The torus and the disk

by R. L. Adler

This paper is a survey of a coherent program of mathematics spanning 28 years. It begins with questions concerning classification and structure in ergodic theory and abstract dynamical systems and describes the author's involvement with toral automorphisms, topological entropy, iteration of maps on the interval, symbolic dynamics, and ultimate engineering applications. It serves as a case study of how unplanned-for practical applications can result from the pursuit of mathematics for its own sake.

The first item in the title refers to a mathematical abstraction, while the second is a successful product of the computer industry.

The torus is a compact group and its automorphisms preserve Haar measure. These are classic examples of *dynamical systems with invariant probability measures*, the objects of study in ergodic theory. The basic abstract object of this subject is designated by (X, α, μ), where X is a Lebesgue space (that is, a space endowed essentially with the measure-theoretic structure of the unit interval), α is a measurable mapping of X onto itself, and μ is a probability measure with the property $\mu(E) = \mu(\phi^{-1}E)$ for any measurable subset E of X. A principal question of this subject is one of isomorphism: When does there exist a measure-preserving change of variables? In ergodic theory two dynamical systems (X, α, μ), (Y, β, ν) are said to be *measure-theoretically isomorphic* (*metrically isomorphic* for short) if there exists a mapping γ of X onto Y such that

$\nu(E) = \mu(\gamma^{-1}E)$ and the *conjugacy relation*

$$\gamma\alpha\gamma^{-1} = \beta \text{ (a.e.)}$$

holds. For invertibility of γ all that is required here is that γ^{-1} exists almost everywhere.

Dynamical systems can be considered from other points of view. For example, in the subject of *topological dynamics* we designate a dynamical system[1] by (X, α), where X is a compact metric space and α a continuous map of X onto itself. Here two dynamical systems (X, α), (Y, β) are said to be *topologically isomorphic* (or alternately *homeomorphically conjugate*) if γ in the conjugacy relation is a homeomorphism of X onto Y. This is the strongest sense of equivalence from a purely topological point of view. But later we elaborate on another slightly weaker one, more in the spirit of measure theory, in the sense that we do not insist that the conjugacy relation hold everywhere. Similarly in the theory of *smooth dynamical systems*, the spaces in question are manifolds, the mappings diffeomorphisms, and we would call the notion of isomorphism *diffeomorphic conjugacy*.

The other object in the title is the magnetic storage disk. More generally, we are referring to any data storage or transmission system. In information theory these are portrayed within the framework of a *channel*, as in **Figure 1.** The basic question here concerns the construction of finite state automata which encode and decode data in order to pass them through input-restricted channels. Later we discuss some typical channel constraints.

It is not difficult to suspect a vague connection between the isomorphism question of dynamical systems and the coding problem of information theory: After all, in both subjects one set is being transformed into another. The discovery that these are really different interpretations of the same problem is a consequence of what turned out to be a

[1] For a comprehensive treatment of such systems see [1].

coherent program of research spanning 27 years and serves as a good example of how unplanned-for practical applications can result from the pursuit of mathematics for its own sake.

It began for me as a graduate student in the late fifties. The central problem in ergodic theory was one of metric isomorphism between *Bernoulli shifts.* These are dynamical systems representing stochastic processes like coin-tossing experiments. The problem was: When could the sequences of independently identically distributed results from one probabilistic experiment be coded into another in an invertible and measurable way so that corresponding events under the coding have equal probability? A major breakthrough occurred when Kolmogorov [2] indicated how Shannon's concept of entropy might be utilized as a metric isomorphism invariant, and Sinai [3] supplied proofs necessary to calculate the entropy of Bernoulli shifts. This established in an effective way that shifts of different entropy are not metrically isomorphic. A decade later Ornstein [4] was to prove the converse, that Bernoulli shifts with the same entropy are metrically isomorphic. This led to tremendous progress and a profound understanding of the basic structure of stationary stochastic processes.

As a graduate student I had come across a similar problem concerning the automorphism of the torus. Toral automorphisms are given by members of $GL(n, Z)$, i.e., matrices of integers with determinant ± 1. They preserve Haar measure and, therefore, are metrically isomorphic if they are *algebraically conjugate*—i.e., they are conjugate elements in the group $GL(n, Z)$. Naturally one would be tempted to prove the converse. I managed to prove such a converse if metric conjugacy was replaced by diffeomorphic conjugacy [5]. A few years later Richard Palais showed me how to improve this to homeomorphic conjugacy [6]. In the meantime this was also proved by Arov [7]. But the original metric conjugacy conjecture turned out to be false.

In the early sixties, the notion of topological entropy was suggested to me by Kolmogorov's notion of ε-entropy [8], which measures complexity of function spaces. I realized that a dynamical invariant could be defined for continuous maps by formal analogy with the Kolmogorov-Sinai probabilistic entropy for measure-preserving transformations (see [9]). This is done by replacing measurable partitions with open covers and the number, called the *entropy of the partition*, $\Sigma P(A_i) \log P(A_i)$ with the log of the cardinality of a minimum sub-cover. The *topological entropy* of a continuous map on a compact space can then be defined as the largest possible growth rate of this number as covers are successively refined by action of the map's inverse. There was no more of an idea to it than that.[2] Originally I thought it a mere curiosity. Its main property is that continuous maps which are homeomorphically conjugate have the same topological entropy. But I knew of no maps that could not be distinguished with other invariants more easily; and the converse, as to whether maps with the same topological entropy are homeomorphically conjugate, could be easily shown to be false for any interesting class of continuous maps one would care to mention. Yet there was a striking fact: namely, the Kolmogorov-Sinai entropy (with respect to Haar measure) and the topological entropy are equal for toral automorphisms. The entropy of a toral automorphism is the logarithm of the product of eigenvalues of modulus ≥ 1 of the associated integer matrix.

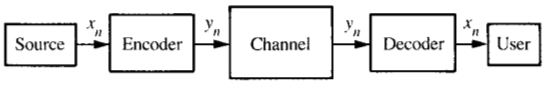

Figure 1

Information channel.

