compactness of the space of Borel probability measures, let $\{m_1, m_2, ..., m_t\}$ be a finite $\varepsilon/3$-dense set of measures. Any measure which is ε away from λ must be within $\varepsilon/3$ of some m_i which is at least $2\varepsilon/3$ away from λ. Thus along some subsequence of values k we know that a fraction of at least ε/t of the measures $S^i(\mu)$, $i \in I_k$ are within $\varepsilon/3$ of some fixed m_i which is at least $2\varepsilon/3$ away from λ. Split these averages into two pieces, those near m_i and those not, and choose a further subsequence where both pieces converge. We are left with two T_p invariant measures which average to λ, one of which is at least $\varepsilon/3$ away from λ. But T_p acts ergodically on λ and we have a conflict. ∎

Compactness of the space of Borel probability measures tells us that Theorem 1.2 is equivalent to the following.

THEOREM 1.5. *For μ satisfying the hypotheses of Theorem* 1.2, *consider I_k with $l(I_k) \to \infty$ and with μ_{I_k} converging weak* to some measure $\hat{\mu}$. Then $\hat{\mu} = \lambda$.*

The rest of this paper concerns itself with proving Theorem 1.5. From now on the sequence of intervals I_k will be fixed, so we write them as $I(k)$, I: $\mathbb{N} \to \{\text{intervals in } \mathbb{N}\}$ and we abbreviate $\lim_{k \to \infty} \hat{\mu}_{I(k)} = \hat{\mu}$ by $\mu_I \to \hat{\mu}$.

In Section 2 we will represent the 2-dimensional dynamical system symbolically. The main idea of the proof of Theorem 1.5 is to investigate the origin of the entropy of $\hat{\mu}$ within that of the original μ. This leads us to examine the conditional expectations of preimage symbols under the map T. In Section 3 we will discuss how such preimages lie in the second quadrant of symbols in our symbolic system and how our maps act on them. In Section 4 we will show that if $\mu_I \to \hat{\mu}$, then $\hat{\mu}$ can be written as $\alpha\lambda + (1-\alpha)\mu_0$ where $h_{\mu_0}(T) = 0$ and $\alpha > 0$. In order to show that in fact $\alpha = 1$ we will partition the second quadrant of our symbolic representation into a finite number of *staircases* descending to the origin, as in [J], with slopes determined by the numbers p and q. If $\alpha \neq 1$ we show that μ-a.s. the symbols to the right of each staircase determines the symbols to the right of the next steepest staircase, which inductively tells us the first quadrant of symbols determines the second. But this implies $h_\mu(T) = 0$ completing the proof.

In Section 8 we will return to the two areas of interest mentioned earlier, measures invariant for both T and S, and under what conditions on a T-invariant measure μ will we be able to say most points relative to μ are normal to the base q. We will give some partial insight into these questions that follows from Theorem 1.2.

2. THE SYMBOLIC REPRESENTATION

We will first consider how to lift the two maps T and S from the circle to symbolic representations and then how to lift the measure μ to this symbolic space. These ideas can be found in [R] and thus proofs of the basic results will be omitted.

Let P be the partition (up to overlapping boundaries) of $[0,1)$ into subintervals $I_j = [j/pq, (j+1)/pq]$, $0 \leq j < pq-1$. Clearly $\{I_j\}$ is a Markov partition for both T and S. Let $V = \{x \in [0,1): x = t/p^n q^m, n, m \in \mathbb{N}\}$. V contains all points x for which $T^{n-1}S^{m-1}(x)$ is a boundary point of P.

Associate to each map a $pq \times pq$ transition matrix of zeros and ones: $M_T = [a_{ij}]$ where $a_{ij} = 1$ iff $I_j \subset T(I_i)$ and similarly define M_S. Let $\Sigma = \{0, 1, ..., pq-1\}$ be the state space associated with these matrices.

Let $Y \subseteq \Sigma^{\mathbb{N}^2}$ consists of all 2-dimensional arrays which are M_T allowed on rows and M_S allowed on columns. We can think of a point $y \in Y$ as a 'first quadrant' of symbols, where there is a symbol at each nonnegative lattice point. Since the left shift on Y corresponds to T on $[0,1)$ we will again call it T. Similarly, let S represent the downshift on Y. More precisely, for $y \in Y$, $Ty(i,j) = y(i+1, j)$ and $Sy(i,j) = y(i, j+1)$.

To any point $x \in [0,1) \backslash V$ there corresponds a unique point $y(x) \in Y$. Just set $y(x)(n,m) = j$ iff $T^n S^m x \in I_j$. For $x \in V$ there is some n, m with $T^n S^m x$ on the boundary between two I_j's. The symbol $y(x)(n,m)$ could be chosen to be either the interval to the right or that to the left of x. However, if we specify the left (or right) interval then in order to obey the transition rules we must take the left (or right) interval for all larger n, m. Thus there are exactly two points in Y that represent each $x \in V$.

Define the map $\varphi: Y \to [0,1)$ by $\varphi(y) = \bigcap_{i=1}^{\infty} T^{-i}S^{-i}(I_{y(i,i)})$. This is 1-1 everywhere except the countable set V where it is 2-1. With the product topology on Y, φ is continuous and $\varphi(y(x)) = x$.

Let $\hat{Y} \subseteq \Sigma^{\mathbb{Z}^2}$ be those doubly infinite arrays where all rows are M_T allowed and all columns are M_S allowed. Thus a point $\hat{y} \in \hat{Y}$ is an entire 2-dimensional integer lattice of symbols. Let $\hat{\varphi}: \hat{Y} \to [0,1)$ be the map which sends $\hat{y}$ to the point in $[0,1)$ associated by φ with its first quadrant. Let T and S again represent the left and downshift. Notice that on $\hat{Y}$ these maps are invertible and that $T\hat{\varphi} = \hat{\varphi}T$, $S\hat{\varphi} = \hat{\varphi}S$.

$\varphi^{-1}(P)$ is the time zero partition of $\hat{Y}$ and will also be denoted by P. Then $\bigvee_{i=0}^{n} T^{-i}P$ is the partition of $\hat{Y}$ into cylinder sets determined by symbols at positions in the block $[(0,0), ..., (n,0)]$. Such a block corresponds to an interval of length $1/p^{n+1}q$ in $[0,1)$. In particular, the positive horizontal axis of symbols determines the point $x = \hat{\varphi}(\hat{y})$.

Now we want to describe how to lift the measure μ to this symbolic space. Recall that μ is invariant for T but not necessarily for S.

LEMMA 2.1. $\mu(V) = 0$.

Proof. Since T is ergodic and of positive entropy for μ, μ must be non-atomic, and of course V is countable. ∎

Since there is a 1-1 μ-a.e. correspondence between the first quadrant of symbols and $[0, 1)$ we can lift μ to a measure on Y. In particular, if C is a cylinder set from $\bigvee_{i=0}^{n} T^{i}P$ then $\mu(C)$ is the measure of the associated interval in $[0, 1)$. This measure on cylinders extends to a T-invariant measure on $\hat{Y}$ as it extends to all cylinders indexed on $\mathbb{Z}$ and any finitely additive measure on cylinders of such a symbolic space extends to a unique T-invariant Borel measure. Now consider a cylinder set $\tilde{C} \in \bigvee_{i=0}^{n} T^{-i}S^{-k}P$. Then $C = S^{k}(\tilde{C}) \in \bigvee_{i=0}^{n} T^{-i}P$, and $\hat{y} \in \tilde{C}$ iff $S^{k}(\hat{y}) \in C$. So $\mu(\tilde{C}) = \mu\{\hat{y}\colon S^{k}\hat{y} \in C\} = \mu(S^{-k}C) = S^{k}\mu(C)$. Thus the measure μ on cylinder sets in row k is just $S^{k}\mu$ of the same cylinder but sitting in row zero.

