ROCKY MOUNTAIN
JOURNAL OF MATHEMATICS
Volume 28, Number 3, Fall 1998

MARKOV PARTITIONS FOR HYPERBOLIC TORAL AUTOMORPHISMS OF $\mathbf{T}^2$

E. RYKKEN

ABSTRACT. Using continued fractions, we give a direct and constructive proof for the fact that every matrix in $GL(2,\mathbf{Z})$ whose eigenvalues lie off the unit circle is similar over the integers to a matrix with all nonnegative or all nonpositive entries. This was first proven indirectly by R.F. Williams in 1970 [**8**]. Using this result, we give a constructive proof that there always exists a Markov partition with two connected rectangles for a hyperbolic toral automorphism on the two-dimensional torus.

0. Introduction. Our goal is to construct and study Markov partitions with two connected rectangles for all hyperbolic toral automorphisms on the two-dimensional torus. In their paper, *Similarity of automorphisms of the torus*, [**1**], Adler and Weiss give different constructions for specific cases of these automorphisms. They do not, however, include one for the case when the determinant of the automorphism is positive and the trace is negative. We present this in Section 4. In order to do this, we give a constructive proof in Section 2 that, if $\mathcal{A}$ is a matrix in $GL(2,\mathbf{Z})$ whose eigenvalues lie off the unit circle, then $\mathcal{A}$ is similar over the integers to a matrix with all nonnegative or all nonpositive entries. The proof uses the following fact: given such a matrix, we can consider the convergents of the continued fraction expansion of the slope of the unstable eigenvector as lattice points. Under the map $\mathcal{A}$, they will eventually map to other convergents. We prove this in Section 3. In fact, the desired similarity matrix is given by a consecutive pair of these convergents. Section 1 provides some necessary background. Adler has also continued his work in this area and has different unpublished proofs of the same results.

1. Hyperbolic toral automorphisms. Let $\mathbf{T}^n = \mathbf{R}^n/\mathbf{Z}^n$ be the n-dimensional torus. An automorphism of $\mathbf{T}^n$ is determined by a linear automorphism Φ of $\mathbf{R}^n$ whose matrix has integer entries and

Received by the editors on March 24, 1995, and in revised form on July 10, 1996.
AMS *Mathematics Subject Classification.* 58F15.

determinant equal to ± 1, that is, $\Phi \in GL(n, \mathbf{Z})$. A toral automorphism Φ is called *hyperbolic* if none of the eigenvalues of the matrix has modulus 1, that is, $|\lambda| \neq 1$ for every eigenvalue λ. For definitions, see [**5**].

Let $\mathcal{A} : \mathbf{T}^2 \to \mathbf{T}^2$ be a hyperbolic toral automorphism. Let us call the eigenvalues λ_u and λ_s. They are both real and irrational and satisfy $|\lambda_u| > 1 > |\lambda_s|$. The slope of the unstable eigenvector, m_u, is also real and irrational (since λ_u is).

Lemma 1.1. *Let $\mathcal{A} : \mathbf{T}^2 \to \mathbf{T}^2$ be a hyperbolic toral automorphism. Let $\pi : \mathbf{R}^2 \to \mathbf{T}^2$ be the projection from $\mathbf{R}^2 \to \mathbf{R}^2/\mathbf{Z}^2$. Let $x \in \mathbf{T}^2$. Then $W^u(x)$, the unstable manifold of x, is the projection of a line through $\pi^{-1}x$ parallel to v_u, where v_u is an unstable eigenvector for $\mathcal{A}$. Likewise, $W^s(x)$, the stable manifold of x, is the projection of a line through $\pi^{-1}x$ parallel to v_s, where v_s is a stable eigenvector for $\mathcal{A}$.*

Proof. See [**5**]. □

We are interested in looking at Markov partitions with two rectangles for hyperbolic toral automorphisms of $\mathbf{T}^2$. We would like to define rectangles differently than Bowen [**2**] did in order to allow us to use larger rectangles to partition $\mathbf{T}^2$. Let R be a closed, connected region in $\mathbf{T}^2$, and let $\tilde{R}$ be a closed, connected region in $\mathbf{R}^2$ such that $\pi : \operatorname{int} \tilde{R} \to \operatorname{int} R$ is one-to-one and onto and $\pi : \tilde{R} \to R$ is finite-to-one and onto. Suppose $x \in R$ with $\tilde{x} \in \tilde{R}$ such that $\pi(\tilde{x}) = x$, define $W^u(x, R) = \pi(W^u(\tilde{x}) \cap \tilde{R})$ and $W^s(x, R) = \pi(W^s(\tilde{x}) \cap \tilde{R})$. Note that, if $x \in \partial R$, then $W^u(x, R)$ and $W^s(x, R)$ may depend on the choice of lift for x. Choose a consistent lift. If more than one choice for $W^u(x, R)$ or $W^s(x, R)$ exists, then the rectangle is wrapping around in $\mathbf{T}^2$ and two of its ends meet (see Figure 1). A closed, connected set R is a *rectangle* if $R = \overline{\operatorname{int} R}$ and, given $x, y \in \operatorname{int} R$, then $W^s(x, R) \cap W^u(y, R)$ is exactly one point and this point is in R. This is equivalent to saying that R is a rectangle if R lifts to a parallelogram with sides in the directions of the stable and unstable eigenvectors, $\tilde{R}$, in $\mathbf{R}^2$ such that $\pi : \operatorname{int} \tilde{R} \to \operatorname{int} R$ is one-to-one and onto.

Definition 1.2. A *Markov partition* of $\mathbf{T}^2$ is a finite covering

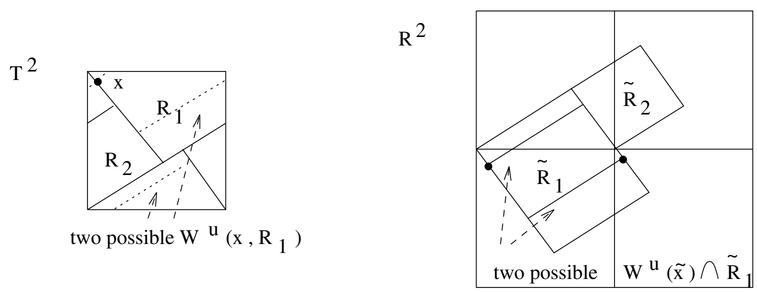


FIGURE 1.

$\{R_1, \dots, R_n\}$ of $\mathbf{T}^2$ by rectangles such that:

1. For $i \neq j$, $\operatorname{int} R_i \cap \operatorname{int} R_j = \varnothing$.
2. If $x \in \operatorname{int} R_i$, $f(x) \in \operatorname{int} R_j$, then $f(W^u(x, R_i)) \supset W^u(f(x), R_j))$ and $f(W^s(x, R_i)) \subset W^s(f(x), R_j))$.

Definition 1.3. We define the *Markov matrix* for a Markov partition $\mathcal{P}$ with n rectangles to be the $n \times n$ matrix given by

$$M_{ij} = \text{the number of times } \operatorname{int} \mathcal{A}(R_j) \text{ crosses } \operatorname{int} R_i$$

for $1 \leq i, j \leq n$.