The significance of this emerged a few years later. In the fall of 1966, Leopold Flatto told Benjamin Weiss and me of a new problem which has since gained enormous notoriety: What is the dynamical behavior of the map $x \to ax(1 - x)$ on the unit interval for choices of the parameter a, $1 \leq a \leq 4$? For instance, when is the orbit of the critical point ½ infinite? This has yet to be answered and perhaps is the type of problem that can never be completely settled. We tried our hand on a simpler version—namely, analyze $x \to a - 2a \mid x - \frac{1}{2} \mid$, $1 \leq a \leq 4$. Weiss and I noticed that for certain values of the parameter a, there exists a partition having Markov behavior under the map. This gives rise to a symbolic expansion for the points on the interval which totally describes the dynamical behavior of the map in much the same way as the binary expansion of numbers describes the dynamical behavior of multiplication by two.

We examine this situation in more detail. Let (X, α) be a dynamical system where $\alpha: x \to (2x)$ and X is the unit interval with 0 and 1 identified to make α continuous. Let Σ_2 denote the set of all binary expansions of numbers in the unit interval or equivalently the set of all infinite paths

[2] Some years later Bowen [10] and Dinaburg [11] independently showed the equivalence of the above definition with one derived more directly from Kolmogorov's ε-entropy; namely, the largest growth rate as $\varepsilon \to \infty$ of the number of ε-separated orbits of length n (two orbits of length n are ε-separated if the distance between some pair of corresponding members is $\geq \varepsilon$). Furthermore, this definition brings into clearer focus the fact that topological entropy is a natural generalization of Shannon's noiseless channel capacity.

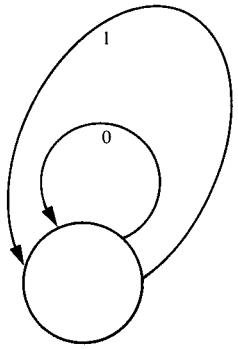

Figure 2

Directed graph of Σ_2.

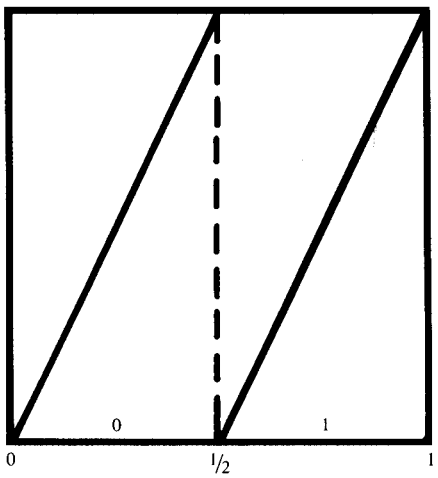

Figure 3

Plot of α.

(sequences of edges) on the directed graph G illustrated in **Figure 2.**

Let σ denote the shift transformation which shifts the symbol sequences in Σ_2 by one to the left and drops off the initial digit. It is easy to define a metric on Σ_2 which makes sequences closer the longer their initial segments agree. This makes Σ_2 a compact metric space, in fact the Cantor discontinuum, and σ a continuous map. We make the elementary observation that the map π of Σ_2 onto X defined by π(binary expansion of x) = x is continuous onto and commutes in the sense that $\sigma\pi = \pi\alpha$. Such maps are called

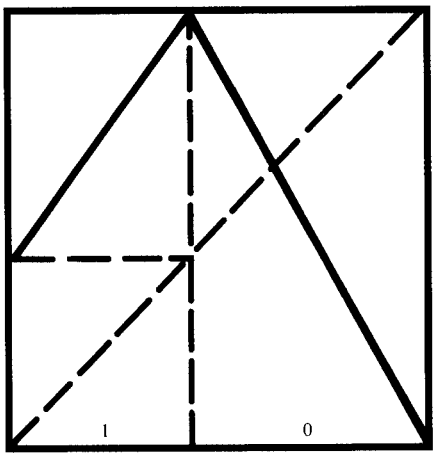

Figure 4

Plot of β.

factor maps (though "quotient" would be a better term), the system (X, α) a *factor* of (Σ_2, σ), and (Σ_2, σ) an *extension* of (X, α). Furthermore, we call π a *finite* factor map since it is nowhere infinite-to-one; in fact it is at most two-to-one, and we call it *essentially one-to-one* since it is one-to-one except for a certain set of rationals which is "negligible" compared to the totality of all numbers. The existence of an extension by a symbolic system which represents a dynamical system in such a simple fashion arises from certain geometrical properties of the map. For example, consider the partition of X into the intervals [0, ½] and [½, 1] (see **Figure 3**). If the first interval is labeled 0 and the second 1, then orbits of the system (X, α) have histories through the partition identical with sequences of Σ_2 which are described by the directed graph of Figure 2.

This happens here because the image of each element of the partition is a union of some others. Partitions that behave like this with respect to a map are called *Markov.* Ambiguities occur when the orbit of a point hits a boundary point of one of the intervals in the partition. Such an occurrence is atypical and is a reflection of the same fact that certain rationals have more than one expansion. In order to get a simple description of the set of allowable expansions, one pays the price by having nonuniqueness of symbolic representation. This is a characteristic feature of decimal expansions in arithmetic and symbolic representations of orbits in dynamical systems.

Consider another example (X, α) where β has a plot as in **Figure 4.** Here the image of the left interval is the right one while the image of the right is the union of both. This

Markov partition gives a symbolic extension (Σ_G, σ) where G is the directed graph in **Figure 5** and Σ_G is the set of infinite paths (here sequences of nodes) on G. This set can be topologized just as Σ_2 and the defined shift σ is continuous. Furthermore, there is an obvious finite essentially one-to-one factor map π which maps histories to points in the interval.

Unfortunately, in analyzing the dynamical behavior of the maps $x \to ax(1-x)$ or $x \to a - 2a|x - \frac{1}{2}|$, Weiss and I found the above considerations useful only for special values of the parameter a. So we abandoned the problem in favor of trying to understand why the two entropies for toral automorphisms yield identical numbers. To our surprise we discovered (or rather rediscovered what K. Berg [12] had found shortly before in research for his Ph.D. thesis) that two-dimensional hyperbolic toral automorphisms have simple Markov partitions. These give rise to symbolic representations, paths on directed graphs, just like those we had been playing with a short time before.

