Version 3.1 22.11.11

Lectures on symbolic dynamics

Kari Eloranta

kve@math.hut.fi

Introduction

These lectures are an attempt to present an overview of certain facets of the theory of dynamical systems. This is already a vast subject and currently under intense research effort. Our objects of interest are in the general sub areas of ergodic theory and symbolic dynamics. Broadly speaking the former is the probabilistic form of dynamics while the latter deals with the asymptotic properties of sequences from a finite alphabet satisfying suitable local restrictions. The "suitability" is determined often by an other dynamical system that gives rise to the symbolic system. Some of the sequences arise naturally from smooth dynamical systems like differential equations defining a flow on a manifold whereas others come from lattice models in statistical mechanics. Apart from resolving questions in the initial (say smooth) framework the sequence space formulation has turned out to be of interest by itself.

The material here is mostly introductory and emphasis is on treatment of a few key examples. To get a coherent picture we need to present along the way some basic theory from the core the dynamical systems. The part that deals with the thermodynamic fromalism is naturally motivated and quite close to ideas in theoretical physics. We also show the relation of symbolic dynamics to automata theory as well as some connections to information theory.

Symbolic dynamical systems that will be considered later on are $\mathbf{Z}^d$-actions and within them the cellular automata. These form an extremely rich and deep class of systems. They provide the full spectrum of examples from simple transient dynamics to chaotic and complex self-organizing dynamics therby providing a testing ground for advanced theoretical notions. Their "physics-like" behavior has stimulated a large amount research part of which will be reviewed.

Caveat: This is still just a set of lecture notes. Typos and mild illogicalities are part of the deal. Effort is made to avoid major blunders but if any such is spotted it should in Chung's ([Ch]) spirit be part of the challenge to the reader...

References. At the end of the sections we collect some key references to the material presented. The actual literature addresses are found at the end in the Bibliography section.

0.1. A few problems

To get the flavor of the topic let us consider some of the basic problems.

Example 0.1.: Suppose that we form a set X of bi-infinite sequences on the integers by assigning one of the symbols 0 and 1 at each integer with the extra requirement that two ones cannot be next to each other. X is clearly non-empty but how big is it? It is easy to see that it is uncountable but is there a natural way of measuring it's size? If one considers the number of such *finite* sequences on $\{-N, \ldots, N\}$ is the growth of their number polynomial or exponential in N? Note that without the exclusion of consecutive ones the growth is exponential so it cannot be any faster in the restricted case. Turns out that there is a general procedure to resolve this question. Another line of questions could be this: what is a typical element like? Are there elements which imitate all of the others in the sense of having a dense orbit under the coordinate shift?

Example 0.2.: On the square integral lattice in the plane, $\mathbf{Z}^2$, we could do something analogous to the first example: allow all arrangements of 0's and 1's on this lattice except those where two ones are next to each other either horizontally of vertically. Again the set of these configurations is uncountable and one can ask the same question about the size as in Example 1. Nobody knows the answer! (except numerically) Interestingly the difficulty is not just in the two-dimensionality: the same rule on the triangular lattice can be analyzed.

The growth rate that we were after in these two examples is closely related to the entropy of the system. It is a key global characteristic of a dynamical system since it captures the degree of (pseudo) randomness in the system.

Example 0.3.: Consider again the set of all sequences of of zeros and ones one the integers. Define a block map by requiring that it maps 001, 011, 100, and 110 $\mapsto$ 1 and the other four nearest neighbor binary triples to 0. Apply this rule at each consecutive (overlapping) block of three in a given bi-infinite sequence of zeros and ones to obtain an other such sequence. Clearly the map from sequences to sequences is uniquely determined by the block description. The global map is onto but not 1-1. It is example of a one-dimensional cellular automaton. Turns out that the evolution under the iteration of this cellular automaton is "chaotic" in the usual sense of being sensitive to initial state. This that can (and will) be made rigorous. This rule is as a factor in many other cellular automata exhibiting more complex behavior and its higher dimensional version can be analyzed to a good extent. An illustration of an evolution from a uniform Bernoulli distributed initial state on a torus is in Figure 1 (top row is the initial state and time runs downwards).

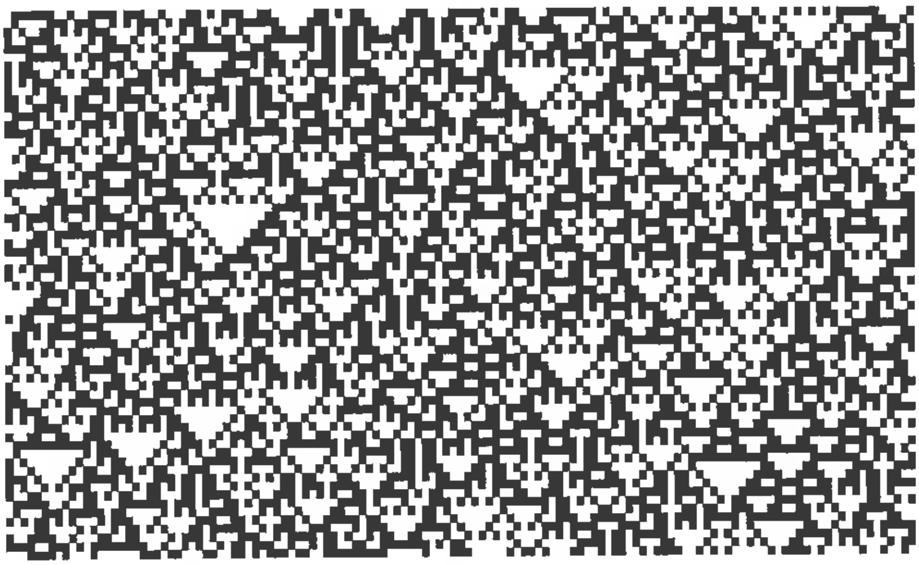

Figure 0.1. Evolution of the elementary cellular automaton Rule 90.

Example 0.4: The Game of Life ([BCG]). With Conway's insight one can propose the following cellular automaton rule even without experimenting with it. Consider again zeros and ones on $\mathbf{Z}^2$ and let the update of the center in a 3×3 neighborhood be 1 if it was 0 and exactly three of it's eight neighbors are ones. If not keep the zero. If the center was one and either two or three of its neighbors are ones keep it as one otherwise flip. Apart from having an entertaining aspect to it (watch it!) this rule has depth. It has been shown to be capable of universal computation i.e. it is possible to implement a Turing machine in this rule thereby computing any recursive function (again see [BCG], the implementation is not efficient but serves the purpose of theoretical analysis). Since the halting problem of the Turing machine is undecidable we thus see that decidability issue enters to the study of symbolic dynamics early on.

1. Basic ergodic theory

1.1. First notions and the ergodic theorem

Ergodic theory is the study of asymptotic properties of measurable dynamical systems. In order to do symbolic dynamics we need the basic concepts of ergodic theory.