Proposition 1.4. *Let $\mathcal{P}$ be a Markov partition for $\mathcal{A}$, a hyperbolic toral automorphism, with Markov matrix M. Then $\mathcal{P}$ is a Markov partition for $\mathcal{A}^{-1}$ with Markov matrix M^T.*

Proof. Left to reader. $\square$

Proposition 1.5. *Let $\Phi \in GL(n, \mathbf{Z})$ be such that $\Phi^{-1}\mathcal{A}\Phi = \mathcal{B}$ where $\mathcal{A}$ and $\mathcal{B}$ are hyperbolic toral automorphisms. Then, if $\mathcal{P}$ is a Markov partition for $\mathcal{B}$ with Markov matrix M, $\Phi\mathcal{P}$ will be a Markov partition for $\mathcal{A}$ with Markov matrix M.*

Proof. Left to reader. □

Let R be a rectangle. Define $\partial_u R \equiv \{x \in R : x \notin \text{int } (W^s(x,R))\}$ and $\partial_s R \equiv \{x \in R : x \notin \text{int } (W^u(x,R))\}$. Interior here refers to the interior of $W^s(x,R)$ relative to $W^s_{2\varepsilon}(x)$ and the interior of $W^u(x,R)$ relative to $W^u_{2\varepsilon}(x)$, where ε is chosen such that $W^s(x,R) \subseteq W^s_\varepsilon(x)$ and $W^u(x,R) \subseteq W^u_\varepsilon(x)$. Let $\mathcal{P} = \{R_1, \dots, R_n\}$ be a partition for $\mathbf{T}^2$. Define the *unstable boundary of a partition* $\mathcal{P}$ to be $\partial_u \mathcal{P} = \cup_{i=1}^n \partial_u R_i$ and the *stable boundary of a partition* $\mathcal{P}$ to be $\partial_s \mathcal{P} = \cup_{i=1}^n \partial_s R_i$.

Definition 1.6. A point where $\partial_u \mathcal{P}$ and $\partial_s \mathcal{P}$ intersect is called a *crossing* if the line segments cross each other completely. If they do not, then this point is called an *endpoint*.

Snavely has also done work with finding Markov partitions for hyperbolic toral automorphisms on $\mathbf{T}^2$ with two rectangles. From his thesis [**6**] we have the following proposition.

Proposition 1.7. *If $\mathcal{P}$ is a partition of $\mathbf{T}^2$ with connected rectangles, then the number of rectangles is equal to the number of crossings plus two.*

From the definition of a Markov partition, we have that:

1. $\mathcal{A}(\partial_u \mathcal{P}) \supset \partial_u \mathcal{P}$ and
2. $\mathcal{A}(\partial_s \mathcal{P}) \subset \partial_s \mathcal{P}$.

The following proposition is also known and not difficult to prove.

Proposition 1.8. *If we partition the torus into rectangles such that* $\text{int}\, R_i \cap \text{int}\, R_j = \varnothing$ *if* $i \neq j$, *then in order to show that this partition is a Markov partition, it suffices to show* $\mathcal{A}(\partial_u \mathcal{P}) \supseteq \partial_u \mathcal{P}$ *and* $\mathcal{A}(\partial_s \mathcal{P}) \subseteq \partial_s \mathcal{P}$.

From Proposition 1.5, it is clear that, if we can prove that every hyperbolic toral automorphism on the two-dimensional torus is similar over the integers to a matrix with all nonnegative or all nonpositive entries, then we are free to restrict our attention to such matrices.

2. Conjugacy to a nonnegative or a nonpositive matrix.

Theorem 2.1. *Every $\mathcal{A} \in GL(2, \mathbf{Z})$ whose eigenvalues lie off the unit circle is similar over the integers to a matrix B, all of whose entries have the same sign* (0 *allowed*). *A similarity is given by consecutive convergents of the continued fraction expansion of the slope of the unstable eigenvector.*

In order to prove this, we will need some results about continued fractions. Given any real, irrational number α, we can write α in the form

$$\alpha = a_1 + \cfrac{1}{a_2 + \frac{1}{a_3 + \frac{1}{a_4 + \cdots}}}$$

where $a_1 \in \mathbf{Z}$ and $a_i \in \mathbf{Z}^+$ for $i \geq 2$. This is called the simple continued fraction expansion of α. We can write $\alpha = [a_1, a_2, a_3, \ldots]$. The finite simple continued fraction $[a_1, a_2, \ldots, a_n]$ has a rational value $c_n = (p_n/q_n)$ and is called the nth convergent to α.

Lemma 2.2.

$$p_{i+1}q_i - p_i q_{i+1} = \pm 1.$$

Proof. Left to reader. □

Following [**3**], we say that a fraction p/q, $q > 0$, is a best approximation to a real, irrational number α if, for all fractions p'/q' with $0 < q' \leq q$, $|q\alpha - p| < |q'\alpha - p'|$ unless $q = q'$ and $p = p'$. In his paper, Irwin proves that the best approximations are precisely the nth convergents, where either $n \geq 1$ or $n \geq 2$. Since $q_{i+1} > q_i$ for $i \geq 2$, we have that if p_n/q_n and p_{n+i}/q_{n+i} are convergents with $n \geq 2$, then $|q_{n+i}\alpha - p_{n+i}| < |q_n\alpha - p_n|$ for all $i \in \mathbf{Z}^+$. This also gives us the inequality $|\alpha - p_{n+i}/q_{n+i}| < |\alpha - p_n/q_n|$, so each convergent is nearer to the value of α than the preceding convergent.

Convergents have the following geometric significance. Given a convergent p_n/q_n for α, we can associate it with the lattice point

(q_n, p_n). Consider the line $y = \alpha x$. If we imagine pegs at each of the lattice points and consider two strings lying on $y = \alpha x$ that are fixed at infinity in one direction, then if we pull one string to the right to the first convergent (q_1, p_1) and the other to the left to the second convergent (q_2, p_2), the pegs that are touched by the string pulled to the left are exactly the upper convergents (those greater than α) and the pegs that are touched by the string pulled to the right are exactly the lower convergents (those less than α). This fact was given by F. Klein, *Ausgewaehlte Kapitel der Zahlentheorie*, in 1907. The explanation can be found in Olds [**4**, pp. 77–79].

Consider the simple continued fraction expansion of m_u, the slope of the unstable eigenvector. We need the following theorem which we will prove in the next section.

Theorem 2.3. *Let $\mathcal{A}$ be an element of $GL(2, \mathbf{Z})$ such that the eigenvalues of $\mathcal{A}$ lie off the unit circle and such that $\operatorname{tr}\mathcal{A} > 0$. Let p_n/q_n be the convergents for m_u. Then there is an $M \in \mathbf{Z}^+$ such that, if $m \geq M$, then $\mathcal{A}\begin{bmatrix} q_m \\ p_m \end{bmatrix}$ corresponds, in the manner described above, to another convergent of m_u, p_{m+i}/q_{m+i} for some $i \geq 1$. There is also an $\tilde{M} \in \mathbf{Z}^+$ such that $\mathcal{A}^{-1}\begin{bmatrix} q_m \\ p_m \end{bmatrix}$ is another convergent for all $m \geq \tilde{M}$.*

Using these results, we are able to prove the theorem.

Proof of Theorem 2.1. If $\operatorname{tr}\mathcal{A} < 0$, then $P^{-1}\mathcal{A}P$ has all nonpositive entries if and only if $P^{-1}(-\mathcal{A})P$ has all nonnegative entries. Since $\det(-\mathcal{A}) = \det\mathcal{A}$, and $\operatorname{tr}(-\mathcal{A}) = -\operatorname{tr}\mathcal{A}$, if we prove the case when $\operatorname{tr}\mathcal{A} > 0$, then the case when $\operatorname{tr}\mathcal{A} < 0$ will follow. Assume $\operatorname{tr}\mathcal{A} > 0$, in which case we will have $\lambda_u > 0$.