A brief account of our result is as follows. Let $X = R^2/Z^2$ denote the two-dimensional torus and α an automorphism of X. Here points $(x + m, y + n)$ in the plane for $m, n \in Z$ are identified, and α is given by

$$\alpha(x, y) = (ax + cy, bx + dy) = (x, y)A,$$

where A is a matrix with integer entries and determinant ± 1. Haar measure here is merely the projection of area measure in the plane, and area measure is preserved by α because $|\det A| = 1$. The matrix A has two eigenvalues λ and κ with $\lambda\kappa = \pm 1$. This transformation is called hyperbolic if, say, $|\lambda| > 1$, which forces $|\kappa| < 1$. Only the hyperbolic case in dimension 2 is of interest from the dynamical point of view. The geometry of a hyperbolic automorphism is as follows. In the plane there are two distinct directions: one in which distances expand by a factor of $|\lambda|$ under the action of A, and the other in which they contract by $|\kappa|$. Because of this fact one can construct Markov partitions and hence a symbolic extension given by a directed graph. For example, consider the case

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}.$$

In **Figure 6** we draw a Markov partition for the automorphism and in **Figure 7** the associated directed graph G. The idea here is the following. Instead of the unit square another fundamental region for the torus is drawn with sides parallel to the expanding and contracting eigenvectors. This region is then partitioned into two parallelograms I and II. Under the action of the automorphism, these get stretched in one direction and shrunk in the other. Weiss and I did a simple-minded thing. On one sheet of transparent graph paper we drew the fundamental region and on a second the image of it under the automorphism. We placed one sheet on top of the other and slid them around to see how the two set partitions got refined. New lines appeared in the

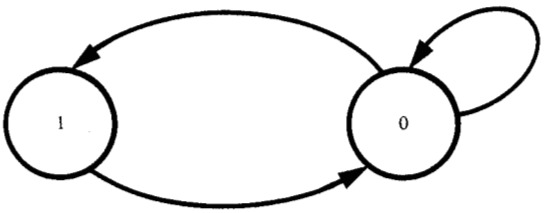

Figure 5

Directed graph of G.

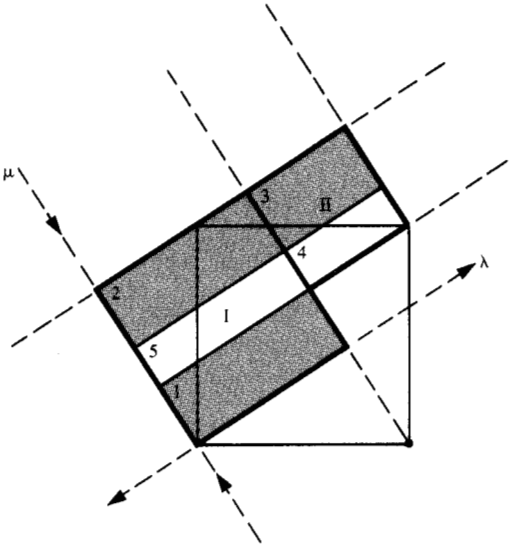

Figure 6

Markov partition.

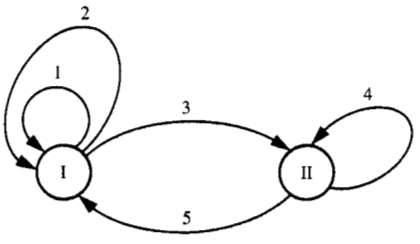

Figure 7

Directed graph G.

expanding direction, but to our amazement no new ones in the shrinking direction. This has a simple geometric explanation and profound consequences. As shown in Figure 6, the two basic parallelograms I and II get refined into five smaller ones. The image of rectangle 1 is a collection of rectangles stretching across 1, 2, and 3; similarly for the images of 2 and 5; the image of 3 stretches across 4 and 5; similarly for 4. These facts are summarized by the transitions allowed in the graph depicted in Figure 7.

We now make a slight change in our notion of Σ_G. We assume from now on that it consists of bi-infinite paths in the graph G, which makes the shift σ an invertible map on Σ_G. One of the consequences of the fact that there are no new lines in the shrinking direction under repeated applications of the automorphism is the existence of a finite essentially one-to-one factor map π of Σ_G onto X. This map associates a bi-infinite path in G to a unique point of X having that path as a history through the partition under the action of α. We shall not give the proof of this, but suffice it to say that it follows from some elementary plane geometry.

The areas of the parallelograms in the Markov partition are numbers with special meaning. We found that a symbolic system (Σ_G, σ) which is an extension of a toral automorphism satisfies a variational principle: namely, the topological entropy is the same as the maximum probabilistic entropy[3], which in turn is the same as the entropy of the toral automorphism. Also encouraged by Meshalkin's [14] success in coding between certain Bernoulli shifts, we found that we could construct metric conjugacies (i.e., measure-preserving changes of variables) by coding between these symbolic systems representing toral automorphisms whenever they had the same entropy. Here a simplification occurred. Inherent in the power of our method we merely had to construct a measurable change of variables: The measure-preserving property was forced to accompany it by virtue of the fact that topological entropy and maximum probabilistic entropy coincide. Answering the question on which I had been stuck as a graduate student, we were able to prove the following.

Theorem *Two 2-dimensional hyperbolic toral automorphisms are metrically isomorphic if and only if they have the same entropy, i.e., the same corresponding* $|\lambda|$.

This was the first natural class of dynamical systems to be classified by entropy. In the early seventies Ornstein [15] made a vast generalization.

Our work [16] combined two new important ideas: finding Markov partitions for smooth dynamical systems and coding between symbolic systems associated with the partitions. Each idea has stimulated mathematical activity.

With respect to the first one, R. Bowen [17, 18] and Ya. Sinai [19] established the existence of Markov partitions for a more general class of smooth dynamical systems. In particular these results give Markov partitions for hyperbolic toral automorphisms of any dimension. Furthermore, Sinai [20] made an application of Markov partitions to some basic questions in statistical mechanics. For an entrance into the literature on the use of Markov partitions in smooth systems, one can consult [21].