Let X be a metric space and $\mathcal{B}$ its Borel σ-algebra i.e. the collection of all subsets which are Borel measurable. Suppose that the space X carries a measure μ so that the triple $(X, \mathcal{B}, \mu)$ is a measure space.

Definition 1.1.: *A map $T : X \to X$ is a* **measurable transformation** *of the space if $T^{-1}B = \{x \in X | \, Tx \in B\}$ is in $\mathcal{B}$ for any $B \in \mathcal{B}$. If moreover $\mu\left(T^{-1}B\right) = \mu(B)$ for all $B \in \mathcal{B}$ the transformation is said to be* **measure preserving** *and μ is an* **invariant measure**. *The quadruple $(X, \mathcal{B}, \mu, T)$ is a (measurable)* **dynamical system**.

Note that the Definition does not assume that T is invertible. The key ingredient here is the preservation of the space and the measure since they allow arbitrary iteration of the map. Indeed one is usually interested in the asymptotic distribution of the **(bi-infinite) orbit** of a point x, $O(x) = \{\ldots, T^{-2}x, T^{-1}x, x, Tx, T^2x, \ldots\}$, in the space X (or perhaps e.g. the one-sided orbit like the **forward orbit** $O_+(x) = \{x, Tx, T^2x, \ldots\}$). If an orbit satisfies $T^k x = x$ for some x and $k > 1$ we call x a **periodic point** with period k. If $k = 1$ then x is a **fixed point.** Physically one should think X as the **phase space** of a system and T as the action of the physical process on it.

If the measure is removed from the Definition we obtain just a topological dynamical system. Even then interesting questions can be asked, typically about the orbit structure. Periodic orbits are often of significance and so are dense orbits. A system in which there is $x \in X$ with a dense orbit is called **topologically transitive.**

If the space X carries a differentiable structure say it is a Riemannian manifold and T is a diffeomorphism the study of the system is usually called smooth dynamics. In this context in particular it may become useful to replace T by a one-parameter semigroup of transformations $\{T_t\}$ each preserving the space X. $\{T_t\}$ is called a **flow** on X.

To be able to physically interpret the results one usually prefers the situation where the measure μ is a probability measure i.e. $\mu(X) = 1$. To get this it is natural to require that X is compact. The existence of an invariant measure is a consequence of Markov-Kakutani Fixed Point Theorem (or more elementary arguments). But the problem of finding *all* invariant measures is in general highly non-trivial.

Example 1.2.: Consider the mapping of the unit interval $X = [0,1)$ (with the usual Borel σ-algebra) defined by $T : x \mapsto 2x \pmod 1$. It is easy to see that the map preserves the Lebesque measure λ. But it also preserves many singular measures for example δ_0, the unit mass at origin. Can you find them all?

Example 1.3.: Let X be the unit circle and T rotation by an angle α. Alternatively one can think of multiplication of complex numbers of absolute value one (or more generally multiplication on any compact group). Again the Lebesque measure is preserved. Depending on whether α is a rational or irrational multiple of π singular measures may be preserved. The behavior of the system is critically dependent on the irrationality of α/π. It is straightforward to see that rational orbits $\{T^i x\}_{i=-\infty}^{\infty}$ will always be finite. But if α/π is irrational is the orbit dense in X? Clearly if it is dense for one $x \in X$ it is for all x. And one can ask more: is the orbit uniformly distributed in X i.e. does the orbit visit two intervals of equal length in X with equal frequency?

Example 1.4.: Let X be as in Example 2.1. and

$$T : x \mapsto \begin{cases} 0, & \text{if } x = 0 \\ \mathrm{Frac}(1/x), & \text{if } x \neq 0. \end{cases}$$

Here Frac denotes the decimal part of the argument: $1/x - \lfloor 1/x \rfloor$. This example is motivated by continued fractions expansion. Since this transformation is important in Diophantine approximation it was studied already by Gauss and although measures were not known at his time (around 1845) he figured out the absolutely continuous invariant measure (or rather the invariant density i.e. the Radon-Nikodym derivative of the measure with respect to the Lebesque measure). Can you do that?

Example 1.5.: Suppose that $\{(q_i, p_i) \mid i = 1, \ldots, n\}$ are a particle's coordinates and momentum in $\mathbf{R}^n$ and the particle obeys Hamiltonian dynamics i.e. there exists a function $H : \mathbf{R}^{2n} \to \mathbf{R}$ (which is typically the sum of kinetic and potential energies) and

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}, \qquad \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i} \qquad i = 1, \ldots, n.$$

By fixing the total energy H one can restrict the action to a compact surface. The equations above define a flow on this manifold. Moreover a standard result, Liouville's theorem, provides us with a (uniform) probability measure invariant under the semigroup action. Hamiltonian mechanics is a rich source of dynamical systems perhaps the simplest of them being the geodesic flow and the billiards.

In a way the starting point of modern dynamics is in the studies by Henri Poincaré in the late 19^{th} century. He was instrumental in shifting the attention from the behavior of individual orbits to the behavior of **ensembles of orbits**. For him one in particular owe the Poincaré Recurrence Theorem:

Theorem 1.6.: *Let $(X, \mathcal{B}, \mu, T)$ be a dynamical system and $B \in \mathcal{B}$, $\mu(B) > 0$. Then μ-almost all points of B return infinitely often to B under positive iteration of T.*

The notions above are some of the most fundamental ones which the ergodic theory tries to develop further. We now proceed to introduce the key theorem of the subject, the pointwise ergodic theorem.

Suppose that we are observing a physical system and we are interested in how much time the system spends in a particular state B in the phase space X. We do not in general know the probability of the system being in the state B but we can count the frequency that it visits the state. Formally $B \in \mathcal{B}$ and we can record the value of the frequency (temporal density)

$$A_n(x) = \frac{1}{n} \sum_{i=0}^{n-1} \chi_B(T^i x)$$

where χ_B is the indicator function of the set B and x is the initial state of the evolution. Suppose that we have (some) invariant probability measure μ and thereby we can talk about a dynamical system $(X, \mathcal{B}, \mu, T)$. It seems that in order for this invariant measure to be physically relevant the frequency above should approach the size of the set B as measured by μ i.e. $\mu(B)$. Obviously this cannot be true for any invariant measure (think e.g. Example 1.2.) but it could hold under some assumptions on the system. It was Ludwig Boltzmann's idea that the 'ergodic hypothesis' should be true i.e. the spatial and temporal averages should agree under mild assumptions on the system.

Definition 1.7.: *The dynamical system $(X, \mathcal{B}, \mu, T)$ is* **ergodic** *if the only invariant sets i.e. sets for which it holds that $T^{-1}B = B$ have μ-measure zero or one. Equivalently the only invariant measurable functions i.e. $f(Tx) = f(x)$ for a.e. x are constant functions.*

The equivalence is simple: on one hand the function χ_B is invariant if B is so if all invariant functions are constant $\mu(B)$ is zero or one. On the other hand if T is ergodic then the set $B_r = \{x \in X |\ f(x) > r\}$ is measurable and invariant if f is such. So $\mu(B_r)$ is zero or one. But if f is not constant $\mu(B_r)$ can be non-trivial, a contradiction.