We want to find $P \in GL(2, \mathbf{Z})$ such that $P^{-1}\mathcal{A}P$ has all nonnegative entries, that is, $P^{-1}\mathcal{A}P$ (first quadrant) $\subseteq$ (first quadrant) or $\mathcal{A}P$ (first quadrant) $\subseteq P$ (first quadrant). $P = \begin{bmatrix} q & \tilde{q} \\ p & \tilde{p} \end{bmatrix}$ with $q\tilde{p} - p\tilde{q} = \pm 1$ can be thought of as a sector in the plane that is bounded by two rays that originate at the origin and pass through (q, p) and $(\tilde{q}, \tilde{p})$ and hence have rational slopes p/q and $\tilde{p}/\tilde{q}$, respectively. The first quadrant can thus be represented by the identity matrix, hence P (first quadrant) $= P$ and we want $\mathcal{A}P \subseteq P$ where P and $\mathcal{A}P$ are thought of as sectors. So,

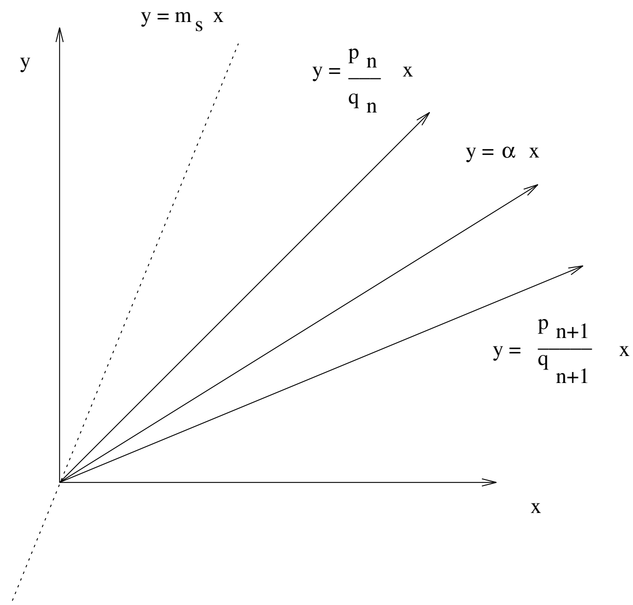


FIGURE 2.

if we can find such a sector that maps into itself under $\mathcal{A}$, then we will be done. Consider an unstable eigenvector that lies in the $x > 0$ half plane; call it v_u. Let v_s be the stable eigenvector with slope m_s. The line $y = m_s x$ divides the plane into two halves. We need a sector that contains v_u and lies completely within one of the half planes determined by $y = m_s x$. This is necessary since points in this sector are a linear combination of v_u and v_s with a positive coefficient for v_u. Under $\mathcal{A}$ the component in the unstable direction will be stretched by $\lambda_u > 1$ and the component in the stable direction will be shrunk by $\lambda_s = \pm 1/\lambda_u$. In order to pick p/q and $\tilde{p}/\tilde{q}$, consider the convergents to $m_u =$ slope of v_u. Since $|\alpha - p_{i+1}/q_{i+1}| < |\alpha - p_i/q_i|$ for every $i \geq 1$, we can find consecutive convergents p_n/q_n and p_{n+1}/q_{n+1} such that the rays that originate from the origin and pass through (q_n, p_n) and (q_{n+1}, p_{n+1}) lie completely within the half plane determined by $y = m_s x$ that contains v_u. Moreover, since the convergents are consecutive, the rays will lie on opposite sides of v_u, hence the sector they form will contain v_u. (See Figure 2.) By Lemma 2.2, we have $q_{n+1}p_n - p_{n+1}q_n = \pm 1$. Moreover, by Theorem 2.3, there is an $M \in \mathbf{Z}^+$ such that

$$\mathcal{A}\begin{bmatrix} q_n \\ p_n \end{bmatrix} = \begin{bmatrix} q_{n+i} \\ p_{n+i} \end{bmatrix}$$

is another convergent of α for some $i \geq 1$ for every $n \geq M$. Since $|\alpha - p_{n+i}/q_{n+i}| \leq |\alpha - p_{n+1}/q_{n+1}|$, we have that $\mathcal{A}P \subseteq P$, thus

$$P = \begin{bmatrix} q & q_{n+1} \\ p_n & p_{n+1} \end{bmatrix}$$

will satisfy our requirements and $P^{-1}\mathcal{A}P$ will have all nonnegative entries. □

3. Convergents will eventually map to other convergents. In order to prove Theorem 2.3 we will need the following background.

Notice that the distance $|q\alpha - p|$ can be thought of as the vertical distance from the point (q, p) to the point $(q, \alpha q)$. We also have the following theorem from Stark [**7**, p. 214].

Theorem 3.1. *Suppose that α is irrational and p_n/q_n and p_{n-1}/q_{n-1} are consecutive convergents that satisfy $0 < q_{n-1} < q_n$ (this is always true if $n \geq 3$). If (q, p) is a lattice point that is not one of the lattice points associated with these convergents and $0 < q \leq q_n$, then the vertical distances of (q, p) and (q_{n-1}, p_{n-1}) from the line $y = \alpha x$ satisfy the inequality $|q_{n-1}\alpha - p_{n-1}| < |q\alpha - p|$.*

In addition, we have the following two lemmas.

Lemma 3.2. *Let $\mathcal{A} \in GL(2, \mathbf{Z})$ with eigenvalues that lie off the unit circle and $\operatorname{tr} \mathcal{A} > 0$. Let $\alpha = m_u$. Consider a convergent p/q for α. Let*

$$\mathcal{A}\begin{bmatrix} q \\ p \end{bmatrix} = \begin{bmatrix} q' \\ p' \end{bmatrix}.$$

Then p' and q' are relatively prime.

Proof. Given a convergent p/q we know by Lemma 2.2 that $(p, q) = 1$. Hence, if we consider $y = (p/q)x$, (q, p) must be the closest integer lattice point to the origin that $y = (p/q)x$ passes through. Now $\mathcal{A}$ maps the line $y = (p/q)x$ to the line $y = (p'/q')x$. Thus (q', p') must be the closest integer lattice point to the origin that $y = (p'/q')x$ passes through. If not, then there is another point $(\tilde{q}, \tilde{p})$ that is closer, but then $\mathcal{A}^{-1}\begin{bmatrix}\tilde{q}\\ \tilde{p}\end{bmatrix}$ would be closer to the origin than (q, p) is on the line $y = (p/q)x$ and this would be a contradiction. Hence, $(p', q') = 1$. ◻

Lemma 3.3. *Let $\mathcal{A} \in GL(2, \mathbf{Z})$ with eigenvalues that lie off the unit circle and $\operatorname{tr}\mathcal{A} > 0$. Let $\alpha = m_u$. Consider the convergents p_m/q_m for α. Let*

$$\mathcal{A}\begin{bmatrix} q_m \\ p_m \end{bmatrix} = \begin{bmatrix} q'_m \\ p'_m \end{bmatrix}.$$

Then there is an $M \in \mathbf{Z}^+$ such that $q'_{m+1} > q'_m > q_m$ if $m \geq M$, that is, the order of the x-coordinates of consecutive convergents will be preserved by their images under $\mathcal{A}$, and the image of the x-coordinate will be greater than the x-coordinate.