Weiss and I concentrated our further work on the second idea. We observed that the codes which we were constructing between symbolic systems were stronger than metric conjugacies yet weaker than homeomorphic ones. Also we could code between examples of symbolic systems with the same topological entropy. The codes we were constructing were almost but not quite invertible. They failed to be one-to-one on a small set of exceptional symbol sequences. This is also the case for metric conjugacies in general, but our set of exceptional points was universally negligible with respect to any regular invariant probability measure rather than a fixed one. Later (see [22, 23]) when their nature was better understood we called these codes "almost-homeomorphic conjugacies." Fashioned from the relationship of a binary expansion and the number it represents, almost-homeomorphic conjugacy is a relation, between topological systems, which has the appearance of being only slightly weaker than homeomorphic conjugacy. Two topological dynamical systems (X, α) and (Y, β) are said to be *almost homeomorphically conjugate* if they are factors of a common extension, say (Z, ρ), and the factor maps are finite and essentially one-to-one. Here *essentially one-to-one* means that the factor maps are one-to-one on the doubly transitive points—that is, the points whose future orbits and past orbits are both dense. In systems which satisfy a standard irreducibility condition, the nondoubly transitive points comprise a negligible set in the sense of measure and category just like those numbers which have more than one binary expansion. Two basic facts can be proved: *Almost-homeomorphic conjugacy is an equivalence relation,* and *topological entropy is an invariant.*

The symbolic systems with which we were dealing (namely, bi-infinite paths on directed graphs) we called *topological Markov shifts* because they could be specified by nonnegative *transition* matrices. The name was chosen because these matrices resemble stochastic ones except that the positive transition probabilities have been replaced by nonnegative integers. The relevant transition matrix is one with entries 0, 1 and is specified as follows: There is a 1 in the ith row and jth column if and only if edge i leads next to edge j. We could just as easily label nodes, in which case: There is an n in the ith row and jth column if and only if there are n paths from node i to node j. Sometimes it is

[3] Another rediscovery. The topological entropy of (Σ_G, σ) is the same thing as Shannon's noiseless channel capacity, and the fact that it equals the maximum probabilistic entropy was known to him for Markov measures. Parry [13] rediscovered this fact and generalized it to arbitrary measures.

more convenient to work with nodes, sometimes edges: e.g., if n assumes values other than 0 or 1, then edge labeling avoids an ambiguity which arises using nodes. A zero-one transition matrix specifies a space of admissible sequences of symbols (the labels of nodes or edges), and along with the shift transformation we get a *symbolic dynamical system.* These systems go under other names: *topological Markov chains* [13] and *shifts of finite type* [24]. We call a directed graph as well as its transition matrix and the dynamical system it specifies *irreducible* if any pair of nodes is connected by a directed path and *aperiodic* if any pair of nodes is connected by a directed path of the same length. The topological entropy of a topological Markov shift is the log of the largest eigenvalue of its transition matrix. This eigenvalue is called the *Perron* value.

We considered first the case where the Perron value was an integer N. A row sum N matrix has Perron value N, but not conversely. We conceived of a proof consisting of two parts: Part 1, to prove that for a system whose matrix has Perron value N there exists a code to a new system where the associated matrix has row sum N; Part 2, to prove that there exists a code between a row sum N system and the system given by an $N \times N$ matrix of all ones (such a system is called the *full N-shift* and its space of sequences is denoted by Σ_N). We could prove Part 1 by a method which has come to be known as *state splitting*, but the proof of Part 2 was elusive. This problem became known as the "road problem."

Here is the simplest version of the road problem. A group of cities is connected by an aperiodic network of one-way roads, each city having two exit roads (a city having a road leading to itself is not excluded). The highway department has two colors, say red and blue, with which to paint the roads. Each city has one red exit and one blue. Is it possible to color the roads in such a way that there is a sequence of colors that leads everyone simultaneously to the same city, say city 1, no matter where he starts?

It is still unsolved; but the most general result to date, I believe, has been done by O'Brien [25]. In 1975 at an ergodic theory symposium at the University of Warwick, L. W. Goodwyn found a way to bypass the problem by observing that it was sufficient to solve the road problem for a higher-order edge graph, which is easier to do than for the original one. This is described in [26].

At that same symposium Furstenburg proved that two irreducible nonnegative integer matrices A, B (not necessarily of the same dimension) have the same Perron value if and only if there exists a positive integer matrix F such that $AF = FB$. On the basis of this result, Brian Marcus and I were able to prove a topological analogue to Ornstein's isomorphism theorem, to wit:

Theorem[4] *Two aperiodic topological Markov shifts are almost-homeomorphically conjugate if and only if they have the same topological entropy* [22].

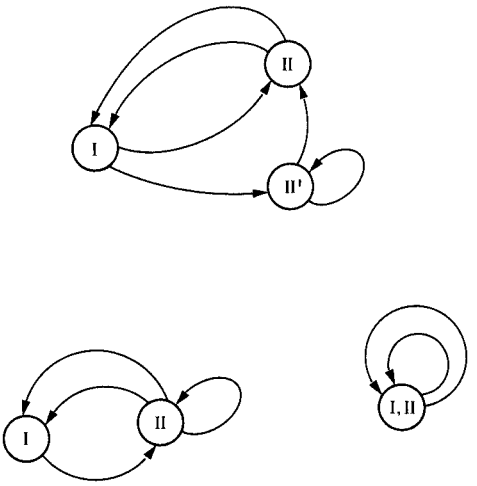

Figure 8

A common extension.

If the transition matrix is irreducible but not aperiodic, another invariant, the period, must be included with the entropy. Our method is based on a technique which we call "filling in tableaux," which constructs a new graph from two others with the same Perron value. In the new graph the outgoing edges at the various nodes are like those of one of the original graphs, while the incoming edges are like those of the other. This new graph gives a dynamical system which is a common extension of the original ones and the almost-homeomorphism is constructed from factor maps. The factor maps are defined by merging appropriate nodes. (See **Figure 8.**)

A corollary of this and Bowen's result [17, 18] is the fact that the theorem of Adler and Weiss can be generalized to hyperbolic toral automorphisms of all dimensions.

Theorem Two hyperbolic toral automorphism are almost-homeomorphically conjugate (hence metrically conjugate) if and only if they have the same entropy.