3. Preimages

Recall that p and q are multiplicatively independent and thus can be written as

$$p = p_0 \pi_1^{n_1} \cdots \pi_h^{n_h},$$

$$q = q_0 \pi_1^{m_1} \cdots \pi_h^{m_h},$$

where $p_0, q_0, \pi_1, ..., \pi_h$ are all relatively prime integers, all $m_i > 0$ and

$$\frac{n_1}{m_1} > \cdots > \frac{n_h}{m_h}$$

and either $p_0 \neq 1$, $q_0 \neq 1$, or $h \geqslant 2$. This is obtained just by grouping the primes in the prime power decompositions of p and q in the proper way.

We will now discuss the preimages of $x \in [0, 1)$ and how these correspond to symbols in $\hat{Y}$. A more detailed description can be found in [J].

Fix a point $\hat{y} \in \hat{Y}$. Let $x = \hat{\varphi}(\hat{y})$. Then

$$T^{-r}(x) = \left\{\frac{x}{p^r} + \frac{i}{p^r}\right\}_{i=0}^{p^r-1}.$$

In terms of $\hat{y}$ these points correspond to the p^r possible words that could be seen in the block of positions $[(-r, 0), (-r+1, 0), ..., (-1, 0)]$ that are consistent with the first quadrant of $\hat{y}$.

Definition 3.1. Let k_{0r} be the smallest integer such that $k_{0r}m_1 \geqslant rn_1$.

The n_i's and m_i's have been ordered so that it now follows that $k_{0r}m_i \geqslant rn_i$ for all $i = 1, ..., h$. It follows that

$$S^{k_{0r}}T^{-r}(x) = \left\{\frac{q^{k_{0r}}x}{p^r} + \frac{i'}{p_0^r}\right\}_{i'=0}^{p_0^r-1}.$$

In terms of $\hat{y}$ these points correspond to the p_0^r possible words in the block of positions $[(-r, k_{0r}), ..., (-1, k_{0r})]$ that are consistent with the first quadrant of symbols of $\hat{y}$. We will refer to these as the p_0^r preimages at position $(-r, k_{0r})$. Further application of S yields a 1-1 map between these preimages and ones at positions $(-r, k_{0r}+i), i \geq 1$. Notice that $S^{k_{0r}}$ acts as a $\pi_1^{n_1 r} \cdots \pi_h^{n_h r}$ to 1 map from the preimages at $(-r, 0)$ to those at $(-r, k_{0r})$.

DEFINITION 3.2. Set

$$\mathscr{D}_0 = \bigvee_{i=0}^{\infty} T^{-i}(P) \vee \bigvee_{i=1}^{\infty} S^{-k_{0i}} T^i(P).$$

Thus $\mathscr{D}_0(\hat{y})$ is the array of symbols from $\hat{y}$ occupying the first quadrant and the blocks $[(-r, k_{0r}), ..., (-1, k_{0r}), r \geq 1$. Think of this as specifying $\hat{\varphi}(\hat{y}) = x \in [0, 1)$ and a choice for the preimages associated with positions $(-r, k_{0r})$.

PROPOSITION 3.3. *Given* $\mathscr{F} = \bigvee_{i=0}^{\infty} T^{-i}(P)$ *and the preimage associated to the position* $(-r, k_{0r})$, *the preimages at positions* $(-n, k_{0n})$ *for* $1 \leq n \leq r$ *are determined.*

Note: We are using probabilistic vocabulary here. "Given $\bigvee_{i=0}^{\infty} T^{-i}(P)$" means "conditioning on the fixed values of the symbols in the first quadrant of some $\hat{y}$."

Proof. Knowing the preimage associated with $(-r, k_{0r})$ trivially tells us the preimage associated with $(-n, k_{0r})$, $1 \leq n < r$. Since $k_{0r} \geq k_{0n}$, there is a 1-1 corrrespondence between the preimages associated with $(-n, k_{0_n})$ and $(-n, k_{0r})$ and this gives the result. ∎

PROPOSITION 3.4. *For each fixed* $\hat{y} \in \hat{Y}$, *the symbols* $\mathscr{D}_0(\hat{y})$ *and the single symbol at* $(-1, 0)$ *determines all the symbols* $\mathscr{D}_0(T^{-1}\hat{y})$.

Proof. Recall that the preimages at $(-r, 0)$ are a coset of the additive group $\{i/p^r\}_{i=0}^{p^r-1}$ and at $(-r, k_{0r})$, a coset of $\{i/p_0^r\}_{i=0}^{p_0^r-1}$. If we are given $\mathscr{D}_0(\hat{y})$, we know which of the i/p_0^r occurs and this narrows the possibilities at $(-r, 0)$ down to a coset of $\{i/\pi_1^{rn_1} \cdots \pi_h^{rn_h}\}$. (From here on we will omit the obvious domain of possible values for i in these subgroups.) If in addition we are told which of the cosets of $\{i/\pi_1^{n_1} \cdots \pi_h^{n_h}\}$ actually occurs at $(-r, 0)$ this further reduces the possibilities at $(-r, 0)$ to a coset of $\{i/\pi_1^{(r-1)n_1} \cdots \pi_h^{(r-1)n_h}\}$. We know that $k_{0(r-1)}$ applications of S will collapse these to one preimage and this determines the symbols of $\mathscr{D}_0(T^{-1}\hat{y})$. ∎

Here we have considered preimages and how they collapse under the map S. Next we will consider their movement under $S^{n_a}T^{-m_a}$, $1 \leqslant a \leqslant h$. We call this the movement of the preimages "up the staircase."

DEFINITION 3.5. For $a = 1, ..., h-1$ and $j \geqslant 1$, let k_{aj} be the smallest integer such that

$$jn_{a+1} - k_{aj}(m_{a+1}n_a - n_{a+1}m_a) \leqslant 0.$$

Define $k_{hj} = 0$ for all $j \geqslant 1$.

DEFINITION 3.6. For $a = 1, ..., h$ define

$$\mathscr{D}_{a-1} = \bigvee_{i=0}^{\infty} T^{-i}(P) \vee \bigvee_{i=1}^{\infty} S^{-k_{0i}}T^{i}(P) \vee \bigvee_{i=1}^{\infty} \bigvee_{j=1}^{a-1} S^{-k_{ji}n_j}T^{i+k_{ji}m_j}(P).$$

Here $\mathscr{D}_{a-1}(\hat{y})$ is the array of symbols in the first quadrant of $\hat{y}$ and the blocks $[(-r-k_{jr}m_j, k_{jr}n_j), ..., (-1, k_{jr}n_j)]$ for $r \geqslant 1, j = 1, ..., a-1$. The choice of this block position is clarified in the next theorem.

THEOREM 3.7. *Given $\mathscr{D}_{a-1}(\hat{y})$, the possible preimages of $\hat{\varphi}(\hat{y})$ at position $(-r, 0)$ are a coset of $\{i/\pi_a^{rm_a} \cdots \pi_h^{rm_h}\}$. Apply the map $S^{n_a}T^{-m_a}$ iteratively to these. After k_{ar} iterations we have only a coset of $\{i/\pi_a^{rm_a}\}$ left, and so $S^{k_{ar}n_a}T^{-k_{ar}m_a}$ is a $\pi_{a+1}^{rm_{a+1}} \cdots \pi_h^{rm_h}$ to 1 map. These are the preimages associated to position $(-r-k_{ar}m_a, k_{ar}n_a)$. Further applications of $S^{n_a}T^{-m_a}$ act on these preimages as a 1-1 map.*

Proof. See section 4, lemma 5 of [J]. ∎

Thus, for example, $\mathscr{D}_1(\hat{y})$ can be thought of as specifying first of all the point $\hat{\varphi}(\hat{y}) \in [0, 1)$, the preimages at positions $(-r, k_{0r})$, $r \geqslant 1$ as before, and also now the preimages at positions $(-r-k_{1r}m_1, k_{1r}n_1)$, $r \geqslant 1$. A preimage at a specific $(-r-k_{1r}m_1, k_{1r}n_1)$ then determines the preimage at $(-r-km_1, kn_1)$, $k \geqslant k_{1r}$. These determine a periodic staircase pattern in the second quadrant with slope $-n_1/m_1$. Notice that the n_i's and m_i's are arranged so at each step the slope of these staircases increases. In particular $\mathscr{D}_{a-1} \subseteq \mathscr{D}_a$ and finally $\mathscr{D}_h$ determines all the symbols in the second quadrant.