Remarks: 1. Intuitively ergodicity means that non-trivial subsets of the space cannot be fixed by the transformation but they are moved around in some fashion. This is a weak assumption but clearly a prerequisite for the ergodic hypothesis – if nontrivial parts of the space could be fixed the principle could be immediately violated. Alternatively one can think of ergodicity as indecomposability of the transformation: it cannot be split into simpler disjoint actions on the same space.
2. It can be shown that the set of invariant measures is the convex hull of the ergodic ones. If this set is a singleton the system is called **uniquely ergodic**.

Theorem 1.8.: *Consider the dynamical system $(X, \mathcal{B}, \mu, T)$ and let $B \in \mathcal{B}$. Let the frequency of visits to B by time n, $A_n(x)$ be as above. Then for μ-almost every $x \in X$ the limit*

$$A(x) = \lim_{n\to\infty} A_n(x)$$

exists. If the system is ergodic then $A(x) = \mu(B)$ μ-almost surely.

Remarks: 1. Note that $A(x)$ is necessarily an invariant function. Therefore if the system is ergodic $A(x)$ must be constant.
2. Note that if the system $(X, \mathcal{B}, T, \mu)$ is ergodic and $\mu(B) > 0$ for any open ball (suppose X is metric) then in particular μ-almost every orbit visits any such ball infinitely often and the orbit is dense in X. The converse is of course not true – in general the density of an orbit says little about the frequency of its visits to an open set.

Proof: Let $\overline{A}(x) = \limsup_{n\to\infty} A_n(x)$ and $\underline{A}(x)$ analogously. We will show that

$$\int_X \overline{A}(x)\mu(dx) \leq \mu(B). \tag{1.1}$$

Note if this can be shown then by considering the complement of B in X we will also get

$$\int_X \underline{A}(x)\mu(dx) \geq \mu(B).$$

But these two inequalities together imply that

$$\int_X \left(\underline{A}(x) - \overline{A}(x)\right)\mu(dx) \geq 0$$

which implies $\underline{A} = \overline{A}$ μ-almost surely as by the definition it holds that $\underline{A}(x) \leq \overline{A}(x)$. Along the way here we also get that

$$\int_X A(x)\mu(dx) = \mu(B)$$

as well as the claim that $A(x)$ almost surely agrees with $\mu(B)$ in the ergodic case (this is the only place where the ergodicity enters the argument).

Given $\epsilon > 0$ define for any $x \in X$ a convergence epoch as follows

$$\tau(x) = \min\left\{n > 0 \mid A_n(x) \geq \overline{A}(x) - \epsilon\right\}.$$

Depending on whether $\tau(x)$ is uniformly bounded or not we argue two cases.

1. Suppose that $\tau(x) < M$ for μ-almost every $x \in X$ for some finite M. Given $n > 0$ divide the interval $\{0, 1, 2, \ldots, n-1\}$ into non-overlapping consecutive blocks

$$\{x, Tx, T^2x, T^3x, \ldots, T^{\tau(x)-1}x\}, \{y, Ty, T^2y, T^3y, \ldots, T^{\tau(y)-1}y\}, \ldots$$

where $y = T^{\tau(x)}x$ and $T^{\tau(y)}y$ is the starting point of the next block etc. On each of these the density of visits to B is at least $\overline{A}(x) - \epsilon$. Note that e.g. on the second block the density is at least $\overline{A}(y) - \epsilon$, but by the T-invariance of $\overline{A}$ this equals to $\overline{A}(x) - \epsilon$. By our assumption we can choose the lengths of these blocks to be bounded by M. So on the entire interval the density of visits to B satisfies

$$A_n(x) \geq \frac{n-M}{n}\left(\overline{A}(x) - \epsilon\right),$$

where the term $-M$ enters since we have no control over the length of the last block except that we know it cannot exceed M. Integrating over the space this yields

$$\frac{1}{n}\sum_{i=0}^{n-1}\int_X \chi_B\left(T^i x\right)\mu(dx) \geq \left(1 - \frac{M}{n}\right)\left(\int_X \overline{A}d\mu - \epsilon\right).$$

As T preserves μ all the integrals on the left equal to $\int_X \chi_B d\mu = \mu(B)$. So by letting n pass to infinity we get

$$\mu(B) \geq \int_X \overline{A}d\mu - \epsilon \tag{1.2}$$

and (1.1) follows.

2. In case τ is not uniformly bounded we fix ϵ, define the bad set $C = \{x \in X \mid \tau(x) > M\}$ where M is such that $\mu(C) < \epsilon$. We extend the test set B by defining $B' = B \cup C$ and define $\tau'(x)$ to equal to $\tau(x)$ on C^c and 1 on C. Using τ' in the the block argument above we decompose the orbit on $\{0, 1, 2, \ldots, n-1\}$. Through this argument we again obtain (purely combinatorially i.e. without the measure) for $A'_n(x) = \frac{1}{n}\sum_{i=0}^{n-1}\chi_{B'}(T^i x)$ and for the original $\overline{A}$ the lower bound

$$A'_n(x) \geq \frac{n-M}{n}\left(\overline{A}(x) - \epsilon\right).$$

Integrating as above we get (1.2) with B replaced by B' on the left. Taking into account that $\mu(B') \leq \mu(B) + \epsilon$ we capture (1.2) with 2ϵ. ∎

Remark: This is neither the original proof by Birkhoff (1931) nor the classical proof using maximal lemma by Garcia (1965). In some sense it is the most natural proof and follows an idea of Kamae. With some further, more standard bookkeeping this result can be generalized to its full form (first prove to non-negative, bounded functions and then to the possibly unbounded positive and negative parts of a general L^1-function using monotone convergence). What results is known as the Pointwise Ergodic Theorem (of Birkhoff):

Theorem 1.9.: *Given a dynamical system $(X, \mathcal{B}, \mu, T)$ and an integrable function f on $(X, \mathcal{B}, \mu)$ ($\int_X |f| d\mu < \infty$) it holds that*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} f(T^i x) = f^*(x) \qquad \mu\text{-a.s.}$$

where f^* is an invariant integrable function. Moreover $\int_X f^* d\mu = \int_X f d\mu$ and if the system is ergodic then the function f^* is μ-almost surely constant.

Birkhoff's theorem was preceded by a string of related results which are all implied by this synthesis. As an illustration we present one of them, Borel's Theorem on Normal Numbers.

Corollary 1.10.: *Almost all numbers on $[0,1)$ are normal to the base 2 i.e. the frequency of ones in the binary expansion is 1/2.*

Proof: Consider the dynamical system $(X, \mathcal{B}, \lambda, T)$ defined in Example 1.2. T is ergodic (exercise). Let $X' \subset X = [0,1)$ be the set of numbers with unique binary expansion. Since $X \setminus X'$ is countable X' is of full measure.