An isomorphism theory (at least the type we are discussing) has three elements: an equivalence relation, an invariant, and a special class of systems for which the invariant is a complete one. For ergodic theory they are metric conjugacy, probabilistic entropy, and Bernoulli shifts;

[4] Likewise Parry [27] used Furstenburg's result to obtain a version of this theorem which stops short of getting the one-to-one condition in almost-homeomorphic conjugacy. Thus he obtains a weaker relation, which he calls *finite equivalence*, between two topological Markov shifts with the same entropy.

while for topological dynamics they are almost-homeomorphic conjugacy, topological entropy, and topological Markov shifts. What about an isomorphism theory based on homeomorphic conjugacy? The outstanding problem concerning shifts of finite type is to give an algorithm (or prove its nonexistence) for determining homeomorphic conjugacy. For material concerning this problem, see [28–32]. If two dynamical systems are homeomorphically conjugate, then they have the same topological entropy [9]. However, it is not hard to demonstrate that the converse is not true even for shifts of finite type. Williams [32] gave an algebraic characterization of topological isomorphism for shifts of finite type. Some algebraic invariants, such as the Jordan form away from 0-eigenvalues, result which are stronger than topological entropy. However, all of the currently known computable invariants are inadequate to completely classify these systems with respect to topological isomorphism. The trouble seems to be that homeomorphic conjugacy is too strong an equivalence relation. It is the weaker one, almost-homeomorphic conjugacy, with respect to which an isomorphism theory with a simple description can be established.

The class of dynamical systems to which our isomorphism theorem applies can be enlarged to include sofic systems (the term "sofic," introduced by Weiss [30], is derived from the Hebrew word for finite and is supposed to suggest the finitary character of these systems). It is the set of output sequences from a finite state automaton. Shannon [33] called them "transducers." A *sofic* system (S, σ) is defined by choosing S to be the space of sequences of symbols gotten from bi-infinite paths on directed graphs just like topological Markov shifts except that perhaps the nodes (edges) are not distinctly labeled. See [34–36]. If the nodes (edges) have distinct labels, the sofic system is a subshift of finite type, but if not, then it may or may not be homeomorphically conjugate to a topological Markov shift. Generally it is not. In any case topological entropy is a complete invariant with respect to almost-homeomorphic conjugacy for aperiodic members of this larger class of symbolic systems.

Marcus [37] improved the state-splitting method introduced by Weiss and me to show that a topological Markov shift with Perron value N is actually homeomorphically conjugate to the one whose transition matrix has row sum N. The row sum N system is a common extension between the original topological Markov shift and the full N-shift. From this follows a stronger statement than the isomorphism theorem for the special case of rational integer Perron values: namely, a topological Markov shift with Perron value N is almost-homeomorphically conjugate to a full N-shift via a factor map. His method has practical implications which we mention later.

The practical applications of the isomorphism theory in topological dynamics were first recognized by Martin Hassner [38]. While doing research for his Ph.D. in electrical engineering, he was struck by the fact that our notion of almost-topological conjugacy is a coding given by a finite algorithm, so that an engineering application ought to be possible. (The metric conjugacies for Bernoulli shifts given by Ornstein are not anything like finite algorithms.) He also was aware that information channels such as ones describing magnetic data storage devices were modeled precisely by topological Markov shifts and sofic systems. As a result of his insight, the attention of mathematicians in ergodic theory and dynamical systems was directed to an area of engineering of which they had previously been oblivious.

Digital information usually takes the form of long binary sequences which we can assume to be arbitrary. When data are to be transmitted or stored, a system doing so may force constraints on the binary sequences. For example, in storing binary data on magnetic surfaces (tapes, disks, drums, etc.) the symbol 1 is ascribed to a transition in the magnetic state of the surface and 0 to a nontransition. During the read-back process, a transition between magnetic states will cause a voltage pulse while a nontransition will result in an absence of signal in the read head. In such recording systems, the separation of transitions is measured in terms of some basic bit duration unit and constraints are introduced for the following reasons. 1) If the transitions are too close, interference between adjacent voltage pulses in the circuits attenuates signals and shifts peaks—trouble for a peak-detection scheme. This places a lower limit on the minimum separation between successive transitions. 2) If the adjacent transitions are too far apart, the absence of a signal may cause a data-based clocking scheme, used to correct for drift, to lose synchronization, giving a false measurement of the number of bit duration units. This places an upper limit on the maximum separation between transitions.

This translates into the so-called (d, k) constraints for binary sequences, where d is the minimum number of zeros between ones and k the maximum. Consequently, a device is needed to code between arbitrary binary sequences and constrained ones. In order for such a device to be practical, it can only process small amounts of data at a time. From its point of view the data that it is processing, though finite, may just as well be infinite both with respect to the past as well as the future. Therefore, symbolic dynamical systems provide a perfect model for such a situation. Arbitrary data are modeled by the full 2-shift (or the full 2^k-shift if one wants to consider blocks of data) and (d, k)-constrained data by a topological Markov shift. Furthermore, certain constraints are modeled by sofic systems. For example, in magnetic recording problems may develop due to a dc component in the electrical signal, especially if the head is coupled by an induction coil to the rest of the system. Besides heating the coil, errors could be introduced in the read-back process if the spectrum of the binary sequence contained any power at the frequency zero. Therefore, we

may be forced to place spectral constraints on our sequences, and such constraints usually lead to sofic systems [39].

The engineer is now faced with two problems. 1) He must code arbitrary data into constrained, and thereby lose some rate. What coding rates are possible? 2) Given a possible coding rate, how does one construct practical finite state automata to do the coding and encoding?

The answer to the first question is provided by the notion of topological entropy, which is the same as Shannon's noiseless channel capacity.[5] This determines the possible coding rates. The tableau and state-splitting methods construct common extensions and factor maps. This gives a solution to the second problem in the case that topological entropy of input data matches capacity of the channel. To explain this we must discuss the concept of the factor map[6] in more detail.