PROPOSITION 3.8. *Given $\mathscr{D}_{a-1}(\hat{y})$, the preimages at $(-r-k_{ar}m_a, k_{ar}n_a)$ determine the preimages at $(-s-k_{as}m_a, k_{as}n_a)$ for $1 \leqslant s < r$.*

Proof. We will show this for $s = r-1$ and the rest will follow. The preimages at $(-r-k_{ar}m_a, k_{ar}n_a)$ certainly determine the preimages at $(-r+1-k_{ar}m_a, k_{ar}n_a)$. But then the result follows from the 1-1 relationship between the preimages at $(-r+1-k_{ar}m_a, k_{ar}n_a)$ and $(-r+1-k_{a(r-1)}m_a, k_{a(r-1)}n_a)$. ∎

PROPOSITION 3.9. *Given $\mathcal{D}_{a-1}(\hat{y})$, the symbol at* $(-1, 0)$ *determines* $\mathcal{D}_{a-1}(T^{-1}\hat{y})$.

Proof. We need to show that $T^{-1}\hat{y}(-i-k_{bi}m_b, k_{bi}n_b)$ is determined for $i \geqslant 1$, $b = 1, ..., a-1$. Fix some arbitrary choice for i and b in these sets. The above is equivalent to showing $\hat{y}(-i-1-k_{bi}m_b, k_{bi}n_b)$ is determined. Given $\mathcal{D}_{a-1}(\hat{y})$ we know the possible preimages at $(-i-1-k_{bi}m_b, 0)$ are a coset of

$$\left\{\frac{i}{\pi_a^{(i+1+k_{bi}m_b)n_a} \cdots \pi_h^{(i+1+k_{bi}m_b)n_h}}\right\}.$$

If in addition we know the symbol at $(-1, 0)$ this reduces the above set to a coset of $\{i/\pi_a^{(i+k_{bi}m_b)n_a} \cdots \pi_h^{i+k_{bi}m_b)n_h}\}$. Now act by $S^{k_{bi}n_b}$ and we get just those terms of the form

$$\left\{\frac{i\pi_a^{k_{bi}n_b n_a} \cdots \pi_h^{k_{bi}n_b m_h}}{\pi_a^{(i+1+k_{bi}m_b)n_a} \cdots \pi_h^{(i+1+k_{bi}m_b)n_h}}\right\}.$$

But we chose k_{bi} so that in fact all terms here cancel and the preimage is uniquely determined. ∎

4. THE DECOMPOSITION OF $\hat{\mu}$

Recall that what we want to show is that if $\mu_I \to \hat{\mu}$, then $\hat{\mu}$ is Lebesgue measure.

PROPOSITION 4.1. *$\hat{\mu}$ is both T and S invariant and*

$$h_{\hat{\mu}}(T) \geqslant h_\mu(T).$$

Proof. Recall that $\mu_I = 1/l(I) \sum_{d \in I} S^d\mu$. So $S\mu_I = 1/l(I) \sum_{d \in I} S^{d+1}\mu = \mu_{I+1}$. For f any continuous function and $I = \{i, ..., j\}$ we can write then

$$\int f\, d\mu_I - \int f\, d\mu_{I+1} = \frac{1}{j-i+1}\left\{\int f\, d(S^i\mu) - \int f\, d(S^{j+1}\mu)\right\}$$

$$\leqslant \frac{1}{j-i+1} \|f\|_\infty.$$

This converges to 0 as $l(I) \to \infty$. The first integral converges to $\int f\, d\hat{\mu}$ and the second to $\int f\, d(S\hat{\mu})$. Thus $\hat{\mu}$ is S invariant, and as it is the limit of T invariant measures it is T invariant.

The system (Y, T, μ) is a q^i-point extension of $(Y, T, S^i\mu)$ and so must have the same entropy. Hence μ_I is the average of measures all of which have entropy $h_\mu(T)$ and thus also has entropy $h_\mu(T)$. Using upper semi-continuity in the weak* topology of entropy on such shift spaces, we have

$$h_\mu(T) \leqslant \overline{\lim}\, h_{\mu_l}(T) \leqslant h_{\hat\mu}(T). \quad \blacksquare$$

This tells us $h_{\hat\mu}(T) > 0$ as we assumed $h_\mu(T) > 0$. From [J] we know that any measure satisfying Proposition 4.1 that further is ergodic for the pair $\{T, S\}$ must be Lebesgue measure. The measure $\hat\mu$ is not necessarily ergodic but can be decomposed into its ergodic components. Thus we can write

$$\hat\mu = \alpha\lambda + (1-\alpha)\mu_0$$

where $\alpha \neq 0$ and $h_{\mu_0}(T) = 0$. In the remaining sections we will show that if in fact $\alpha \neq 1$ then $\alpha = 0$ completing the proof of Theorem 1.5.

5. From $\hat\mu$ back to μ

As said in section 3, given a point $x \in [0, 1)$, $T^{-n}(x)$ consists of p^n distinct points. Their names in Y agree except on their left-most n positions. These p^n words correspond to elements in the partition $\bigvee_{i=-1}^{-n} T^{-i}P$ consistent with x. Recall that if $\{z_j\}$ is a list of these words then

$$D\left(\bigvee_{i=-1}^{-n} T^{-i}P \mid \mathscr{F}\right)(\hat y)$$

is a p^n-dimensional probability vector with entries $E(z_j \mid \mathscr{F})(\hat y)$. We know that $h_{\hat\mu}(T) = 0$ on a T and S invariant set of measure $1-\alpha$ so on this set our map T is essentially invertible. Thus there must be exactly one j with

$$E_{\hat\mu}(z_j \mid \mathscr{F})(\hat y) = 1.$$

In this case we say $D_{\hat\mu}(\bigvee_{i=-1}^{-n} T^{-i}P \mid \mathscr{F})(\hat y)$ is a *trivial* vector. We will write it as $\vec e(\hat y)$, a standard basis vector consisting of a single 1 and all other terms zero. Thus

$$\hat\mu\left\{\hat y : D_{\hat\mu}\left(\bigvee_{i=-1}^{-n} T^{-i}P \mid \mathscr{F}\right)(\hat y) = \vec e(\hat y)\right\} \geqslant 1-\alpha.$$

By the martingale convergence theorem we can pick $M = M(n, \varepsilon)$ so that

$$\hat\mu\left\{\hat y : D_{\hat\mu}\left(\bigvee_{i=-1}^{-n} T^{-i}P \,\middle|\, \bigvee_{i=0}^{M} T^{-i}P\right)(\hat y) = \left(1-\frac{\varepsilon}{2}\right)\vec e(\hat y) + \frac{\varepsilon}{2}\vec v(\hat y)\right\} \geqslant 1-\alpha-\frac{\varepsilon}{2},$$

where $\vec{v}(\hat{y})$ is some arbitrary probability vector, meant to denote the small amount of weight that $\hat{\mu}$ may now put on other preimages. By conditioning on a large but finite amount of the future, we have lost part of our set of measure $1-\alpha$ and the past distribution is only almost trivial. But the distribution function now only depends on a finite cylinder set. Thus if we go out in our sequence of measures μ_I far enough,

$$\mu_I\left\{\hat{y}\colon D_{\mu_I}\left(\bigvee_{i=1}^{-n} T^{-i}P \,\middle|\, \bigvee_{i=0}^{M} T^{-i}P\right)(\hat{y}) = (1-\varepsilon)\,\vec{e}(\hat{y}) + \varepsilon\vec{v}(\hat{y})\right\} \geqslant 1-\alpha-\varepsilon.$$

Since μ_I is the average of a block of $S^d\mu$'s and $\vec{e}(\hat{y})$ is an extreme point in the set of probability vectors, the above statement about μ_I gives a very similar statement concerning some individual $S^d\mu$. This will be shown in the next two propositions. Since we are now conditioning on $\bigvee_{i=0}^{M} T^{-i}P$, we can refer to this as a function of an $(M+1)$-name η instead of a point $\hat{y}$. In other words we write

$$D(Q \mid \eta) = D\left(Q \,\middle|\, \bigvee_{i=0}^{M} T^{-i}\right)(\hat{y})$$

where $\hat{y}$ has the name η at positions $[0, ..., M]$. This will simplify notation somewhat.