Proof. The convergents can be written as a linear combination of v_u and v_s, that is, $(q_n, p_n) = a_n v_u + b_n v_s$. The line $y = m_s x$ divides the plane into halves. Consider the half plane that contains v_u where v_u lies in the $x > 0$ plane. Since the convergents have the property that $q_{n+1} > q_n$ for $n \geq 2$ and $|\alpha - p_{n+1}/q_{n+1}| < |\alpha - p_n/q_n|$, after some N_1 all the convergents will lie in this half plane and hence $a_n > 0$ for $n \geq N_1$. Furthermore, by similar triangles, since $|\alpha q_{n+1} - p_{n+1}| < |\alpha q_n - p_n|$, $|b_{n+1}| < |b_n|$ for all n, see Figure 3. Hence, the b_n are bounded in absolute value. Let $x(v_s)$ denote the x-coordinate of v_s and $x(v_u)$ denote the x-coordinate of v_u. We know that $\lim_{n\to\infty} q_n = \infty$, and $q_n = a_n x(v_u) + b_n x(v_s)$. Since the b_n's are bounded, we must have $\lim_{n\to\infty} a_n = \infty$. We have $(q'_n, p'_n) = a_n\lambda_u v_u + b_n\lambda_s v_s$. We would like to show that $a_n\lambda_u x(v_u) + b_n\lambda_s x(v_s) > a_n x(v_u) + b_n x(v_s)$, that is, $a_n(\lambda_u - 1)x(v_u) + b_n(1 - \lambda_s)(-x(v_s)) > 0$. But $\lim_{n\to\infty} a_n = \infty$, $\lambda_u > 1$, $x(v_u) > 0$, and the b_n are bounded so this is clearly true after some M_1. Next we would like to show that $a_{n+1}\lambda_u x(v_u) + b_{n+1}\lambda_s x(v_s) > a_n\lambda_u x(v_u) + b_n\lambda_s x(v_s)$. It will suffice to show that $(a_{n+1} - a_n)\lambda_u x(v_u) > (|b_{n+1}\lambda_s| + |b_n\lambda_s|)x(v_s)$. Since $q_{n+1} - q_n \geq q_{n-1}$, we know that $\lim_{n\to\infty}(q_{n+1} - q_n) = \infty$. Now $q_{n+1} - q_n = (a_{n+1} - a_n)x(v_u) + (b_{n+1} - b_n)x(v_s)$ and since b_{n+1} and

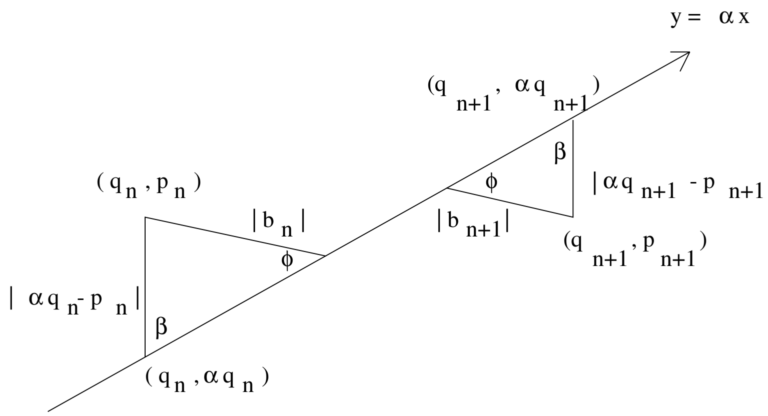


FIGURE 3.

b_n are bounded so is their difference, hence $\lim_{n\to\infty}(a_{n+1} - a_n) = \infty$. Because $\lambda_u > 1$, and $x(v_u) > 0$, this shows that the inequality will hold after some M_2. Choose $M = \max\{M_1, M_2\}$. $\square$

We now proceed with the proof of Theorem 2.3.

Proof of Theorem 2.3. We will first show that eventually the inverse images of the convergents are convergents. Consider the region, R, of the plane determined by the segments connecting the upper convergents, the segments connecting the lower convergents, the vertical segment from the first convergent (q_1, p_1) to $y = \alpha x$, the vertical segment from the second convergent (q_2, p_2) to $y = \alpha x$, and the part of the line $y = \alpha x$ that connects $(q_1, \alpha q_1)$ to $(q_2, \alpha q_2)$ (see Figure 4). By Klein's observation with the strings, the interior of this region contains no lattice points. Consider its image under $\mathcal{A}$. Since lattice points and only lattice points map to lattice points, its image must also contain no lattice points.

We can write $(q_n, p_n) = a_n v_u + b_n v_s$, where the b_n's for the lower convergents all have the same sign and the b_n's for the upper convergents all have the opposite sign. Since $(q_n', p_n') = a_n\lambda_u v_u + b_n\lambda_s v_s$, we

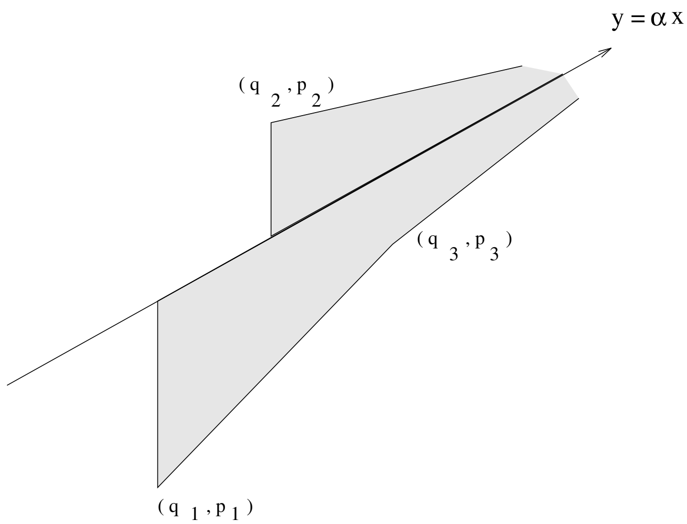


FIGURE 4.

have the images of all the lower convergents lie on one side of the line $y = \alpha x$ and the images of all the upper convergents lie on the opposite side.

Consider the trapezoid formed by $(q_1, p_1), (q_3, p_3), (q_1, \alpha q_1)$ and $(q_3, \alpha q_3)$. Its image will be another trapezoid, with the segment from $(q_1, \alpha q_1)$ to $(q_3, \alpha q_3)$ mapping to another segment on the line $y = \alpha x$, the two vertical segments mapping to two parallel segments, the segment between (q_1, p_1) and (q_3, p_3) mapping to the segment between their images and the interior mapping to the interior of the new trapezoid (see Figure 5). This is true for all trapezoids formed this way by two consecutive lower convergents or two consecutive upper convergents. The trapezoid formed by (q_n, p_n) and (q_{n+2}, p_{n+2}) and the one formed by (q_{n+2}, p_{n+2}) and (q_{n+4}, p_{n+4}) will share the vertical segment from (q_{n+2}, p_{n+2}) to $(q_{n+2}, \alpha q_{n+2})$; hence their images will share the image of this segment. Since the trapezoid maps to another trapezoid and $x(\mathcal{A}(q_n, \alpha q_n)) = \lambda_u q_n < \lambda_u q_{n+2} = x(\mathcal{A}(q_{n+2}, \alpha q_{n+2}))$, we have that $q'_n < q'_{n+2}$ for every $n \geq 1$. Furthermore, since the images of adjacent trapezoids share a boundary, there exist $Q > 0$ such that the image of our region will completely contain the vertical line segment from the boundary to the line $y = \alpha x$ for segments of vertical lies $x = q$

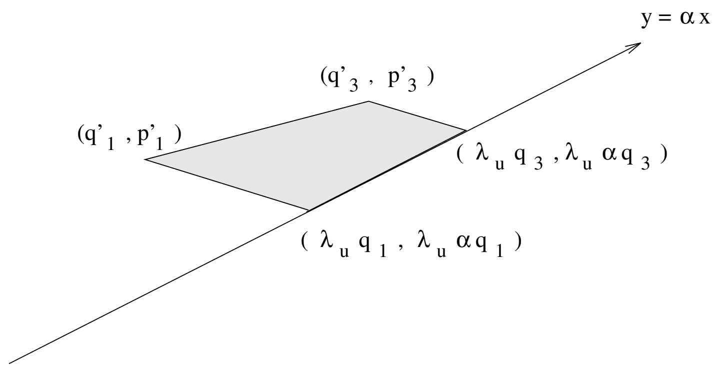


FIGURE 5.

where $q \geq Q$.