Let L_A and L_B denote an ordered set of symbols for two topological Markov shifts (Σ_A, σ) and (Σ_B, σ), respectively. Also let $\pi:\Sigma_A \to \Sigma_B$ be a factor map. It follows from continuity and the shift-commuting property that π is a *k-block* map for some integer k. This means that there is a fixed function $\pi:L_A \times \cdots \times L_A \to L_B$ of k variables (using π again by a slight abuse of notation) such that $y_i = \pi(x_{i-j}, x_{i-j+1}, \cdots, x_{i-j+k-1})$ for some fixed $j \in Z$. (By a suitable change of the symbol set L_A and the transition matrix A we can always arrange $j = 0$, $k = 1$ so as to obtain a 1-block map.) If the mapping π is also invertible, then π^{-1} is also a k-block map, for perhaps a different k. Finite factor maps[7] are not invertible. They do, however, possess a weak type of invertibility. Assuming, without loss of generality, that $\pi x = y$ is a 1-block factor map, then it is finite if and only if $(x_{i-p}, \cdots, x_{i+q})$ is uniquely determined from $(y_{i-p}, \cdots, y_{i+q})$, x_{i-p} and x_{i+q} for all i, p, q. Another way of saying this is: If $\pi(u) = \pi(v)$ and the distance between $\sigma^i u$ and $\sigma^i v$ goes to 0 as $i \to \mp\infty$, then $u = v$ ("there are no diamonds in pre-images"). A stronger kind of this weak invertibility is the following. A factor map π is said to be *right-closing*[8] if there are fixed integers $p, q \geq 0$, $r \geq 1$ such that if $\pi x = y$, then x_i is uniquely determined from $(y_{i-p}, \cdots, y_{i+q})$ and the coordinates $(x_{i-r}, \cdots, x_{i-1})$ which are to the left of x_i. Another way of saying this is: If $\pi(u) = \pi(v)$ and the distance between $\sigma^i u$ and $\sigma^i v$ goes to 0 as $i \to -\infty$, then $u = v$ ("there are no right forks in pre-images"). We call a right-closing map *right-resolving* if it is a 1-block map such that $r = 1$, $p = q = 0$. *Left-closing* and *left-resolving* are similarly defined.

The concept of a right-resolving factor map can be given the following engineering interpretation. $\pi:L_A \to L_B$ determines how to construct a finite state automaton which encodes sequences Σ_B to Σ_A. The present output symbol depends on the present input and the present internal state. The present internal state depends on the past input and the past internal state. The encoding proceeds from left to right on the sequences, but this specification of the direction of time is merely a convention. The general isomorphism theorem for topological Markov shifts states that given (Σ_A, σ) and (Σ_B, σ) of equal entropy, there is a common extension (Σ_C, σ) and two essentially one-to-one 1-block factor maps $\pi_1:\Sigma_C \to \Sigma_A$ and $\pi_2:\Sigma_C \to \Sigma_B$. One of these maps is right-resolving and one is left-resolving. This means that the finite state automaton which encodes sequences from Σ_A via Σ_C to Σ_B proceeds from right to left, whereas the one that decodes goes from left to right.

In the case of Perron value N, an integer, we have a special situation. Encoding from (Σ_N, σ) to (Σ_A, σ) is derived, as in general, from a right-resolving factor map $\pi_1:\Sigma_C \to \Sigma_N$, where Σ_C is a common extension of Σ_A and Σ_N. The graph given by C defines the automaton and the factor map $\pi_2:\Sigma_C \to \Sigma_A$ specifies its output. However, from Marcus's theorem the factor map $\pi_2:\Sigma_C \to \Sigma_A$ from the common extension to Σ_A is invertible, which means that the decoding automaton is just the image of a continuous map—namely, $\pi_1\pi_2^{-1}$—and does not depend on an internal state. Engineers call this type of decoder a *sliding block decoder.* Thus, while errors in the input of the encoder might cause infinite output errors, error propagation is limited for a sliding block decoder. This is just what is required for encoding and decoding data for the magnetic recording information channel. The channel coder is not responsible for user errors, only channel errors. These he wants to limit in order not to overwhelm an *error-correcting code*, another level of coding which has not been part of our discussion.

To sum the last two paragraphs, the engineering significance of almost-homeomorphic conjugacy (in fact, even the weaker relation, finite equivalence) is that automata can readily be constructed to encode and decode arbitrarily long sequences from one subshift of finite type to another of equal entropy. When the first is the full N-shift, then the decoder can be made sliding block, something not true in general. An added quality of almost-homeomorphic conjugacy over finite equivalence is the existence of a finite input which resets the encoding automaton independent of

[5] It is instructive at this point to compare Shannon's noiseless coding theorem [33] with the results presented here. His theorem states that coding is possible between source and channel data if the source entropy is strictly less than the channel capacity and impossible if the inequality is reversed. He does not treat the case of equality. Furthermore, his theorem says nothing about how coding can be done with automata. On the other hand, his source entropy is a probabilistic one, something more general than topological entropy.

[6] For initial literature on factor maps in symbolic dynamics see the paper of Hedlund [40].

[7] Kitchens [41] found an important algebraic property of factor maps of topological Markov shifts: *If $\pi:\Sigma_A \to \Sigma_B$ is a finite factor map, then the Jordan form of A, apart from the 0-eigenvalues, contains that of B.*

[8] This concept was known to B. McMillan [42] and was called *unifilar* by him.

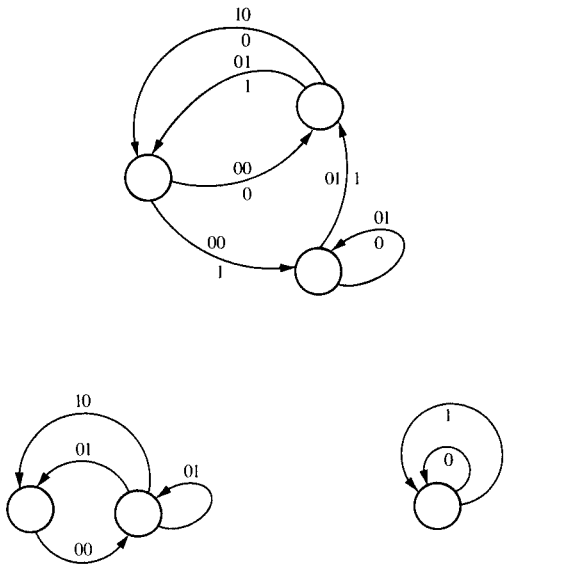

Figure 9

Automata.

the current internal state (likewise for the decoder). There may be situations where it is important to know the current internal state: for instance, an initial state is an ingredient in specifying an actual device.