Definition 5.1.

$$W(\eta) = \left\{d \in I\colon D_{S^d\mu}\left(\bigvee_{i=-1}^{-n} T^{-i}P \mid \eta\right) = (1-\sqrt{\varepsilon})\,\vec{e}(\eta) + \sqrt{\varepsilon}\,\vec{v}(\eta)\right\}.$$

Proposition 5.2.

$$\frac{1}{l(I)} \sum_{d \in W(\eta)} S^d\mu(\eta) \geqslant (1-\sqrt{\varepsilon})\,\mu_I(\eta).$$

Proof. Assume not. Let $A = \{d\colon d \in W(\eta)^c \cap I\}$. Then

$$\frac{1}{l(I)} \sum_{d \in A} S^d\mu(\eta) \geqslant \sqrt{\varepsilon}\,\mu_I(\eta).$$

Let $\gamma \in \bigvee_{i=-1}^{-n} T^{-i}P$ be that preimage on which μ_I puts all but ε of its weight. Then

$$\mu_I(\eta \cap \gamma^c) \geqslant \frac{1}{l(I)} \sum_{d \in A} S^d\mu(\eta \cap \gamma^c)$$

$$\geqslant \frac{1}{l(I)} \sum_{d \in A} \sqrt{\varepsilon}\, S^d\mu(\eta)$$

$$> \sqrt{\varepsilon}\,\sqrt{\varepsilon}\,\mu_I(\eta).$$

Thus $\mu_I(\eta \cap \gamma^c)/\mu_I(\eta) > \varepsilon$, yet $\mu_I(\eta \cap \gamma)/\mu_I(\eta) \geqslant 1-\varepsilon$ which is a contradiction. ▮

The choice of indices $W(\eta)$ may vary with η. Let $\mathcal{N}$ be the union of all η such that

$$D_{\mu_I}\left(\bigvee_{i=-1}^{-n} T^{-i}P \mid \eta\right) = (1-\varepsilon)\,\vec{e}(\eta) + \varepsilon\vec{v}(\eta).$$

We know

$$\mu_I(\mathcal{N}) \geqslant 1-\alpha-\varepsilon, \text{ i.e. } \sum_{d \in I} \frac{S^d\mu(\eta)}{l(I)} \geqslant 1-\alpha-\varepsilon.$$

Let $\mathcal{G} = \{(\eta, d)\colon d \in W(\eta), \eta \in \mathcal{N}\}$. Proposition 5.2. says

$$\sum_{(\eta, d) \in \mathcal{G}} \frac{S^d\mu(\eta)}{l(I)} \geqslant (1-\sqrt{\varepsilon})\,\mu_I(\mathcal{N}),$$

or equivalently,

$$\sum_{\substack{(\eta, d) \notin \mathcal{G} \\ n \in \mathcal{N}}} \frac{S^d\mu(n)}{l(I)} \leqslant \sqrt{\varepsilon}\,\mu_I(\mathcal{N}).$$

DEFINITION 5.3.

$$A = \left\{d\colon \sum_{\{\eta\colon (\eta, d) \in \mathcal{G}\}} S^d\mu(\eta) \geqslant (1-\varepsilon^{1/4})\, S^d\mu(\mathcal{N})\right\}.$$

PROPOSITION 5.4.

$$\sum_{d \in A} \frac{S^d\mu(\mathcal{N})}{l(I)} \geqslant (1-\varepsilon^{1/4})\,\mu_I(\mathcal{N}).$$

Proof. Assume not, and

$$\sum_{d\in A^c} \frac{S^d\mu(\mathcal{N})}{l(I)} > \varepsilon^{1/4}\mu_I(\mathcal{N}).$$

Consider

$$\sum_{d} \sum_{\substack{(\eta,d)\notin\mathcal{G}\\ \eta\in\mathcal{N}}} \frac{S^d\mu(\eta)}{l(I)} \geqslant \sum_{d\in A^c} \sum_{\substack{(\eta,d)\notin\mathcal{G}\\ \eta\in\mathcal{N}}} \frac{S^d\mu(\eta)}{l(I)}.$$

By definition of A^c, the above is strictly larger than

$$\sum_{d\in A^c} \frac{\varepsilon^{1/4}S^d(\mathcal{N})}{l(I)} > \sqrt{\varepsilon}\,\mu_I(\mathcal{N}).$$

But this is a contradiction. ∎

So

$$\frac{1}{l(I)} \sum_{d\in A} S^d\mu(\mathcal{N}) \geqslant (1-\varepsilon^{1/4})\,\mu_I(\mathcal{N})$$

and thus

$$\frac{1}{\#(A)} \sum_{d\in A} S^d\mu(\mathcal{N}) \geqslant (1-\varepsilon^{1/4})\,\mu_I(\mathcal{N}).$$

So for at least one $d\in A$, $S^d\mu(\mathcal{N}) \geqslant (1-\varepsilon^{1/4})\,\mu_I(\mathcal{N})$. From the definition of A we conclude

$$\sum_{\{\eta:\,(\eta,d)\in\mathcal{G}\}} S^d\mu(\eta) \geqslant (1-\varepsilon^{1/4})^2\,(1-\alpha-\varepsilon).$$

Since $d\in W(\eta)$, we can rewrite this as

$$S^d\mu\left\{\hat{y}: D_{S^d\mu}\left(\bigvee_{i=-1}^{-n} T^{-i}P \,\middle|\, \bigvee_{i=0}^{M} T^{-i}P\right)(\hat{y}) = (1-\sqrt{\varepsilon})\,\vec{e}(\hat{y}) + \sqrt{\varepsilon}\,\vec{v}(\hat{y})\right\}$$

$$\geqslant (1-\varepsilon^{1/4})^2\,(1-\alpha-\varepsilon).$$

Let $M\to\infty$ and $\bigvee_{i=0}^{M} T^{-i}P \to \mathcal{F}$. Then we have

$$S^d\mu\left\{\hat{y}: D_\mu\left(S^{-d}\left(\bigvee_{i=-1}^{-n} T^{-i}P\right) \,\middle|\, S^{-d}\mathcal{F}\right)(S^{-d}\hat{y}) = (1-\sqrt{\varepsilon}\,\vec{e}(\hat{y}) + \sqrt{\varepsilon}\,\vec{v}(\hat{y})\right\}$$

$$\geqslant (1-\varepsilon^{1/4})^2\,(1-\alpha-\varepsilon).$$

Now let $\hat{z} = S^{-d}\hat{y}$, so $S^d\hat{z} = \hat{y}$. Rewrite the above as

$$S^d\mu\left\{S^d\hat{z}: D_\mu\left(S^{-d}\left(\bigvee_{i=-1}^{-n} T^{-i}P\right)\bigg| S^{-d}\mathscr{F}\right)(\hat{z}) = (1-\sqrt{\varepsilon})\,\vec{e}(S^d\hat{z}) + \sqrt{\varepsilon}\,\vec{v}(S^d\hat{z})\right\}$$