By Lemma 3.3, after some N the image of the convergents (q_n, p_n) with $n \geq N$ will have the property that they lie further to the right than their preimage and, given two convergents (q_n, p_n) and (q_{n+i}, p_{n+i}), the order of their x coordinate will be preserved by their images under $\mathcal{A}$ if $n \geq N$.

Suppose (q, p) is an upper convergent with $q \geq \max\{q'_2, q'_N, Q\} = \tilde{M}$; then $\mathcal{A}^{-1}(q, p)$ is another convergent. This can be seen as follows. Suppose not. Since $q > q'_2$, q will lie between the x-coordinates of the images of either two consecutive lower convergents or two consecutive upper convergents. Pick the images that lie above the line $y = \alpha x$ and call them (q'_n, p'_n) and (q'_{n+2}, p'_{n+2}). (q'_n, p'_n) and (q'_{n+2}, p'_{n+2}) are lattice points and hence they must lie outside or on the boundary of our original region, R. Since $q > q'_N$, we have $q'_n \geq q'_N \geq q_2$. By Klein's string argument, the line segment joining (q'_n, p'_n) and (q'_{n+2}, p'_{n+2}) will lie outside or on the boundary of our region, R. Moreover, since (q, p) is a vertex of R, the segment between (q'_n, p'_n) and (q'_{n+2}, p'_{n+2}) will not contain (q, p). This segment will, however, be on the boundary of the image of R. Finally, since $q \geq Q$, the vertical line segment of the line $x = q$ from the boundary of the image of the region to the line $y = \alpha x$ will contain (q, p) and be completely contained in the image of the region (see Figure 6). Since the interior of the image cannot contain any lattice points, this is a contradiction. Thus $\mathcal{A}^{-1}(q, p)$ must

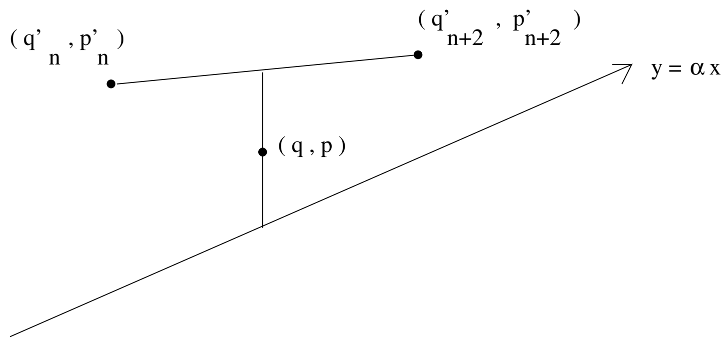


FIGURE 6.

be a convergent. By a similar argument, the same will be true if (q,p) is a lower convergent with $q \geq \max\{q'_1, q'_N, Q\}$.

By Lemma 3.3 there exists a $Q^* \in \mathbf{Z}^+$ such that, if (q',p') and $(\tilde{q}',\tilde{p}')$ are images of convergents with $q' > \tilde{q}'$, then $x(\mathcal{A}^{-1}(q',p')) > x(\mathcal{A}^{-1}(\tilde{q}',\tilde{p}'))$ for every $q', \tilde{q}' \geq Q^*$. This can be seen as follows. There are only a finite number of convergents such that $q'_{m+1} \leq q'_m$, at most the set $\{(q_1,p_1),\ldots,(q_{N-1},p_{N-1})\}$. If we take $Q^* > \max\{q'_1, q'_2, \ldots, q'_{N-1}\}$, then we are done. Let $Q^* = q'_{N+1}$ since $q'_{N+1} > q'_N$ and $q'_{n+2} > q'_n$ for all n, this satisfies our condition.

Let (q_F, p_F) be the first convergent with $q_F \geq \max\{q'_2, q'_{N+1}, Q\}$; then $\mathcal{A}^{-1}(q_f, p_f)$ is another convergent for every $f \geq F$. Let $\mathcal{A}^{-1}(q_F,p_F) = (q_M, p_M)$; then

$$\mathcal{A}\begin{bmatrix} q_m \\ p_m \end{bmatrix} = \begin{bmatrix} q'_m \\ p'_m \end{bmatrix}$$

must be another convergent for every $m \geq M$. This can be seen as follows. Suppose not; then there is an $\tilde{m} \geq M$ such that $(q'_{\tilde{m}}, p'_{\tilde{m}})$ is not a convergent. Consider the convergent (q_L, p_L) such that $q_L < q'_{\tilde{m}} < q_{L+1}$. Since $q_F \geq q'_{N+1} = Q^*$, we have $\tilde{m} \geq M \geq N+1$, and hence $q'_{\tilde{m}} > q_F$, and thus $(q_L,p_L) = \mathcal{A}(q_K,p_K)$ for some convergent (q_K,p_K) with $K < \tilde{m}$ (since $q'_{\tilde{m}} > q_L \geq q_F \geq Q^*$). Thus, $|q_K\alpha - p_K| > |q_{\tilde{m}}\alpha - p_{\tilde{m}}|$. Consider the triangle formed by a convergent $(q,p), (q,\alpha q)$, and a segment from (q,p) to the line $y = \alpha x$ in the direction of the stable eigenvector (this will intersect $y = \alpha x$ since the slope of the stable eigenvector is $(\lambda_s - a)/b$ which is not equal to α), and the one formed by the convergent's image $(q', p'), (q', \alpha q')$, and a segment from (q', p') to the line $y = \alpha x$ in the direction of the stable eigenvector. These two triangles are similar. Since the side in the first triangle that lies in the stable direction must contract by $|\lambda_s| = 1/|\lambda_u|$ under $\mathcal{A}$, all sides of the second triangle must have length $1/|\lambda_u|$ times the length of their corresponding sides in the first triangle. Thus, $|q'\alpha - p'| = (1/\lambda_u)|q\alpha - p|$. Hence, $|q'_K\alpha - p'_K| = |q_L\alpha - p_L| > |q'_{\tilde{m}} - p'_{\tilde{m}}|$. But this contradicts Theorem 3.1. Hence (q'_m, p'_m) must be a convergent for every $m \geq M$. Moreover, since $|q'_m\alpha - p'_m| = (1/\lambda_u)|q_m\alpha - p_m| < |q_m\alpha - p_m|$, for every m, we must have $p'_n/q'_n = p_{n+i}/q_{n+i}$ for some $i \geq 1$. □

4. Markov partitions with two rectangles.

Theorem 4.1. *Let $\mathcal{A} : \mathbf{T}^2 \to \mathbf{T}^2$ be a hyperbolic toral automorphism. Then there exists a Markov partition for $\mathcal{A}$ with two rectangles.*

Proof. By Theorem 3.1, we know that $\mathcal{A}$ is conjugate by $\Phi \in GL(2, \mathbf{Z})$ to a matrix with all nonnegative or all nonpositive entries. By Proposition 1.5, it suffices to find a Markov partition with two rectangles for such matrices. Hence, without loss of generality, assume either $\mathcal{A} \geq 0$ or $\mathcal{A} \leq 0$. If $\mathcal{A} \geq 0$, then by the Perron-Frobenius theorem, we have $m_s < 0 < m_u$. If $\mathcal{A} \leq 0$, then $\mathcal{A}^2 \geq 0$ and we are back with the previous case. We will consider four cases. When $\operatorname{tr}\mathcal{A} < 0$, we may assume $\mathcal{A} \leq 0$. In this case, let $\mathcal{A} = \begin{bmatrix} -a & -b \\ -c & -d \end{bmatrix}$ where $a, b, c, d \geq 0$. When $\operatorname{tr}\mathcal{A} > 0$, we may assume $\mathcal{A} \geq 0$. In this case, let $\mathcal{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ where, again, $a, b, c, d \geq 0$. Given $\tilde{x}, \tilde{y} \in R^2$, let $[\tilde{x}, \tilde{y}]$ denote the unique point $W^s(\tilde{x}) \cap W^u(\tilde{y})$.