Let $x \in X'$ have the binary representation $\sum_{j=1}^{\infty} a_j 2^{-j}$. Then

$$\chi_{[1/2,1)}\left(T^i x\right) = \begin{cases} 1, & \text{if } a_i = 1 \\ 0, & \text{if } a_i = 0 \end{cases}$$

and the summation

$$\sum_{i=0}^{n-1} \chi_{[1/2,1)}\left(T^i x\right)$$

gives the number of ones within the first n digits in the representation. But Theorem 1.8. implies that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\chi_{[1/2,1)}\left(T^i x\right)=\int_X \chi_{[1/2,1)}(x)dx=\frac{1}{2}$$

almost surely with respect to the uniform (Lebesque) measure. ∎

Remark: There is nothing special about base 2, the same result holds for any integral base. However one can ask subtle questions about the simultaneous normality to several bases. And one would also be interested in base-free analysis of the reals. One vehicle for this is the continued fraction expansion above.

The Corollary also illustrates the important notion of genericity. A point $x \in X$ is **generic** for $(X, \mathcal{B}, \mu, T)$ if the ergodic theorem holds along the the orbit from x i.e. it gives the right average. In other words

$$\frac{1}{n}\sum_{i=0}^{n-1}\delta_{T^i x}\longrightarrow \mu$$

in the sense of weak convergence of measures. So normality is a generic property of a real number. But our result does not say anything about any particular number – for example we do not know whether the decimals of π or e are equally frequent!

There is no general rate result to sharpen the ergodic theorem. If the dynamical system is uniquely ergodic then the convergence is uniform.

Theorem 1.11.: *$(X, \mathcal{B}, \mu, T)$ is uniquely ergodic if every $x \in X$ is generic for μ and the Ergodic Theorem holds uniformly:*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}f\left(T^i x\right)=\int_X f d\mu$$

for all $f \in C(X)$ uniformly in x.

Proof: If uniformity fails in the ergodic theorem for some g then for some $\epsilon > 0$ and for all n there is $x(n)$ such that

$$\left|\frac{1}{n}\sum_{i=0}^{n-1}g\left(T^i x(n)\right)-\int g d\mu\right|\geq\epsilon .$$

Let $\nu_n = \sum_0^{n-1} \delta_{T^i x(n)}$ so $\left|\int g d\nu_n - \int g d\mu\right| \geq \epsilon$. Suppose ν is some weak limit of $\{\nu_n\}$. It is invariant but also $\left|\int g d\nu - \int g d\mu\right| \geq \epsilon$ which contradicts the uniqueness. ∎

Example 1.12.: Consider the rotation (on the complex unit circle) of Example 2.3. Another way of declaring the rationality of the rotation is to say that $a = e^{i\alpha}$ is a root of unity i.e. there is p s.t. $a^p = 1$. If $f(z) = z^p$ then $f \circ T = f$ yet f is not almost everywhere constant so the system isn't ergodic.

If on the other hand a in not a root of unity pick an invariant function in L^2 of the circle. Every L^2 function has a Fourier series so let $f(z) = \sum_{\mathbf{Z}} a_n z^n$. Then $f(Tz) = f(az) = \sum_{\mathbf{Z}} a_n a^n z^n$. From the invariance of f we conclude that $a_n(a^n - 1) = 0$ for all n i.e. $f(z) = a_0$ almost surely. Since indicator functions are in L^2 we deduce that all invariant sets have measure either zero or one.

So the rotation on the circle is ergodic iff α/π is irrational. This result answers all the questions posed in Example 1.3. If α/π is irrational all the orbits are uniformly distributed on the circle and in particular dense. In fact the system is uniquely ergodic (show this!). Later in the context of entropy we will see how much this transformation mixes the points on the circle.

Without the measure one cannot talk about ergodicity but something analogous can be defined. A homeomorphism T on a compact metric space X is **minimal** if all bi-infinite orbits are dense i.e. $\overline{O(x)} = X \ \forall x \in X$. It can be easily shown that (see [Wa])

Proposition 1.13.: *Minimality of T is equivalent to the only closed and T-invariant subsets of X being $\emptyset$ and X. Moreover if f is continuous on X and $f \circ T = f$ then f is constant.*

Ergodicity is near the bottom level in the hierarchy mixing/chaos. By this we mean the ordering of dynamical systems according to how faithful their evolutions are to the initial data. The higher the system is in the hierarchy the more it mixes up the order in the initial state. Alternatively one can think of random or pseudorandom components being more dominant in the behavior as one ascends in the hierarchy. Purely random systems are at the very top.

In order to define the next level above ergodicity in this hierarchy it is useful to recognize clearly the following fact.

Proposition 1.14.: *Suppose that $(X, \mathcal{B}, \mu, T)$ is a dynamical system and $A, B \in \mathcal{B}$. Then the system is ergodic iff*

$$\frac{1}{n}\sum_{i=0}^{n-1} \mu\left(T^{-i}A \cap B\right) \to \mu(A)\mu(B) \qquad \mu\text{–a.s.} \tag{1.3}$$

Proof: Applying Theorem 1.8. to χ_A and multiplying the result by χ_B gives

$$\frac{1}{n}\sum_{i=0}^{n-1} \chi_A\left(T^{-i}x\right)\chi_B \to \mu(A)\chi_B \qquad \mu\text{–a.s.}$$

which in turn yields $(*)$ by the dominated convergence theorem.

For the reverse let $A = B$ be an invariant set. Then (1.3) implies that

$$\frac{1}{n}\sum_{i=0}^{n-1} \mu(B) \to \mu(B)^2,$$

i.e. necessarily $\mu(B) = 0$ or 1. ∎

The strengthening of ergodicity that we have in mind is **(strong) mixing** which is defined by requiring that

$$\mu\left(T^{-i}A \cap B\right) \to \mu(A)\mu(B) \tag{1.4}$$

for all measurable subsets A and B and $i \to \infty$. Clearly mixing implies ergodicity by the result that we just established (1.3). The converse is not true – in fact the irrational rotation is an example of a dynamical system which is ergodic but not strong mixing (exercise).

In terms of probability concepts mixing is the same as asymptotic independence. Independence would be simply $\mu\left(T^{-i}A \cap B\right) = \mu(A)\mu(B)\ \forall i \neq 0$. In the context of deterministic dynamics this notion would seem to be too strong to prevail to any reasonable extent but that isn't actually the case. However it is usually hard work to show the existence of such component in a dynamical system.

A natural extension of (1.4) would be the ***k*-fold mixing**

$$\mu\left(T^{-i_1}A_1 \cap T^{-i_2}A_2 \cap \cdots \cap T^{-i_{k-1}}A_{k-1} \cap A_k\right) \to \mu(A_1)\mu(A_2)\cdots\mu(A_k)$$

for $(i_1, i_2, \ldots, i_{k-1})$ escaping to infinity. Rokhlin conjectured that this should hold for all k if it holds for $k = 2$. This is one of the most important outstanding problems in Ergodic Theory. Little is known how widely this is true. In the context of $\mathbf{Z}^d$-actions we will see an example where the conjecture fails.