$$\geqslant (1-\varepsilon^{1/4})^2(1-\alpha-\varepsilon).$$

This is the same as

$$\mu\left\{\hat{z}: D_\mu\left(S^{-d}\left(\bigvee_{i=-1}^{-n} T^{-i}P\right)\bigg| S^{-d}\mathscr{F}\right)(\hat{z}) = (1-\sqrt{\varepsilon})\,\vec{e}(S^d\hat{z}) + \sqrt{\varepsilon}\,\vec{v}(S^d\hat{z})\right\}$$

$$\geqslant (1-\varepsilon^{1/4})^2(1-\alpha-\varepsilon).$$

These are the points with the property that if the future on row d is given, the past of length n on row d is almost determined; i.e. μ puts most of its weight on just one of the p^n possible preimages. Let the set of such points $\hat{z}$ be called $\mathscr{U}$. So the above says

$$\mu(\mathscr{U}) \geqslant (1-\varepsilon^{1/4})^2(1-\alpha-\varepsilon).$$

In the next two sections we will use the symbolic machinery of sections 2 and 3 to move this down to the equivalent statement about the 0th row.

6. $\mathscr{D}_0 \subseteq \mathscr{F}$

We know

$$S^{-d}\hat{\varphi}(S^d\hat{z}) = \left\{\hat{\varphi}(\hat{z}) + \frac{j}{q^d}\right\}_{j=0}^{q^d-1},$$

which is to say there are q^d preimages at $(0, 0)$ consistent with $\hat{\varphi}(S^d\hat{z})$. The expectation of any particular preimage at $(-n, d)$ given $\hat{\varphi}(S^d\hat{z})$ is thus the weighted average of its expectations given $\hat{\varphi}(\hat{z}) + (j/q^d)$. So we write

$$D_\mu\left(S^{-d}\bigvee_{i=-1}^{-n} T^{-i}P \mid S^{-d}\mathscr{F}\right)(\hat{z})$$

$$= \sum_{j=0}^{q^d-1}\left\{E_\mu\left(\hat{\varphi}^{-1}\left(\hat{\varphi}(\hat{z}) + \frac{j}{q^d}\right)\bigg| S^{-d}\mathscr{F}\right)(\hat{z})\right.$$

$$\left.\times D_\mu\left(S^{-d}\bigvee_{i=-1}^{-n} T^{-i}P \mid \mathscr{F}\right)(\hat{\varphi}^{-1}(\hat{\varphi}(\hat{z}) + j/q^d))\right\}. \qquad (1)$$

Notice that for a particular $\hat{\varphi}(\hat{z})+j/q^d$, its distribution vector has certain zeros forced topologically in that not all of the preimages from $S^{-d}(\bigvee_{i=-1}^{-n} T^{-i}P)$ are consistent with $\hat{\varphi}(\hat{z})+j/q^d$. Those that are not will of course have zero expectation. If we fix $\hat{z}$ then (1) has the form

$$\vec{w}=\sum_{j=0}^{q^d-1} a_j \vec{w}_j.$$

For $\hat{z}\in\mathscr{U}$, $\vec{w}$ is almost a trivial vector, one entry being nearly one and the others nearly zero. As $\vec{w}$ is in fact a probability vector, it is enough to say one term is nearly 1, and we direct our attention to this entry. Let us say it is the kth entry. We know then $w^k \geqslant 1-\sqrt{\varepsilon}$ or $\sqrt{\varepsilon}\geqslant 1-w^k$. We want to show that most of the $\vec{w}_j$ have the same property.

Definition 6.1. For $\hat{z}\in\mathscr{U}$ fixed, let $\vec{w}$ be as above, and set

$$J=\{j\in\{0,...,q^d-1\} \text{ such that } \varepsilon^{1/4}\geqslant 1-w_j^k\}.$$

Proposition 6.2. *For $\hat{z}\in\mathscr{U}$ and $\vec{w}=\sum_{j=0}^{q^d-1} a_j\vec{w}_j$ as above,*

$$\sum_{j\in J} a_j > 1-\varepsilon^{1/4}.$$

Proof. Suppose not. Then $\sum_{j\in J^c} a_j > \varepsilon^{1/4}$. But then

$$\sum_{j=0}^{q^d-1} a_j(1-w_j^k)\geqslant \sum_{j\in J^c} a_j(1-w_j^k) > \varepsilon^{1/4}\sum_{j\in J^c} a_j > \varepsilon^{1/4}\varepsilon^{1/4}=\sqrt{\varepsilon}.$$

But

$$\sum_{j=0}^{q^d-1} a_j(1-w_j^k)=1-w^k\leqslant\sqrt{\varepsilon}.$$

This is a contradiction. ∎

$S^{-d}\mathscr{F}$ is the sub-algebra of $\mathscr{F}$ obtained by mapping a point $\hat{z}$ to

$$\langle\hat{z}\rangle=\{\hat{y}\colon \varphi(S^d\hat{y})=\varphi(S^d\hat{z})\}.$$

The measure μ can be decomposed into μ restricted to $S^{-d}\mathscr{F}$ which is conjugate to $S^d\mu$, with fiber measures $\mu_{\langle z\rangle}$ which are atomic and supported on $\langle z\rangle$. From (1) the weights on these atoms are precisely the

$$a_j=E_\mu(\hat{\varphi}^{-1}(\hat{\varphi}(\hat{z})+j/q^d)\mid S^{-d}\mathscr{F})(\hat{z}).$$

By definition of $\mathscr{U}$, if $\hat{z} \in \mathscr{U}$ then so is every element in $\langle \hat{z} \rangle$. Thus the set

$$A = \left\{ \langle \hat{z} \rangle : D_\mu \left(S^{-d} \bigvee_{i=-1}^{-n} T^{-i}P \mid S^{-d}\mathscr{F} \right)(\hat{z}) = (1-\sqrt{\varepsilon})\, \vec{e}(S^d\hat{z}) + \sqrt{\varepsilon}\vec{v}(S^d\hat{z}) \right\}$$

is well defined and in $S^{-d}\mathscr{F}$ with measures $S^d\mu(A) = \mu(\mathscr{U})$.

Let

$$B = \left\{ \hat{z}: D_\mu \left(S^{-d} \left(\bigvee_{i=-1}^{-n} T^{-i}P \right) \Bigm| \mathscr{F} \right)(\hat{z}) = (1-\varepsilon^{1/4})\, \vec{e}(S^d\hat{z}) + \varepsilon^{1/4}\vec{v}(S^d\hat{z}) \right\}.$$

Thus B contains the set B' of points with an almost trivial distribution and which arise as one of the preimages of some $\langle \hat{z} \rangle \in A$. So using proposition 6.2 and the fact that B' factors onto A,

$$\mu(B) \geqslant \mu(B') = \int 1_{B'}\, d\mu_{\langle z \rangle}\, dS^d\mu$$

$$> (1-\varepsilon^{1/4}) \int 1_{B'}\, dS^d\mu > (1-\varepsilon^{1/4})^3\,(1-\alpha-\varepsilon).$$

So altogether we have

$$\mu \left\{ \hat{z}: D_\mu \left(S^{-d} \left(\bigvee_{i=-1}^{-n} T^{-i}P \right) \Bigm| \mathscr{F} \right)(\hat{z}) = (1-\varepsilon^{1/4})\, \vec{e}(S^d\hat{z}) + \varepsilon^{1/4}\, \vec{v}(S^d\hat{z}) \right\}$$
$$\geqslant (1-\varepsilon^{1/4})^3\,(1-\alpha-\varepsilon).$$