Case 1. $\det\mathcal{A} = -1$, $\operatorname{tr}\mathcal{A} < 0$. In this case we have $\lambda_u < -1 < 0 < \lambda_s < 1$. Consider the partition of $\mathbf{T}^2$ shown in Figure 7. We obtain this partition by considering two segments in $\mathbf{R}^2$: one from $(0,0)$ to $[(0,0), (-1,1)]$, which lies in the stable direction and one from $[(1,0), (0,0)]$ to $[(0,-1), (0,0)]$, which lies in the unstable direction. We then project these segments to $\mathbf{T}^2$. The projection of the first segment will be $\partial_s\mathcal{P}$, and the projection of the next segment will form $\partial_u\mathcal{P}$.

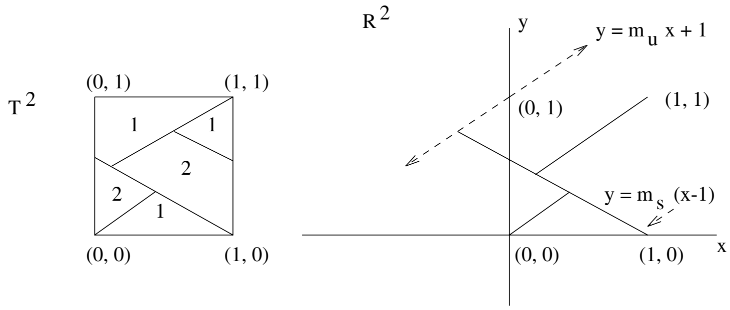


FIGURE 7.

Note that the ends of $\partial_u \mathcal{P}$ lie in $\partial_s \mathcal{P}$. While it is clear that the ends of the stable segment lie in the unstable manifold of the origin, it is also true that the ends of the stable segment lie in $\partial_u \mathcal{P}$. This can be seen by considering Figure 8, where it is clear that in both cases $d' < d$ since $m_s < 0 < m_u$. Also note that the unstable segments and the stable segments do not cross since, considering these segments through any

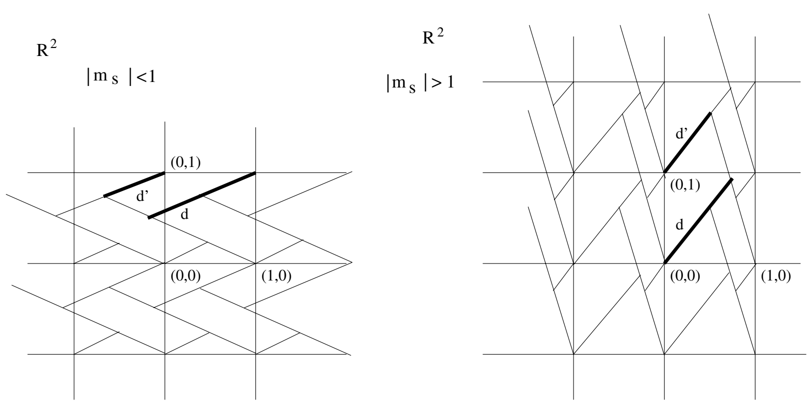


FIGURE 8.

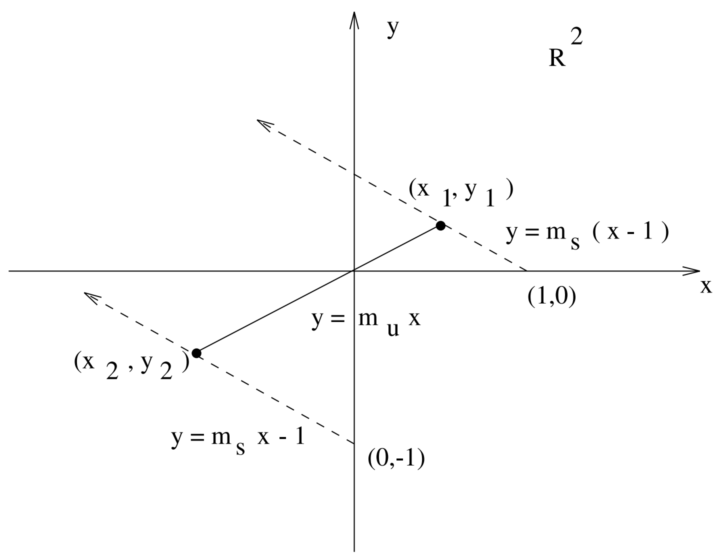


FIGURE 9.

other lattice points does not produce any new crossings (again, since $m_s < 0 < m_u$). Hence, by Proposition 1.7, this partition forms two rectangles on $\mathbf{T}^2$.

Since the origin is a fixed point and the stable segment will contract by $1 > \lambda_s > 0$, we will have $\mathcal{A}(\partial_s \mathcal{P}) \subset \partial_s \mathcal{P}$. Again, since the origin is a fixed point, if we show that $\mathcal{A}(1,0) = (-a,-c)$ satisfies the inequality $y \leq m_s x - 1$ and $\mathcal{A}(0,-1) = (b,d)$ satisfies the inequality $y \geq m_s(x-1)$, then we will have shown that $\partial_u \mathcal{P} \subset \mathcal{A}(\partial_u \mathcal{P})$ (see Figure 9). The first inequality can be written as $c \geq 1 + a m_s$ which clearly holds since $m_s < 0$ and $c \geq 1$ ($b = 0$ or $c = 0$ would imply that the eigenvalues are the integers a and d). The second inequality can be written $d \geq m_s(b-1)$ which also clearly holds since $m_s < 0$ and $b - 1 \geq 0$. Hence, this partition is a Markov partition.

Case 2. $\det \mathcal{A} = -1$, $\operatorname{tr} \mathcal{A} > 0$. In this case we have $\lambda_s < 0 < \lambda_u$. Consider $\mathcal{A}^{-1}$. $\mathcal{A}^{-1}$ has $\lambda_u < 0 < \lambda_s$, $\operatorname{tr}(\mathcal{A}^{-1}) < 0$, $\det(\mathcal{A}^{-1}) = -1$.

Now $\mathcal{A}^{-1}$ is conjugate to an integer matrix $\mathcal{M}$ with all nonpositive entries, that is, $\mathcal{M} = \Phi^{-1}(\mathcal{A}^{-1})\Phi$ for some $\Phi \in GL(2, \mathbf{Z})$. From Case 1, we have a Markov partition $\mathcal{P}$ for $\mathcal{M}$, hence by Proposition 1.5, $\Phi(\mathcal{P})$ will be a Markov partition for $\mathcal{A}^{-1}$. By Proposition 1.4, $\Phi(\mathcal{P})$ will also be a Markov partition for $\mathcal{A}$. We could also directly construct a partition as in Case 1.