Inbetween ergodicity ("rigidity") and independence ("looseness") there is a rich hierarchy of mixing. One can characterise the degree in which the space gets shuffled in a variety of ways. These include correlation decay rate, validity of the Central Limit Theorem, triviality of the tail σ-algebra etc. Some of these we will encounter in subsequent sections.

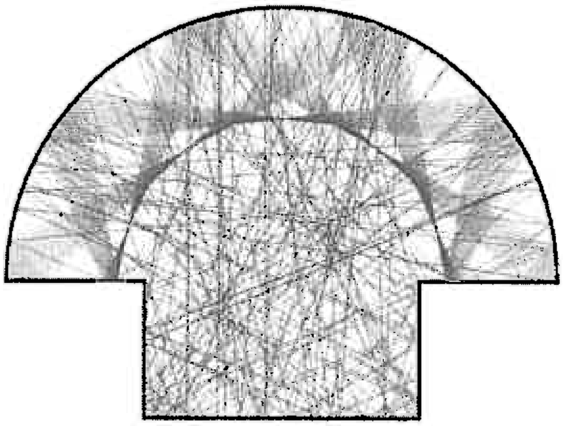

Figure 1.1. Billiards: the coexistence of periodic and chaotic orbits ([DG]).

References. A nice introduction to the basic questions in ergodic theory is Keane's article in [BKS]. Further reading could be a general treatise on ergodic theory such as [Pe] or [Wa]. The ergodic theorem presented here (it's full proof can be found in [KW]) is just the beginning of a long story. A nearly upto date account of ergodic theorems is Krengel's book [Kr]. However the recent results on ergodic theorems along subsequences of the integers are missing from there. In particular it is known that Theorem 1.9. also holds along primes! This is a particularly hard result (due to Bourgain) since it does not satisfy the natural condition that the subsequence along which the iterates are picked has positive density among the integers (by the Prime Number Theorem primes have zero density).

1.2. Entropy

1.2.1. Measure theoretic entropy

Entropy is a fundamental notion related to the degree of disorder in a system. It is defined in slightly different way e.g. in the context of statistical mechanics, graph theory and information theory. Our treatment is closest to the last one.

Consider a probability space $(X, \mathcal{B}, \mu)$. A finite **partition** $\mathcal{P}$ of a space X is its division to a collection of disjoint subsets $\{P_1, \ldots, P_n\}$ whose union is X (on aside note that it is logical to denote a partition with a calligraphic letter as it generates a finite σ-algebra). In order to avoid pathologies we assume that the **atoms** P_i are measurable. If $\mathcal{P}' = \{P'_1, \ldots, P'_m\}$ is another partition of X we say that it **refines** $\mathcal{P}$ if each element of $\mathcal{P}$ is a union of elements in $\mathcal{P}'$ and write $\mathcal{P} \preceq \mathcal{P}'$. The **join** of the partitions

$$\mathcal{P} \vee \mathcal{P}' = \{P_i \cap P'_j |\ 1 \leq i \leq n,\ 1 \leq j \leq m\}$$

is again a partition (which refines both original partitions).

Given a probability space $(X, \mathcal{B}, \mu)$ we define the **entropy of a partition** as follows:

$$H(\mathcal{P}) = -\sum_{i=1}^{n} \mu(P_i) \ln \mu(P_i). \tag{1.4}$$

It is understood here that $0 \ln 0 = 0$. We usually choose the base of the logarithm to be e, sometimes 2 in which case it is indicated.

The functional form (1.4) is motivated by the requirements to find a function on the partition that is (i) non-negative, (ii) non-zero if the partition is non-trivial w.r.t. μ, (iii) continuous and symmetric in the arguments, (iv) has a maximum at the uniform distribution w.r.t. μ. The derivation of the form (2.4) can be found in many of the standard textbooks (check that (i)-(iv) are satisfied or derive the form from the first principles).

Furthermore one can define the **entropy of $\mathcal{P}$ with respect to $\mathcal{P}'$** as

$$H(\mathcal{P}|\mathcal{P}') = -\sum_{j=1}^{m} \mu(P'_j) \sum_{i=1}^{n} \mu(P_i|P'_j) \ln \mu(P_i|P'_j).$$

Here $\mu\left(P_i|P'_j\right) = \mu\left(P_i \cap P'_j\right)/\mu\left(P'_j\right)$ is the conditional probability for the events P_i and P'_j. Conditioning on sets of measure zero is avoided by removing these sets from the partition $\mathcal{P}'$.

The conditional entropy is important because of its intuitive content. Suppose that we are told in which atom of $\mathcal{P}'$ a point x of X is in. The conditional entropy $H(\mathcal{P}|\mathcal{P}')$ tells us how much residual uncertainty there is if we would also like to guess where x is in terms of $\mathcal{P}$. If $\mathcal{P} = \mathcal{P}'$ upto sets of μ-measure zero there is no uncertainty and if $\mathcal{P}'$ is the trivial σ-field $\{\emptyset, X\}$ the uncertainty is maximal i.e. $H(\mathcal{P})$.

The following statements are not hard. The formal proofs are left as exercises and one should also find the intuitive reasons why the statements are true.

Proposition 1.15.: *If $\mathcal{P}, \mathcal{P}'$ and $\mathcal{P}''$ are finite partitions then*
(i) $H(\mathcal{P} \vee \mathcal{P}') = H(\mathcal{P}) + H(\mathcal{P}'|\mathcal{P})$.
(ii) $H(\mathcal{P} \vee \mathcal{P}') \leq H(\mathcal{P}) + H(\mathcal{P}')$.
(iii) $\mathcal{P}' \preceq \mathcal{P}'' \Rightarrow H(\mathcal{P}|\mathcal{P}'') \leq H(\mathcal{P}|\mathcal{P}')$.

If T is measure preserving transformation on X and $\mathcal{P}$ is a partition as above then it is natural to denote by $T^{-1}\mathcal{P}$ the partition $\left\{T^{-1}P_1, \ldots, T^{-1}P_n\right\}$. Note that as T preserves μ H is T-invariant: $H\left(T^{-1}\mathcal{P}\right) = H(\mathcal{P})$.

We use the notation $\bigvee_{i=0}^{k-1} \mathcal{P}_i$ to refer to the k-fold join of the k partitions $\mathcal{P}_i$, $0 \leq i \leq k-1$. A further shorthand used subsequently is $\mathcal{P}_n^m = \bigvee_{i=n}^m T^{-i}\mathcal{P}$. All the $x \in X$ with the property that $x \in P_{i_0}$, $Tx \in P_{i_1}, \ldots,$ $T^{(k-1)}x \in P_{i_{k-1}}$ belong to the same atom of the partition $\mathcal{P}_0^{k-1}$. The index sequence $\{i_0, i_1, \ldots, i_{k-1}\}$ that specifies this (cylinder) set is the **k-name of x**.