Recall from Section 3 that there are p_0^n possible preimages in $S^{-(k_{0n}+r)}\bigvee_{i=-1}^{-n} T^{-i}P$, $r \geqslant 0$, given the first quadrant of symbols. If $d = k_{0n} + r$ for some r then the weights associated with the distribution at $S^{-d}(\bigvee_{i=-1}^{-n} T^{-i}P)$ and $S^{-k_{0n}}(\bigvee_{i=-1}^{-n} T_{-i}P)$ are the same and we can write

$$\mu \left\{ \hat{z}: D_\mu \left(S^{-k_{0n}} \left(\bigvee_{i=-1}^{-n} T^{-i}P \right) \Bigm| \mathscr{F} \right)(\hat{z}) = (1-\varepsilon^{1/4})\, \vec{e}(S^d\hat{z}) + \varepsilon^{1/4}\vec{v}(S^d\hat{z}) \right\}$$
$$\geqslant (1-\varepsilon^{1/4})^3\,(1-\alpha-\varepsilon). \qquad (2)$$

If $d < k_{0n}$ then there are more than p_0^n possible preimages in $S^{-d}(\bigvee_{i=-1}^{-n} T^{-i}P)$ given the first quadrant. However, if the distribution vector at the dth row is almost trivial then is certainly almost trivial if the preimages are collapsed as they move up to the k_{0n}^{th} row (with the same error estimates) and statement (2) is still true.

Recall now how ε entered these calculations. We gained one $\varepsilon/2$ when we estimated $\mathscr{F}$ by $\bigvee_{i=0}^{M} T^{-i}P$, but we have since let $M \to \infty$. Another $\varepsilon/2$ was added when we replaced $\hat{\mu}$ by μ_I and the remaining errors came from "Chebyshev" type arguments on these errors. However our conclusion has now eliminated any dependence on I. So if we let $\ell(I) \to \infty$ and then $\varepsilon \to 0$ we can conclude that defining

$$G_n = \left\{\hat{z}\colon D_\mu\left(S^{-k_{0n}}\left(\bigvee_{i=-1}^{-n} T^{-i}P\right) \middle| \mathscr{F}\right)(\hat{z}) \text{ is a trivial vector}\right\},$$

$$\mu(G_n) \geqslant 1 - \alpha \tag{3}$$

By proposition 3.3 these sets are nested $G_{n+1} \subseteq G_n$. Let

$$G = \bigcap_{n=1}^{\infty} G_n.$$

Now

$$\mu(G) = \lim \mu(G_n) \geqslant 1 - \alpha.$$

For $\hat{z} \in G$, $\hat{\varphi}(\hat{z})$ determines the preimages at all positions $(-n, k_{0n})$, $n \geqslant 1$. Proposition 3.4 tells us then that $T^{-1}(G) = G$ and so by (3) $G = \hat{Y}$ μ-a.s. and we conclude that μ-a.s.

$$\mathscr{D}_0 \subseteq \mathscr{F}.$$

7. $\mathscr{D}_a \subseteq \mathscr{F}$

We will prove the inclusion $\mathscr{D}_a \subseteq \mathscr{F}$ by induction on a, having verified it for $a = 0$ in the last section. Now suppose it is true for $a - 1$. We can rewrite the last statement of section 5 as

$$\mu\left\{\hat{z}\colon D_\mu\left(S^{-d}\left(\bigvee_{i=-1}^{-n} T^{-i}P\right) \middle| S^{-d}\mathscr{D}_{a-1}\right)(\hat{z})\right.$$

$$\left. = (1 - \sqrt{\varepsilon})\, \vec{e}(S^d\hat{z}) + \sqrt{\varepsilon}\, \vec{v}(S^d\hat{z})\right\}$$

$$\geqslant (1 - \varepsilon^{1/4})^2(1 - \alpha - \varepsilon).$$

The set of points measured in this statement is the set of $\hat{z}$ such that given $\mathscr{D}_{a-1}(S^d\hat{z})$ then the past for n steps to the left of $S^d\hat{z}$ is almost determined in that μ puts most of its weight on just one of the possible preimages. Again we want to move this statement about the dth row down to the 0th row. This means we will have to compare $\mathscr{D}_{a-1}(S^d\hat{z})$ and

$\mathscr{D}_{a-1}(\hat{z})$. We are considering the preimages in $\bigvee_{i=-1}^{-n} T^{-i}P$ for some arbitrary but fixed n. To make sure these preimages are not swallowed up in $\mathscr{D}_{a-1}(\hat{z})$ we will move it over using the T-invariance of μ. Let k be that integer such that $kn_a \leqslant d < (k+1)n_a$. Then put $r = km_a$. This is how much we will move the preimages over. We know

$$\mu\left\{T^r\hat{z}: D_\mu\left(S^{-d}T^r\left(\bigvee_{i=-1}^{-n} T^{-i}P\right)\middle| T^rS^{-d}\mathscr{D}_{a-1}\right)(T^r\hat{z})\right.$$

$$\left. = (1-\sqrt{\varepsilon})\,\vec{e}(S^d\hat{z}) + \sqrt{\varepsilon}\,\vec{v}(S^d\hat{z})\right\}$$

$$\geqslant (1-\varepsilon^{1/4})^2\,(1-\alpha-\varepsilon).$$

Let $\hat{y} = T^r\hat{z}$, so we can rewrite this as

$$\mu\left\{\hat{y}: D_\mu\left(S^{-d}T^r\left(\bigvee_{i=-1}^{-n} T^{-i}P\right)\middle| T^rS^{-d}\mathscr{D}_{a-1}\right)(\hat{y})\right.$$

$$\left. = (1-\sqrt{\varepsilon})\,\vec{e}(S^dT^{-r}\hat{y}) + \sqrt{\varepsilon}\,\vec{v}(S^dT^{-r}\hat{y})\right\}$$

$$\geqslant (1-\varepsilon^{1/4})^2\,(1-\alpha-\varepsilon).$$

Call the set of $\hat{y}$ being measured here $\mathscr{U}_{a-1}$.

The information we are given here is $T^rS^{-d}\mathscr{D}_{a-1}(\hat{y}) = \mathscr{D}_{a-1}(T^{-r}S^d\hat{y})$. We want to move down to a statement where the given information is $\mathscr{D}_{a-1}(\hat{y})$. Now given $T^rS^{-d}\mathscr{D}_{a-1}(\hat{y})$, the possibilities for position $(0, 0)$ are $\{\hat{\varphi}^{-1}(\hat{\varphi}(\hat{y}) + p^d i/q^r)\}$, which after cancellation we can write as $\{\hat{\varphi}^{-1}(\hat{\varphi})(\hat{y}) + j/b)\}$ for some $b \geqslant 1$. Now $T^rS^{-d}\mathscr{D}_{a-1}$ is the sub-algebra of $\mathscr{D}_{a-1}$ where a point $\hat{y}$ is mapped to a finite class

$$\langle\hat{y}\rangle = \{\hat{\varphi}^{-1}(\hat{\varphi}(\hat{y}) + j/b)\}.$$

The measure μ as in the previous section can be decomposed as $S^d\mu$ on this sub-algebra with finite fibers, and the atomic fiber measures are $\mu_{\langle\hat{y}\rangle}$ with weights

$$E(\hat{\varphi}^{-1}(\hat{\varphi}(\hat{y}) + j/b) \mid T^rS^{-d}\mathscr{D}_{a-1})(\hat{y}).$$

As before the expectation of a preimage given the factor algebra $T^rS^{-d}\mathscr{D}_{a-1}$ at $\hat{y}$ is the weighted average of the expectations of that preimage given $\mathscr{D}_{a-1}$ for all $\hat{w} \in \langle\hat{y}\rangle$. So we can write

$$D_\mu\left(S^{-d}T^r\left(\bigvee_{i=-1}^{-n} T^{-i}P\right)\bigg| T^rS^{-d}\mathscr{D}_{a-1}\right)(\hat{y})$$
$$=\sum_{j=0}^{b-1}\left\{E_\mu(\hat{\phi}^{-1}(\hat{\phi}(\hat{y})+j/b) \mid T^rS^{-d}\mathscr{D}_{a-1})(\hat{y})\right.$$
$$\left.\times D_\mu\left(S^{-d}T^r\left(\bigvee_{i=-1}^{-n} T^{-i}P\right)\bigg| \mathscr{D}_{a-1}\right)(\hat{\phi}^{-1}(\hat{\phi}(\hat{y})+j/b))\right\}.$$