Case 3. $\det \mathcal{A} = 1$, $\operatorname{tr} \mathcal{A} > 0$. We can use the construction in Case 1 to partition $\mathbf{T}^2$ into two rectangles. Since the origin is a fixed point and the unstable segments will expand by $\lambda_u > 1$ and the stable segments will contract by $1 > \lambda_s > 0$, we will have $\mathcal{A}(\partial_u \mathcal{P}) \supseteq \partial_u \mathcal{P}$ and $\mathcal{A}(\partial_s \mathcal{P}) \subseteq \partial_s \mathcal{P}$, as desired, and the partition will be Markov.

Case 4. $\det \mathcal{A} = 1$, $\operatorname{tr} \mathcal{A} < 0$. In this case we have $\lambda_u < -1 < \lambda_s < 0$. Since both eigenvalues are negative, the previous partition will fail to partition the 2-torus into two rectangles. By the Lefschetz fixed point theorem, the sum of the indices of the fixed points of $\mathcal{A}$ is equal to the alternating sum of the traces on homology, in this case, $2 - \operatorname{tr} \mathcal{A} \geq 5$. This implies that there exists a fixed point other than the origin. Let (p, q) be a fixed point that is not the origin with $0 \leq p$, $q < 1$, then

$$\begin{bmatrix} -a & -b \\ -c & -d \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix} = \begin{bmatrix} p + m \\ q + n \end{bmatrix}$$

for some $m, n \in \mathbf{Z}$ where $m \leq -1$ and $n \leq -1$.

Consider the following partition of $\mathbf{T}^2$ shown in Figure 10. We obtain this partition by considering a segment in $\mathbf{R}^2$ from $[(p, q), (1, 0)]$ to $[(p, q), (0, 1)]$ (note that this segment has slope m_s), and a segment in the direction of v_u from $[(p - 1, q - 1), (0, 0)]$ to $[(p, q), (0, 0)]$. We then project these segments to $\mathbf{T}^2$. The projection of the first segment will be $\partial_s \mathcal{P}$ and the projection of the next segment will form $\partial_u \mathcal{P}$. Note that the ends of $\partial_u \mathcal{P}$ lie in $\partial_s \mathcal{P}$. It is also true that the ends of the stable segment lie in $\partial_u \mathcal{P}$. This can be seen in Figure 11, where we consider the four possible ways in which the stable segment can cross through the unit square given that $m_s < 0$.

Also note that the unstable and stable segments do not cross, since considering the unstable segments through any other lattice points or the stable segments through translates of (p, q) does not produce any

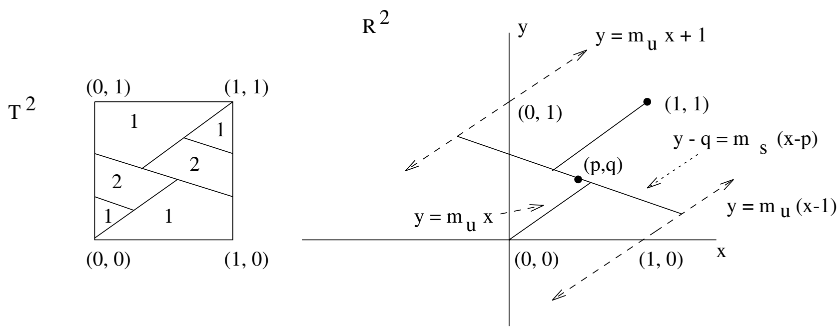


FIGURE 10.

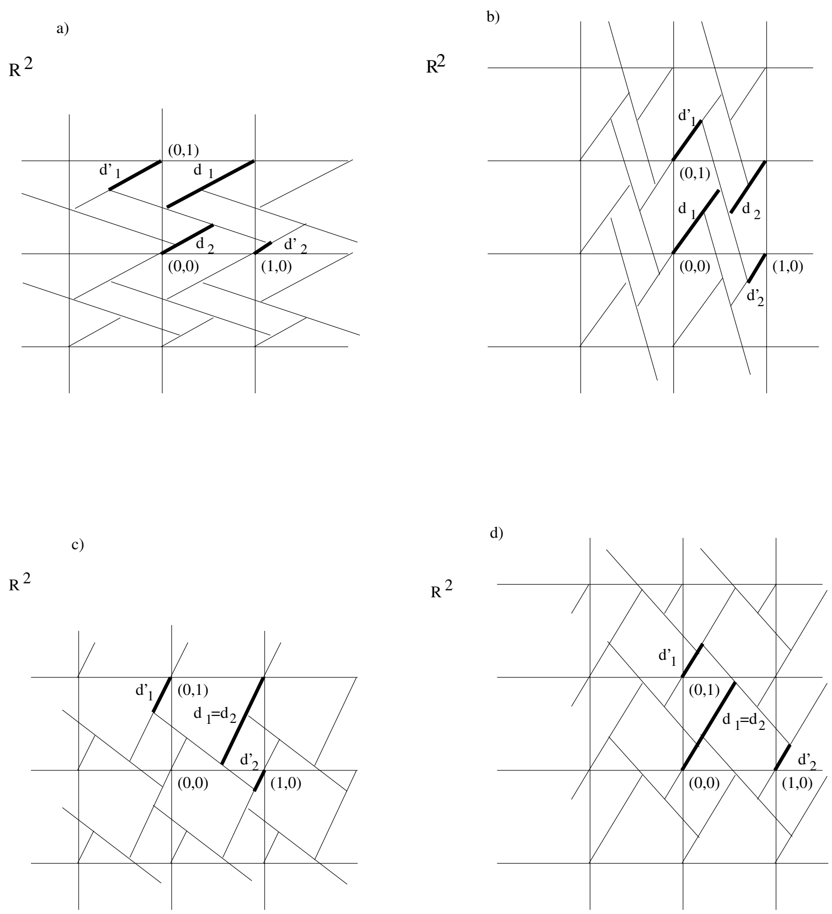
new crossings (again, since $m_u > 0 > m_s$). Hence this partitions $\mathbf{T}^2$ into two rectangles.

Consider the image of the segment from $(0,0)$ to $[(p,q),(0,0)]$; call it $(\partial_u\mathcal{P})^+$. Next, consider the image of the segment from $(0,0)$ to $[(p-1,q-1),(0,0)]$; call it $(\partial_u\mathcal{P})^-$. Since the origin is a fixed point and $\lambda_u < 0$, we must show that $\mathcal{A}(\partial_u\mathcal{P})^+ \supseteq (\partial_u\mathcal{P})^-$ and $\mathcal{A}(\partial_u\mathcal{P})^- \supseteq (\partial_u\mathcal{P})^+$. If we can show that $\mathcal{A}(p,q) = (p+m,q+n)$ satisfies the inequality $y \leq m_s(x-(p-1))+(q-1)$, then all points on $y-(q+n) = m_s(x-(p_m))$ will satisfy it and we will have $\mathcal{A}(\partial_u\mathcal{P})^+ \supseteq (\partial_u\mathcal{P})^-$. So we must show that $q+n \leq m_s(p+m-(p-1))+q-1$, that is, $n+1 \leq m_s(m+1)$. Since $n+1 \leq 0$ and $m+1 \leq 0$, we have $m_s(m+1) \geq 0$, hence the inequality holds.