To illustrate what we have been saying, we return to the previous example in **Figure 9** with edges labeled by data symbols. The graph of the full 2-shift Σ_2 is on the right with its edges labeled by 0 and 1. It represents sequences of arbitrary data. The graph of Σ_A is on the left with its edges labeled by some pairs of zeros and ones. It represents constrained data. The sequences of pairs ostensibly comprise a sofic system. Actually it is topologically isomorphic to the Markov shift Σ_A because there is a one-to-one correspondence between sequences of pairs and sequences of edges. Here sequences of pairs obey constraints which imply run-length ones. One can read from the graph of the common refinement Σ_C on top how to encode arbitrary data to constrained data at a rate 1:2 and also decode. The internal states are six edges of graph C which should be viewed as having distinct labels. Encoding proceeds from left to right (reverse for decoding). On the other hand, this is a case of integer Perron value (two). For this example the state-splitting method and the tableau method construct the same common extension. A consequence of the state-splitting method is that we can decode in the same direction as encoding, namely, from left to right. In this method, to determine a present unconstrained symbol one just looks ahead in the sequence of constrained pairs.

Marcus's result can be used to solve the second engineering problem of achieving error-limited decoding with a sliding block decoder for rate $p:q$ in the case where the Perron value = $2^{p/q}$, the log of which is the entropy. This applies to the case where the input system taken as p-blocks of user data has the same topological entropy as that of q-blocks of channel-constrained data. Constraints rarely have such entropies, so his method had to be extended to include Perron values $\neq 2^{p/q}$; and this was done by Adler, Coppersmith, and Hassner [43]. One can always get the source entropy less than or equal to the channel capacity by choosing p and q properly.

The theorem of Marcus [37] for factoring topological Markov shifts onto the full N-shift was generalized by him [44] to include a special class of sofic systems he called *almost-finite type.* These are described by finite state automata with the property that output sequences lying in some open set determine sequences of internal states—i.e., nodes of the defining graph. This mathematical notion was inspired by engineering applications, which impose constraints on signals representing bit strings, like having no dc component. Spectral constraints of this sort lead to this type of sofic system. Actually Marcus's result [44] for sofic systems was not as strong as the one for topological Markov shifts because it just gives a factoring of (S, σ^k) onto (Σ_N, σ^k) for some k.[9] The following remark will perhaps shed some light on the practical significance of this. Patel [45] invented a code for the so-called *zero-modulation channel* used in the IBM 3850, a mass storage magnetic tape system. It satisfies the (1, 3) run-length constraint along with a spectral constraint to eliminate the dc component in the electrical signals. This makes it a sofic system. Patel discovered a code with a simple sliding block decoder having small error propagation. The encoder, however, is based on a finite factor map which is not right-resolving. In order to design an encoding automaton for this particular system, a small amount of rate had to be sacrificed. As a by-product of Marcus's and Karabed's research on sofic systems, it can now be done without this sacrifice. Admittedly, there is a substantial increase in complexity and error propagation in the decoder.

It is appropriate to mention here the interesting work of Franaszek [46] and of Lempel and Cohn [47]. They developed methods similar to the aforementioned in the sense that all are based on the Perron-Frobenius spectral theory of nonnegative matrices. Mathematical clarity which

[9] A hot new result at IBM Research San Jose by him and a collaborator, R. Karabed, is the fact that $k = 1$, just as in the Markovian case. In addition, Marcus was then able to extend this theorem to arbitrary sofic systems by inventing a new notion he called a *non-catastrophic decoder*, which was also inspired by engineering requirements. It is a generalization of the notion of the sliding block decoder and beautifully fits the mathematics of the situation.

is missing from their work can now be supplied in terms of notions which have become standard in symbolic dynamics.

Finally, patents [48–50] and products (the IBM 9332 Hard Disk File) have accrued as dividends of this kind of work. Moreover, in the past it might take an engineer several months to design, by ingenious *ad hoc* methods, the code tables and the logic for an automaton to code data to fit constraints such as (d, k) ones. Now some of the most complicated cases can be routinely handled; but the main point is that these can often be done in less than a day.