Definition 1.16.: *The* **entropy of the transformation T with respect to the partition $\mathcal{P}$** *is given by*

$$h(T, \mathcal{P}) = \lim_{k \to \infty} \frac{1}{k} H\left(\bigvee_{i=0}^{k-1} T^{-i}\mathcal{P}\right).$$

The **entropy** *of T is the number*

$$h(T) = \sup_{\mathcal{P}} h(T, \mathcal{P})$$

where the supremum is taken over all finite partitions of X.

Remarks: 1. To distinguish the measure theoretic entropy from it's topological counterpart as well as to clarify with respect to which measure it is taken we will use the notation $h_\mu(T)$ when necessary.

2. By our knowledge of H we know that both the entropies defined are non-negative numbers and for the latter infinity is not excluded because of the supremum. We will now proceed to argue the existence and finiteness of the former.

Proposition 1.17.: *The limit* $\lim_{k\to\infty} \frac{1}{k} H\left(\bigvee_{i=0}^{k-1} T^{-i}\mathcal{P}\right)$ *exists for finite* $\mathcal{P}$.

Proof: Let $a_k = H\left(\bigvee_{i=0}^{k-1} T^{-i}\mathcal{P}\right)$. By Proposition 1.15.

$$\begin{aligned} a_{k+p} &= H\left(\bigvee_{i=0}^{k+p-1} T^{-i}\mathcal{P}\right) \\ &\leq H\left(\bigvee_{i=0}^{k-1} T^{-i}\mathcal{P}\right) + H\left(\bigvee_{i=k}^{k+p-1} T^{-i}\mathcal{P}\right) \\ &= a_k + H\left(\bigvee_{i=0}^{p-1} T^{-i}\mathcal{P}\right) = a_k + a_p \end{aligned}$$

i.e. $\{a_k\}$ is a subadditive sequence. But the limit $\lim_{k\to\infty} a_k/k$ exists for all such sequences and equals to $\inf_k a_k/k$ by the following Lemma. ∎

Lemma 1.18.: *A sequence of real numbers* $\{a_k\}_0^\infty$ *is subadditive if the inequality* $a_{k+p} \leq a_k + a_p$ *holds for all indices* k, p. *For such sequence* $\lim_{k\to\infty} a_k/k = \inf_k a_k/k$ *and in particular if the sequence is bounded from below the limit is finite.*

Proof: Given $p > 0$ any natural k can be written as $k = lp + i$ with $0 \leq i < p$. First we note that

$$\frac{a_k}{k} \leq \frac{a_i}{lp} + \frac{a_{lp}}{lp} \leq \frac{a_i}{lp} + \frac{la_p}{lp} = \frac{a_i}{lp} + \frac{a_p}{p}$$

which implies that $\limsup \frac{a_k}{k} \leq \frac{a_p}{p}$. So in particular $\limsup \frac{a_k}{k} \leq \inf \frac{a_p}{p}$. But clearly $\inf \frac{a_p}{p} \leq \liminf \frac{a_n}{n}$ and the inf is sandwiched to a point. ∎

From Proposition 1.15. one can quickly derive properties of the two entropies introduced above. In particular a little algebra with partitions shows that

Proposition 1.19.: *If* $\mathcal{P}$ *and* $\mathcal{P}'$ *are finite partitions then*
(i) $h(T, \mathcal{P}) \leq H(\mathcal{P})$.
(ii) $h(T, \mathcal{P} \vee \mathcal{P}') \leq h(T, \mathcal{P}) + h(T, \mathcal{P}')$.
(iii) $\mathcal{P} \preceq \mathcal{P}' \Rightarrow h(T, \mathcal{P}) \leq h(T, \mathcal{P}')$.
(iv) For $k > 0$, $h(T^k) = kh(T)$.

The following result gives an alternative way of looking at the entropy with respect to a partition.

Theorem 1.20.: $h(T,\mathcal{P}) = \lim_{k\to\infty} H\left(\mathcal{P} \mid \mathcal{P}_1^k\right) = H\left(\mathcal{P}|\mathcal{P}_1^\infty\right)$.

Proof: By Proposition 1.15. (iii) the limit is non-increasing and since it is bounded from below it exists. We will show by induction that

$$H\left(\mathcal{P}_0^{k-1}\right) = H(\mathcal{P}) + \sum_{j=1}^{k-1} H(\mathcal{P}|\mathcal{P}_1^j). \tag{1.5}$$

Clearly the formula is true for $k=1$. Suppose it is true for $k=p$. Then

$$\begin{aligned} H\left(\mathcal{P}_0^p\right) &= H\left(\mathcal{P}\vee\mathcal{P}_1^p\right) = H\left(\mathcal{P}_1^p\right) + H\left(\mathcal{P}|\mathcal{P}_1^p\right) \\ &= H(\mathcal{P}_0^{p-1}) + H\left(\mathcal{P}|\mathcal{P}_1^p\right) \\ &= H(\mathcal{P}) + \sum_{j=1}^{p-1} H(\mathcal{P}|\mathcal{P}_1^j) + H\left(\mathcal{P}|\mathcal{P}_1^p\right) \\ &= H(\mathcal{P}) + \sum_{j=1}^{p} H(\mathcal{P}|\mathcal{P}_1^j) \end{aligned}$$

where we have used Proposition 1.15. (i), T-invariance of μ and the induction assumption. So the formula holds for $k=p+1$ as well.

Dividing (1.5) by k yields

$$\frac{1}{k}H(\mathcal{P}_0^{k-1}) = \frac{1}{k}H(\mathcal{P}) + \frac{1}{k}\sum_{j=1}^{k-1} H(\mathcal{P}|\mathcal{P}_1^j)$$

which implies the result since the Cesaro limit agrees with the usual limit whenever the latter exists. ∎

Partitions can be good or bad (or ugly...). By this we mean that the partition either splits the space in a reasonable way so that knowing in which atom a point is gives some information on where it has been and where it will be going under the iteration of the transformation. If it holds that

$$\bigvee_{i=-\infty}^{\infty} T^{-i}\mathcal{P} = \mathcal{B}$$

where $\mathcal{B}$ is the usual Borel σ-algebra and the equivalence is upto sets of measure zero then we say that $\mathcal{P}$ is a **generating partition.** Hence a generating partition can distinguish the orbits upto a set of initial states of measure zero.

The key results concerning computation of entropy are the following theorems. The first one is due to Kolmogorov and Sinai.

Theorem 1.21.: *If $\mathcal{P}$ is a finite generator then $h(T) = h(T, \mathcal{P})$.*

The existence of the generator required above was resolved by Krieger.

Theorem 1.22.: *If T is ergodic and of finite entropy then a finite generator exists.*

The proofs of both theorems are non-trivial but can be found in the standard treatises of ergodic theory. Actually finding a generator albeit finite may still be tricky. However in the examples that we will consider the generation will be fairly obvious.

To get at least one entropy nailed down we will take a final look at the rotation of the circle.

Example 1.23.: Let us continue the Examples 1.3. and 1.12. If α/π is rational then all orbits are periodic of the same period i.e. $T^p x = x$ for some natural p and all x. But then T^p is identity hence its entropy is zero. Furthermore by Proposition 1.19. (iv) $0 = h(T^p) = ph(T)$. More generally a measure preserving transformation on a finite space has zero entropy.