Now $(1-p,1-q)$ is also a fixed point for $\mathcal{A}$ since

$$\begin{bmatrix} -a & -b \\ -c & -d \end{bmatrix} \begin{bmatrix} 1-p \\ 1-q \end{bmatrix} = \begin{bmatrix} (1-p)-(m+a+b+1) \\ (1-q)-(n+c+d+1) \end{bmatrix}.$$

This means that we could have used $(1-p,1-q)$ as our original fixed point. But then the length of the new $(\partial_u\mathcal{P})^-$ would be the length of the old $(\partial_u\mathcal{P})^+$, and the length of the new $(\partial_u\mathcal{P})^+$ would be the length of the old $(\partial_u\mathcal{P})^-$. By what we have shown above, $|\lambda_u|$ times the length of the new $(\partial_u\mathcal{P})^+$ is greater than or equal to the length of the new $(\partial_u\mathcal{P})^-$. This implies that $|\lambda_u|$ times the length of the old $(\partial_u\mathcal{P})^-$ is



a)
R 2
(0,1)
d' 1
d 1
d 2
d' 2
(0,0)
(1,0)
b)
R2
d'1
(0,1)
d 1
d 2
(0,0)
(1,0)
d'2
c)
R 2
d'1
(0,1)
d 1=d2
d'2
(0,0)
(1,0)
d)
R 2
d'1
(0,1)
d 1=d2
d'2
(0,0)
(1,0)

FIGURE 11.

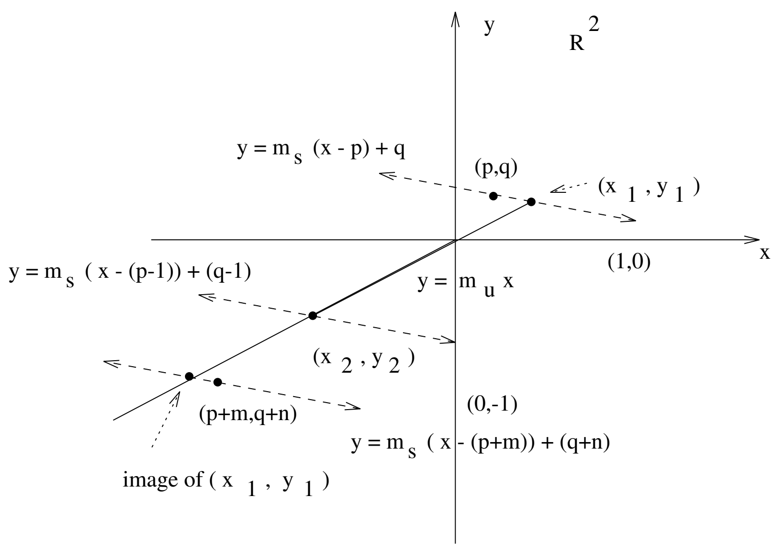


FIGURE 12.

greater than or equal to the length of the old $(\partial_u\mathcal{P})^+$. This shows that $\mathcal{A}(\partial_u\mathcal{P})^- \supseteq (\partial_u\mathcal{P})^+$.

Next consider the image of the segment from (p,q) to $[(p,q),(1,0)]$; call it $(\partial_s\mathcal{P})^+$, and consider the image of the segment from (p,q) to $[(p,q),(0,1)]$; call it $(\partial_s\mathcal{P})^-$. Since (p,q) is a fixed point and $\lambda_s < 0$, we must show that $\mathcal{A}(\partial_s\mathcal{P})^+ \subseteq (\partial_s\mathcal{P})^-$ and $\mathcal{A}(\partial_s\mathcal{P})^- \subseteq (\partial_s\mathcal{P})^+$.

If we can show that $\mathcal{A}(1,0) = (-a,-c)$ satisfies the inequality $y \le m_u x + 1$, then all points on $y = m_u(x+a) - c$ will satisfy it and we will have $\mathcal{A}(\partial_s\mathcal{P})^+ \subseteq (\partial_s\mathcal{P})^-$. We must show $-c \le m_u(-a) + 1 = 1 - a(-a+d+\sqrt{(a+d)^2-4}))/(2b) = (2b+a^2-ad-a\sqrt{(a+d)^2-4}))/(2b)$; in other words, $2b+a^2+2bc \ge ad + a\sqrt{(a+d)^2-4}$. Since $bc = ad-1$, this is equivalent to $2b + a^2 + 2ad - 2 \ge ad + a\sqrt{(a+d)^2-4}$, or $2b + a^2 + ad \ge 2 + a\sqrt{(a+d)^2-4}$. Since $a(a+d) \ge a\sqrt{(a+d)^2-4}$ and $2b \ge 2$, the inequality is true.

If we can show that $\mathcal{A}(0,1) = (-b,-d)$ satisfies the inequality $y \ge m_u(x-1)$, then all points on $y = m_u(x+b) - d$ will satisfy it and we will have $\mathcal{A}(\partial_s\mathcal{P})^- \subseteq (\partial_s\mathcal{P})^+$. We must show $-d \ge m_u(-b-1)$, that is,

$d \le ((-a+d+\sqrt{(a+d)^2-4})/(2b))(1+b)$, or $(b+1)\sqrt{(a+d)^2-4} \ge bd+ab+a-d = b(a+d)+(a-d)$. Since both sides of the inequality are positive, we can square both sides to get the equivalent inequalities,

$$\begin{aligned}&(b^2+2b+1)((a+d)^2-4)\\ &\qquad\qquad \ge b^2(a+d)^2+2b(a+d)(a-d)+(a-d)^2\end{aligned}$$

$$\begin{aligned}&b^2(a+d)^2+2b(a+d)^2+(a+d)^2+2bd^2\\ &\qquad\qquad \ge b^2(a+d)^2+2ba^2+(a-d)^2+4(b+1)^2\end{aligned}$$

$$\begin{aligned}2ba^2+4abd+2bd^2+2ad+2bd^2 &\ge 2ba^2-2ad+4b^2+8b+4\\ 4abd+4bd^2+4ad &\ge 4b^2+8b+4\\ abd+bd^2+ad &\ge b^2+2b+1\\ b^2c+b+bd^2+bc+1 &\ge b^2+2b+1 \quad (\text{since } ad=bc+1)\\ b^2c+bd^2+bc &\ge b^2+b\\ bc+d^2+c &\ge b+1.\end{aligned}$$

This last inequality is obvious since $c>0$. Thus this partition forms a Markov partition and we are done with all possible cases. $\square$

Acknowledgments. I would like to thank John Franks, Dan Zelinsky and Bob Williams for very helpful conversations.

REFERENCES

1. Roy L. Adler and Benjamin Weiss, *Similarity of automorphisms of the torus*, Mem. Amer. Math. Soc. **98** (1970), 1–43.

2. R. Bowen, *Equilibrium states and the ergodic theory of Anosov diffeomorphisms*, Lecture Notes in Math. **470** (1975).

3. Michael Charles Irwin, *Geometry of continued fractions*, Amer. Math. Monthly **96** (1989), 696–703.

4. C.D. Olds, *Continued fractions*, New Mathematical Library, The Mathematical Association of America, Yale University, 1963.

5. Clark Robinson, *Dynamical systems*, CRC Press, Boca Raton, FL, 1995.

6. Mark Snavely, *Markov partitions for hyperbolic automorphisms of the two-dimensional torus*, Thesis, Northwestern University, 1990.

7. Harold M. Stark, *An introduction to number theory*, Markham Publ. Co., Chicago, 1970.

8. R.F. Williams, *The 'DA' maps of Smale and structural stability*, Proc. Sympos. Pure Math. **14** (1970), 329–334.

Department of Mathematics, Indiana University Northwest, Gary, IN 46408
E-mail address: e_rykken@math.nwu.edu