As before, this has the form $\vec{w}=\sum_{j=0}^{b-1} a_j\vec{w}_j$ and we can again show by a 'Chebyshev' type argument that for most j's, $\vec{w}_j$ is 'almost trivial' with error the square root of that for $\vec{w}$ itself. From the definition of $\mathscr{U}_{a-1}$, if $\hat{y}\in\mathscr{U}_{a-1}$ then so are all elements of $\langle\hat{y}\rangle$. Thus the set

$$A=\left\{\langle\hat{y}\rangle: D_\mu\left(S^{-d}T^r\left(\bigvee_{i=-1}^{-n} T^{-i}P\right)\bigg| T^rS^{-d}\mathscr{D}_{a-1}\right)(\langle\hat{y}\rangle)\right.$$
$$\left.(1-\sqrt{\varepsilon})\,\vec{e}(S^dT^{-r}\hat{z})+\sqrt{\varepsilon}\,\vec{v}(S^dT^{-r}\hat{z})\right\}$$

is well defined, is in $T^rS^{-d}\mathscr{D}_{a-1}$ and has measure

$$S^d\mu(A)=\mu(\mathscr{U}_{a-1}).$$

Let

$$B=\left\{\hat{z}: D_\mu\left(S^{-d}T^r\left(\bigvee_{i=-1}^{-n} T^{-i}P\right)\bigg| \mathscr{D}_{a-1}\right)(\hat{z})=(1-\varepsilon^{1/4})\,\vec{e}(\hat{z})+\varepsilon^{1/4}\vec{v}(\hat{z})\right\}.$$

Then B contains the set B' of all points which have an almost trivial distribution with the above error and which arise as one of the elements of $\langle\hat{y}\rangle\in A$. Thus

$$\mu(B)\geqslant\mu(B')=\int 1_{B'}\,d\mu_{\langle\hat{z}\rangle}\,dS^d\mu>(1-\varepsilon^{1/4})^3(1-\alpha-\varepsilon)$$

from proposition 6.2 and the fact that B' factors to A.

Now let $i=n-m_a$. We know from Theorem 3.7 that given the symbols $\mathscr{D}_{a-1}(\hat{z})$ there are $\pi_a^{in_a}\cdots\pi_h^{in_h}$ possibilities at position $(-i,0)$. Move these 'up the staircase' by $S^{n_a}T^{-m_a}$. At $(-i-k_{ai}m_a, k_{ai}n_a)$ these have collapsed to just $\pi_a^{in_a}$ preimages.

B contains those points such that given the symbols $\mathscr{D}_{a-1}(\hat{z})$, the preimage at $(-km_a - n, d)$ is almost determined. Thus so is the preimage at $(-km_a - n, (k+1)n_a)$ (with the same error estimate) which we rewrite as $(-i-(k+1)m_a, (k+1)n_a)$. If $k+1 \geqslant k_{ai}$, then one to one-ness forces the preimage at $(-i-k_{ai}m_a, k_{ai}n_a)$ to also be almost determined with the same error estimate. If $k+1 < k_{ai}$ then the preimages at $(-i-(k+1)m_a, (k+1)n_a)$ will collapse as they move up the staircase to step k_{ai}. If their distribution was almost trivial at step $k+1$ this collapsing can only improve the situation at step k_{ai}. In either case we can write

$$\mu\left\{\hat{z}: D_\mu\left(S^{-k_{ai}n_a}T^{k_{ai}m_a} \bigvee_{j=-1}^{-i} T^{-j}P \mid \mathscr{D}_{a-1}\right)(\hat{z}) = (1-\varepsilon^{1/4})\,\vec{e}(\hat{z}) + \varepsilon^{1/4}\vec{v}(\hat{z})\right\}$$
$$\geqslant (1-\varepsilon^{1/4})^3\,(1-\alpha-\varepsilon).$$

Now let ε go to zero and we have

$$\mu\left\{\hat{z}: D_\mu\left(S^{-k_{ai}n_a}T^{k_{ai}m_a} \bigvee_{j=-1}^{-i} T^{-j}P \mid \mathscr{D}_{a-1}\right)(\hat{z}) \text{ is a trivial vector}\right\} \geqslant 1-\alpha.$$

Call the set being measured here G_i. From Proposition 3.8, these sets are nested, and so set $G = \bigcap G_i$. We conclude that $\mu(G) = \lim \mu(G_i) \geqslant 1-\alpha$. By Proposition 3.9 G is T-invariant and so by ergodicity μ-a.s. $G = \hat{Y}$. Hence for $\mathscr{D}_a = \mathscr{F}\,\mu$-a.s.

Proof of Theorem 1.5. Thus if $\alpha \neq 1$ we inductively conclude $\mathscr{D}_h \subseteq \mathscr{F}$ which is to say that μ-a.s. the first quadrant of symbols determines the second. This is precisely to say that $h_\mu(T) = 0$ conflicting with our original assumption. Hence $\alpha = 1$ and $\hat{\mu} = \lambda$. ∎

8. Conclusions

Theorem 1.2 gives some information concerning both areas mentioned in the introduction, T and S invariant measures, and conditions on T-invariant measures μ that might ensure μ-a.e. point is normal to the base q. In neither case will we give a complete answer to the issue. We will though gain some information.

For the first issue, one possible way to seek a counterexample to Furstenberg's conjecture that the only T and S invariant ergodic measures are either atomic or Lebesgue, would be to select some T-invariant measure μ of zero entropy and consider some limit of averages of the form

$S_{I(k)}\mu$. Our next result says that in order to accomplish this, either μ must be chosen very carefully or the sequence $I(k)$ must be chosen very carefully in that generically for most μ and most such subsequences $I(k)$ the limit is λ.

To set the stage for this, let $(\mathcal{M}_T, w^*)$ be the set of all T-invariant Borel probability measures on Y with w^* some metric giving the weak* topology. Let $\mathcal{M}_T^e \subseteq \mathcal{M}_T$ be those measures which are ergodic for T and let $\mathcal{M}_T^0 \subseteq \mathcal{M}_T^e$ be those of zero entropy for T. It is well known (see for example [R1]) that $\mathcal{M}_T^0$ is a residual subset of $\mathcal{M}_T$.

DEFINITION 8.1. We say a subset $A \subseteq \mathbb{N}$ has full upper density if

$$\overline{\lim_{j \to \infty}} \frac{\#\{a \in A: 0 \leqslant a \leqslant j\}}{j+1} = 1.$$

This is the weakest notion of A being almost every integer where uniform full density is the strongest notion short of cofinite.

THEOREM 8.2. *For a residual set of $\mu \in \mathcal{M}_T$, along a subset $A(\mu)$ of full upper density,*

$$\lim_{\substack{i \to \infty \\ i \in \mathcal{A}}} S^i \mu = \lambda.$$

Proof. Fix an $\varepsilon > 0$ and define

$$\mathcal{O}(n, \varepsilon) = \left\{\mu \in \mathcal{M}_T: \frac{\#\{0 \leqslant i \leqslant n: w^*(S^i\mu, \lambda) < \varepsilon\}}{n+1} > 1 - \varepsilon\right\}.$$

This is w^*-open.

$$\mathcal{O}(\varepsilon) = \bigcup_{n=1}^{\infty} \mathcal{O}(n, \varepsilon)$$

must contain all μ with $h_\mu(T) > 0$ by Theorem 1.2. Hence $\mathcal{O}(\varepsilon)$ is open and dense in $\mathcal{M}_T$. Any μ in the residual set $\bigcap_m \mathcal{O}(1/m)$ must have the desired limiting property. ∎

For the second issue of normalcy to the base q we also will only get weak information. To begin we give a simple lemma about convex sets.