If α/π is irrational the orbits are dense. Pick a partition $\mathcal{P}$ consisting of two semiclosed half-circles. As the sets in the partitions in $\{T^{-i}\mathcal{P}\}$ have their endpoints dense in the circle it is clear that $\mathcal{P}_0^\infty$ generates the Borel algebra. So we have a one-sided generator (which surely is a generator), a strong property of a system! So using theorems 1.20. and 1.21. we get $h(T) = h(T, \mathcal{P}) = H(\mathcal{P}|\mathcal{P}_1^\infty) = H(\mathcal{P}|\mathcal{B})$. But the conditional entropy of any partition with respect to the full Borel algebra must be zero. Hence the rotation on the circle is always of zero entropy. It is in this sense that it is rigid and not chaotic at all.

The argument above on one-sided generators is useful to keep in mind in checking the zero-entropy case in general.

Those familiar with information theory may already have noticed a connection which is precisely formulated in a result by Shannon, MacMillan and Breiman.

Theorem 1.24.: *Given a probability space $(X, \mathcal{B}, \mu)$, an ergodic transformation T on it and a finite partition $\mathcal{P}$ let $B_k(x)$ denote the atom in the partition $\mathcal{P}_0^{k-1}$ where x belongs to. Then*

$$-\frac{1}{k} \ln \mu\left(B_k(x)\right) \longrightarrow h(T, \mathcal{P}) \qquad \mu\text{-a.s.}$$

and in $L^1(\mu)$.

Proof can be found in any standard text on ergodic or information theory. This is a useful approach to entropy. Indeed one can use it as a defining property: entropy with respect to a partition is the exponential rate at which the set of all x with the same k-name shrinks as k increases (given h is positive). For all but a zero measure exceptional set of sequences this rate is equal. This result is also called the **asymptotic equipartition property**.

1.2.2. Topological entropy

The concept of entropy can be introduced without any reference to the probabilities of the events involved. We now briefly show how this topological version is set-up.

Suppose that X is a compact topological space and T is a homeomorphism. Instead of measurable partitions we now deal with open covers. However the logic of the construction is similar. For example if $\mathcal{C}$ and $\mathcal{C}'$ are two covers we say that $\mathcal{C}'$ refines $\mathcal{C}$ ($\mathcal{C} \preceq \mathcal{C}'$) if every set in $\mathcal{C}'$ is contained in some set in $\mathcal{C}$ (i.e. $\mathcal{C}'$ is a **subcover** of $\mathcal{C}$). The join of the covers $\mathcal{C} \vee \mathcal{C}'$ is again defined as $\{C \cap C' |\ C \in \mathcal{C}, C' \in \mathcal{C}'\}$.

Given a cover $\mathcal{C}$ by $N(\mathcal{C})$ we denote the minimum of the cardinalities among its subcovers and suggestively record $H(\mathcal{C}) = \ln N(\mathcal{C})$. As it certainly holds that $N(\mathcal{C} \vee \mathcal{C}') \leq N(\mathcal{C})N(\mathcal{C}')$ we see that $H(\mathcal{C} \vee \mathcal{C}') \leq H(\mathcal{C}) + H(\mathcal{C}')$, a formula analogous to Proposition 1.15. (ii). When k-fold join is understood in the obvious way an argument parallel to the proof of Proposition 1.17. gives

Proposition 1.25.: $h(T, \mathcal{C}) = \lim_{k\to\infty} \frac{1}{k} H\left(\bigvee_{i=0}^{k-1} T^{-i}\mathcal{C}\right)$.

Definition 1.26.: *The* **topological entropy** *of the transformation T is defined as*

$$h_{top}(T) = \sup_{\mathcal{C}} h(T, \mathcal{C})$$

where $\mathcal{C}$ is any open cover of X.

Again $\mathcal{C} \preceq \mathcal{C}'$ implies that $h(T, \mathcal{C}) \leq h(T, \mathcal{C}')$ which is useful to know since if we can find (and we will) a sequence of covers that eventually refines any cover the limit along this sequence will catch the topological entropy.

The actual computation of topological entropy will be performed in the examples in the upcoming sections. To wrap up things here we will point out the close relation between the two entropies defined.

Theorem 1.27.: *Given a homeomorphism T on a compact metric space X the topological entropy is the maximum of the measure theoretic entropy taken over all T-invariant probability measures $\mathcal{M}_T$:*

$$h_{top}(T) = \sup_{\mu \in \mathcal{M}_T} h_\mu(T).$$

A statement of this type is called a **variational principle**. The supremum is actually achieved and an element of $\mathcal{M}_T$ which gives it is a **measure of maximal entropy**. As one might suspect they are of physical significance. We do not prove this result but in the context of Gibbs measures at the end of next section a stronger result will be established.

References. The basic introductory material is covered for example in [Pe] and [Wa]. Further elaboration can be found in [CFS]. A cute introduction to entropy is presented in [Bi]. Entropy was introduced to mathematics by Kolmogorov in his studies on the sizes of function spaces. In the fifties and sixties he an Sinai formulated the most natural definitions in the dynamics context. Part of the reason why the entropy has attained such central position in ergodic theory is due to the path breaking result by Ornstein showing that entropy is a complete isomorphism invariant for Bernoulli shifts. The subsequently polished Bernoulli theory is at the core of all of ergodic and probability theory and plays a key role in the context of systems with positive measure theoretic entropy as they have automatically Bernoulli factors. More on Bernoulli shifts in the next section.

2.1. Symbolic dynamics

We now proceed to present the basic framework of one-dimensional symbolic dynamics.

The fundamental object of study is a sequence space. One might want to think it as a "skeleton" of a richer dynamical set-up, say a flow on a manifold, when one only records in which atom of a partition the orbit is at the times of measurements.

Definition 2.1.: *Let the set $S = \{1, \dots, n\}$ be a finite set of* **symbols** *forming an* **alphabet**. *If we assign one of the symbols at each* **site** *of the one-dimensional integer lattice* $\mathbf{Z}$ *we obtain the set $X = S^{\mathbf{Z}}$ of* **configurations**.

Equipped with the product topology the set X becomes a topological space. Of the many equivalent ways of metrizing this topology the following is perhaps most useful to keep in mind

$$d(x,x) = 0 \ , \qquad d(x,y) = 2^{-\min\{|i| | \ x_i \neq y_i\}} .$$

Here x_i refers to the i^{th} coordinate of the sequence $x \in X$.

The set of configurations is a compact metric space for all n (exercise). Indeed it is homeomorphic to the closed unit interval hence it must be uncountable.

The basic dynamical operation on the configurations is the shift in the "time" direction.