References

1. M. Denker, C. Grillenberger, and K. Sigmund, "Ergodic Theory on Compact Spaces," *Springer Lecture Notes in Mathematics* **527,** 1976.
2. A. N. Kolmogorov, "A New Metric Invariant of Transitive Automorphisms of Lebesgue Spaces," *Dokl. Akad. Nauk SSSR* **119,** No. 5, 861–864 (1958).
3. Ya. G. Sinai, "On the Concept of Entropy of a Dynamical System," *Dokl. Akad. Nauk SSSR* **124,** 768–771 (1959).
4. D. S. Ornstein, "Bernoulli Shifts with the Same Entropy Are Isomorphic," *Adv. Math.* **5,** 337–352 (1970).
5. R. L. Adler, "Diffeomorphic Conjugacy of Automorphisms on the Torus," *Research Report RC-1117,* IBM Thomas J. Watson Research Center, Yorktown Heights, NY, February 1964.
6. R. L. Adler and R. Palais, "Homeomorphic Conjugacy of Automorphisms on the Torus," *Proc. Amer. Math. Soc.* **16,** No. 6, 1222–1225 (1965).
7. D. Z. Arov, "Topological Similitude of Automorphisms and Translations of Compact Commutative Groups," *Uspehi Mat. Nauk* **18,** No. 5, 133–138 (1963).
8. A. N. Kolmogorov and V. M. Tihomirov, "ε-Entropy and ε-Capacity of Sets in Functional Spaces," *Uspehi Mat. Nauk* **14,** No. 2, 3–86 (1959); *Amer. Math. Soc. Transl.* **17,** No. 2, 277–364 (1961).
9. R. L. Adler, A. G. Konheim, and M. H. McAndrew, "Topological Entropy," *Trans. Amer. Math. Soc.* **114,** 309–319 (1965).
10. R. Bowen, "Entropy for Group Endomorphisms and Homogeneous Spaces," *Trans. Amer. Math. Soc.* **153,** 401–414 (1971).
11. E. I. Dinaburg, "The Relation Between Topological Entropy and Metric Entropy," *Dokl. Akad. Nauk SSSR* **190,** No. 1, 13–16 (1970).
12. K. Berg, "On the Conjugacy Problem for K-Systems," Ph.D. thesis, University of Minnesota, 1967.
13. W. Parry, "Intrinsic Markov Chains," *Trans. Amer. Math. Soc.* **112,** 55–66 (1964).
14. L. D. Meshalkin, "A Case of Isomorphism of Bernoulli Schemes," *Dokl. Akad. Nauk SSSR* **128,** 41–44 (1959).
15. D. S. Ornstein, *Ergodic Theory Randomness and Dynamical Systems,* Yale Math. Monographs, Vol. 5, Yale University Press, New Haven and London, 1974.
16. R. L. Adler and B. Weiss, "Similarity of Automorphisms of the Torus," *Mem. Amer. Math. Soc.* **98** (1970).
17. R. Bowen, "Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms," *Springer Lecture Notes in Mathematics,* 1975.
18. R. Bowen, "Markov Partitions for Axiom A Diffeomorphisms," *Amer. J. Math.* **91,** 725–747 (1970).
19. Ya. G. Sinai, "Markov Partitions and C-Diffeomorphisms," *Funct. Anal. & Appl.* **2,** 64–89 (1968).
20. Ya. G. Sinai, "Ergodic and Kinetic Properties of the Lorentz Gas," *Ann. N.Y. Acad. Sci.* **357,** 143–149 (1980).
21. A. Katok, Ya. G. Sinai, and A. M. Stepin, "Theory of Dynamical Systems and General Transformation Groups with an Invariant Measure," *J. Soviet Math.,* 974–1064 (1977).
22. R. L. Adler and B. Marcus, "Topological Entropy and Equivalence of Dynamical Systems," *Mem. Amer. Math. Soc.* **219.**
23. R. L. Gray, "Generalizing Period and Topological Entropy to Transitive Non-Wandering Systems," Master's thesis, University of North Carolina, Chapel Hill, 1978.
24. S. Smale, "Differentiable Dynamical Systems," *Bull. Amer. Math. Soc.* **73,** 747–813 (1967).
25. G. L. O'Brien, "The Road Colouring Problem," *Israel J. Math.* **39,** 145–154 (1981).
26. R. L. Adler, L. W. Goodwyn, and B. Weiss, "Equivalence of Topological Markov Shifts," *Israel J. Math.* **27,** 49–63 (1977).
27. W. Parry, "A Finitary Classification of Topological Markov Chains and Sofic Systems," *Bull. Lond. Math. Soc.* **9,** 86–92 (1977).
28. K. Baker, "Strong Shift Equivalence of 2×2 Matrices of Nonnegative Integers Associated with Topological Markov Chains," *Ergod. Theory & Dynam. Syst.* **3,** 501–508 (1983).
29. K. H. Kim and F. W. Roush, "Some Results on Decidability of Shift Equivalence," *J. Combin., Info. & Syst. Sci.* **4,** 123–146 (1979).
30. B. Weiss, "Subshifts of Finite Type and Sofic Systems," *Monatsh. Math.* **77,** 462–474 (1973).
31. W. Parry and R. F. Williams, "Block Coding and a Zeta Function for Finite Markov Chains," *Proc. Lond. Math. Soc.* **35,** 483–495 (1977).
32. R. F. Williams, "Classification of Shifts of Finite Type," *Ann. Math.* **98,** 120–153 (1973).
33. C. Shannon, "A Mathematical Theory of Communication," *Bell Syst. Tech. J.* **27,** 379–423, 623–656 (1948).
34. E. M. Coven and M. E. Paul, "Endomorphisms of Irreducible Subshifts of Finite Type," *Math. Syst. Theory* **8,** 167–175 (1974).
35. E. M. Coven and M. E. Paul, "Sofic Systems," *Israel J. Math.* **20,** 165–177 (1975).
36. E. M. Coven and M. E. Paul, "Finite Procedures for Sofic Systems," *Monatsh. Math.* **83,** 265–278 (1977).
37. B. Marcus, "Factors and Extensions of Full Shifts," *Monatsh. Math.* **88,** 239–247 (1979).
38. M. Hassner, "A Non-Probabilistic Source and Channel Coding Theory," Ph.D. dissertation, UCLA, 1980.
39. K. Petersen, "Chains, Entropy, and Coding," *Ergod. Theory & Dynam. Syst.* (to appear).
40. G. A. Hedlund, "Endomorphisms and Automorphisms of the Shift Dynamical System," *Math. Syst. Theory* **3,** 320–375 (1969).
41. B. Kitchens, "An Invariant for Continuous Factors of Markov Shifts," *Proc. Amer. Math. Soc.* **83,** 825–828 (1981).
42. B. McMillan, "The Basic Theorems of Information Theory," *Ann. Math. Stat.* **24,** 196–219 (1953).
43. R. L. Adler, D. Coppersmith, and M. Hassner, "Algorithms for Sliding Block Codes," *IEEE Trans. Info. Theory* **IT-29,** 5–22 (1983).
44. B. Marcus, "Sofic Systems and Encoding Data," *IEEE Trans. Info. Theory* **IT-31,** 366–377 (1985).
45. A. Patel, "Zero Modulation Encoding in Magnetic Recording," *IBM J. Res. Develop.* **19,** No. 4, 366–378 (1975).
46. P. Franaszek, "Construction of Bounded Delay Codes for Discrete Channels," *IBM J. Res. Develop.* **26,** 506–514 (1982).
47. A. Lempel and M. Cohn, "Look Ahead Coding for Input Restricted Channels," *IEEE Trans. Info. Theory* **IT-28,** 933–937 (1982).
48. R. L. Adler, M. Hassner, and J. Moussouris, "Method and Apparatus for Generating a Noiseless Sliding Block Code for a (1,7) Channel with Rate 2/3," U.S. Patent 4,413,251, November 1, 1983.
49. D. Coppersmith and B. Kitchens, "Run-Length Limited Code Without dc Level," patent pending.
50. G. Langdon and P. Siegel, "Direction Constrained Ternary Codes Using Peak and Polarity Detection," U.S. Patent 4,566,044, January 21, 1986.

Received August 2, 1986; accepted for publication November 7, 1986

Roy L. Adler *IBM Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598.* Dr. Adler received the following degrees in mathematics: a B.S. from Yale University, New Haven, Connecticut, in 1952; an A.M. from Columbia University, New York, New York, in 1954; and a Ph.D. from Yale University, New Haven, Connecticut, in 1961. In 1960, he joined the IBM Corporation as a Research staff member at the Thomas J. Watson Research Center; he is currently manager of the general mathematical studies group. From graduate school to the present his research has been centered in the subjects of ergodic theory and dynamical systems. Dr. Adler was co-recipient of the 1985 Paper Award, Institute of Electrical and Electronics Engineers, Information Theory Group (Reference [43] in this paper). He has been appointed to serve on the Board of Trustees of the Mathematical Sciences Research Institute and is a Fellow of the New York Academy of Science.