LEMMA 8.3. *Suppose (X, m) is a Banach space with metric m, $X_0 \subset X$ is a compact convex subset and x_0 is an extreme point of X_0.*

Suppose $\{v_i\}$ *is a sequence of Borel probability measures on* X *with*

$$\int m(x, X_0)\, dv_i \xrightarrow{i} 0 \quad \textit{and} \tag{1}$$

$$m\left(\int x\, dv_i, x_0\right) \xrightarrow{i} 0. \tag{2}$$

We conclude

$$\int m(x, x_0)\, dv_i \xrightarrow{i} 0. \tag{3}$$

Proof. Suppose this is false. There is then an $\varepsilon > 0$ and a subsequence $i(k)$ with

$$\lim_{k \to \infty} \int m(x, x_0)\, dv_{i(k)} > \varepsilon.$$

Let $\{a_1, ..., a_t\}$ be an $\varepsilon/2$-dense subset of X_0, and let

$$B_j = \{x: m(x, a_j) < \varepsilon/2\}.$$

Thus $\bigcup B_j$ contains an $\varepsilon/2$ neighborhood of X_0 and

$$\lim_{i \to \infty} v_i\left(\bigcup_j B_j\right) = 1.$$

Set

$$C_j = B_j \cap \left(\bigcup_{k=1}^{j-1} B_k\right)^c,$$

and now the C_j are disjoint sets, each of diameter at most ε and

$$\lim_{i \to \infty} \sum_{j=1}^{t} v_i(C_j) = 1.$$

Choose a subsequence $i(k)$ with

$$v_{i(k)}(C_j) \xrightarrow{k} b_j.$$

Of course $\sum_j b_j = 1$.

For $b_j \neq 0$ consider the sequence of points

$$x_{j,k} = \frac{1}{b_j} \int_{C_j} x\, dv_{i(k)}.$$

From (1) we can choose a subsequence of the k's so that

$$x_{j,k(t)} \xrightarrow{t} x_j \in X_0 \cap \bar{C}_j.$$

From (2), $m(\sum b_j x_j, x_0) = 0$.

As x_0 is an extreme point of X_0 all $x_j = x_0$. Hence for all $b_j \neq 0$ we must have $x_0 \in \bar{C}_j$. We conclude

$$\overline{\lim_{t \to \infty}} \int m(x, x_0)\, dv_{i(k(t))} \leqslant \varepsilon/2,$$

conflicting with our supposition. ∎

Let $(\mathscr{M}, w^*)$ be the space of all signed Borel measures on Y, a Banach space. Let $\mathscr{M}_0$ be the space of S-invariant probability measures, a compact, convex subset. We know $\lambda \in \mathscr{M}$ is an extreme point of $\mathscr{M}_0$.

For any $y \in Y$, let $\delta_i(y)$ be a point mass at $S^i(y)$ and for an interval $I \subset \mathbb{N}$, let

$$\delta_I(y) = \frac{1}{l(I)} \sum_{j \in I} \delta_j(y).$$

Lemma 8.4. *For any sequence $I(k)$ of intervals with $l(I(k)) \xrightarrow{k} \infty$,*

$$\lim_{k \to \infty} w^*(\delta_{I(k)}(y), \mathscr{M}_0) = 0.$$

Proof. The probability measures in $\mathscr{M}$ are a compact convex set. Hence any subsequence of $\delta_{I(k)}(y)$ contains a convergent subsequence. On the other hand, any subsequential limit point of the $\delta_{I(k)}(y)$ must be in $\mathscr{M}_0$. The result follows. ∎

Definition 8.5. A point $y \in Y$ is called normal to the base q if for $I(k) = [0, 1, ..., k]$

$$\lim_{k \to \infty} w^*(\delta_{I(k)}(y), \lambda) = 0.$$

A probability measure $\mu \in \mathscr{M}$ is called normal to the base q if

$$\lim_{k \to \infty} \int w^*(\delta_{I(k)}(y), \lambda)\, d\mu = 0.$$

A probability measure $\mu \in \mathscr{M}$ is called quasi-normal to the base q if along some subsequence $k(i) \to \infty$,

$$\lim_{i \to \infty} \int w^*(\delta_{I(k(i))}(y), \lambda)\, d\mu = 0.$$

This is the standard definition of normalcy of a point. Notice that to say a measure is normal to the base q is not to say that μ-a.e. point is normal as the distance to λ is going to zero in probability, but not necessarily pointwise. Notice that to say y is normal to the base q is the same as to say $\delta_0(y)$ is normal to the base q, and to say μ is normal to the base q does imply that μ-a.e. y is quasi-normal to the base q in that $\delta_0(y)$ is quasi-normal to the base q.

Theorem 8.6. *Suppose p and q are multiplicatively independent.*

(a) *If μ is T_p-invariant and ergodic and $h_\mu(T_p) > 0$ then μ is normal to the base q.*

(b) *For a residual subset of all T_p-invariant measures μ, μ is quasi-normal to the base q.*

Proof. Suppose μ is a T_p invariant measure and $I(k)$ are such that

$$\mu_{I(k)} \xrightarrow{k} \lambda.$$

Define a function

$$\phi_k : Y \to \mathcal{M} \qquad \text{by} \qquad \phi_k(y) = \delta_{I(k)}(y).$$

Let $\nu_k = \varphi_k(\mu)$.

Now

$$\int w^*(m, \mathcal{M}_0)\, d\nu_k = \int w^*(\delta_{I(k)}(y), \mathcal{M}_0)\, d\mu(y) \xrightarrow{k} 0 \tag{1}$$

by Lemma 8.4.

Also

$$w^*\left(\int m\, d\nu_k, \lambda\right) = w^*\left(\int \delta_{I(k)}(y)\, d\mu(y), \lambda\right) = w^*(\mu_{I(k)}, \lambda) \xrightarrow{k} 0. \tag{2}$$

We conclude from Theorem 8.3 that

$$\int w^*(\delta_{I(k)}(y), \lambda)\, d\mu \xrightarrow{k} 0. \tag{3}$$

The two conclusions of the theorem now follow from Theorems 1.2 and 8.2. ▮

In the proofs of all our major results, Theorems 1.2, 8.2, and 8.6, at some point our argument rested on some general fact about convex sets or category. It is in these places that the possible bad subsequences on which convergence cannot be guaranteed creep in. As we have no examples showing such bad sequences can actually exist perhaps it is possible by some more explicit investigation to eliminate these bad subsequences along which convergence to λ fails.

REFERENCES

[F] H. FURSTENBERG, Disjointness in ergodic theory, minimal sets and a problem in Diophantine approximation, *Math. Systems Theory* **1** (1967).

[FS] J. FELDMAN AND M. SMORODINSKY, Normal numbers from independent processes, preprint.

[J] A. JOHNSON, Measures on the circle invariant under multiplication by a nonlacunary subsemigroup of the integers, *Israel J.* **77** (1992), 211–240.

[PK] C. PEARCE AND M. KEANE, On normal numbers, *J. Austral. Math. Soc. Ser. A* **32** (1982), 79–87.

[R] D. J. RUDOLPH, $\times 2$ and $\times 3$ invariant measures and entropy, *Ergodic Theory Dynamical Systems* **10** (1990), 395–406.

[R1] D. J. RUDOLPH, "Fundamentals of Measurable Dynamics," Oxford Univ. Press, New York, 1990.

[S] W. SCHMIDT, On normal numbers, *Pacific J. Math.* **10** (1960), 661–672.

Printed by Catherine Press, Ltd., Tempelhof 41, B-8000 Brugge, Belgium