Definition 2.2.: *The* **left (coordinate) shift** *on a configuration $x \in X$ is defined as $(\sigma x)_i = x_{i+1} \ \forall i$. The topological dynamical system $(X, \mathcal{B}, \sigma)$ is called the* **full n-shift**. *If we call the product measure π obtained by assigning the symbol s the probability p_s (so $p_s \geq 0$ and $\sum p_s = 1$) then the dynamical system $(X, \mathcal{B}, \pi, \sigma)$ is the* **Bernoulli-shift** *$B(p_1, \dots, p_n)$.*

The coordinate shift is often called a shift-action since it is thought that the group $\mathbf{Z}$ is acting as a translation on the configurations. The shift is a homeomorphism of the space X (exercise). Hence we have a continuous action on a space (with a somewhat unusual topology) and the definition just formulates their natural combination.

Example 2.3.: Given the full shift on n symbols we can easily compute the topological entropy. The topological entropy is again given by the entropy with respect to a generating cover. Such a cover is now readily available. Define $\mathcal{C} = \{C_1, \dots, C_n\}$

where $C_s = \{\{x_i\}_{-\infty}^{\infty} | \ x_0 = s\}$ (check that this cover which is also a partition is a generator). Then

$$h_{top} = h(\sigma, \mathcal{C}) = \lim_{k\to\infty} \frac{1}{k} \ln N\left(\mathcal{C}_0^{k-1}\right) = \lim_{k\to\infty} \frac{1}{k} \ln n^k = \ln n.$$

An argument this formal is of course overkill – here it just serves to unwrap the definitions and results given. The idea of topological entropy is that it counts the exponential rate at which the number of orbits of length k grows as k increases. As the the full shift on n symbols has exactly n^k orbits of length k the topological entropy must be the given number.

The measure-theoretic entropy of the Bernoulli shift $B(p_1, \ldots, p_n)$,

$$h_\pi = -\sum_{i=1}^{n} p_i \ln p_i$$

follows directly from Definition 1.14. and Theorem 1.19. (exercise). This set-up is actually so simple that we can immediately verify the variational principle.

Proposition 2.4.: *The full shift on n symbols has a unique measure of maximal entropy, the uniform product measure.*

Proof: Let $\mathcal{P}$ be the natural generator above for the entropy. If μ is any measure of maximal entropy then using the standard notation where $\mathcal{P}_0^{k-1}$ stands for the k-fold join of the partitions we have

$$\ln n \le h_\mu \le \frac{1}{k} H_\mu\left(\mathcal{P}_0^{k-1}\right) \le \frac{1}{k} \ln n^k = \ln n.$$

Here we have used the facts that $H_\mu\left(\mathcal{P}_0^{k-1}\right)/k$ decreases to the entropy and that the entropy function H is maximized at a unique point, the uniform distribution (check). So $H_\mu\left(\mathcal{P}_0^{k-1}\right) = k \ln n$. At the maximum of H each atom in $\mathcal{P}_0^{k-1}$ has measure n^{-k}. The argument is independent on k hence the measure μ must be the product measure. ∎

The full shift is rather simple object. However things become a great deal more interesting if we consider non-trivial closed and shift-invariant subsets of X instead and the shift-action on them. Indeed things get so interesting that as of now only dimension one can be comprehensively dealt with.

Suppose that we are given a $n \times n$ matrix A with each entry a_{ij} being either one or zero. Its powers are of course matrices with non-negative integer entries. When such matrix power is considered it is convenient to refer to its elements by $[A^n]_{ij}$.

Definition 2.5.: *Given the set* $X_A = \{x \in X| \; a_{x_i x_{i+1}} = 1 \; \forall i \in \mathbf{Z}\}$ *the system* $(X_A, \mathcal{B}, \sigma)$ *is called a* **subshift of finite type (SOFT)** *or a* **topological Markov chain.**
If the matrix A *is such that for each* i, j *there exists* $n > 0$ *such that* $[A^n]_{ij} > 0$ *we call the SOFT* **irreducible.** *If there is a universal* n *such that* $[A^n]_{ij} > 0 \; \forall i, j$ *we call the SOFT* **irreducible and aperiodic** *or simply* **primitive.**

Equipped with the inherited topology the set X_A is a compact metric space invariant under the shift-action. The full shift corresponds to all entries of A being one. If all entries of A are zero the space is of course empty but it can be empty for a non-trivial A as well, for example in the case of the matrix on the left (for all the matrices the alphabet is $\{0, 1\}$). The two matrices in the center (b,c) illustrate the non-uniqueness of A: they define the same SOFT.

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \qquad \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \qquad \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

Figure 2.1a, b, c, d.

The rightmost entry in the Figure 2.1. is an example of an irreducible and aperiodic matrix/subshift. Without irreducibility the statespace of the topological Markov chain could split into components which do not communicate, a redundant case which we do not wish to deal with. Aperiodicity rules out the situation where a state can only be visited at certain (say odd) times. For elaboration of these conditions see any text on Markov chains e.g. [KT]. Irreducibility and aperiodicity are subsequently running assumptions.

It is also useful to notice that dependency between the symbols in the sequence over longer but finite ranges can always be reduced to our set-up. If there is a dependency at distance L but no further (here we have $L = 2$) one just defines a larger alphabet from blocks of length $L - 1$. The dependency within this new alphabet can be described in terms of an adjacency matrix A.

With this preparation we are ready to compute the topological entropy of a general SOFT.

Theorem 2.6.: *Given a SOFT with an irreducible matrix* A *its topological entropy is*

$$h_{top} = \ln \lambda \tag{2.1}$$

where λ *is the largest positive eigenvalue of* A.

Proof: Suppose that A is a $n \times n$ matrix. The generator of the full n-shift $\mathcal{C} = \{C_i\}_1^n$, $C_i = \left\{\{x_j\}_{-\infty}^{\infty} \mid x_0 = i\right\}$ generates for the shift on X_A as well. Therefore $N_k = N\left(\mathcal{C}_0^{k-1}\right)$ gives the number of different orbits of length k under the shift. Consider the set all sequences which have a prescribed k-block starting at origin:

$$X_{i_0,\ldots,i_{k-1}} = \left\{\{x_j\}_{-\infty}^{\infty} \in X_A \mid x_0 = i_0, \ldots, x_{k-1} = i_{k-1}\right\}.$$

It is non-empty iff $a_{i_0 i_1} a_{i_1 i_2} \cdots a_{i_{k-2} i_{k-1}} = 1$. So the number of k-tuples for which $X_{i_0,\ldots,i_{k-1}}$ is non-empty is

$$N_k = \sum_{i_0,\ldots,i_{k-1}=1}^{n} a_{i_0 i_1} a_{i_1 i_2} \cdots a_{i_{k-2} i_{k-1}} = \sum_{i_0, i_{k-1}=1}^{n} \left[A^{k-1}\right]_{i_0 i_{k-1}} = \|A^{k-1}\|$$

where the sums are k-fold and we have defined the matrix norm by setting $\|B\| = \sum_{i,j} |b_{ij}|$. But then by the Spectral Radius Formula ([Rud])

$$\frac{1}{k} \ln N_k = \ln\left(\|A^{k-1}\|^{1/k}\right) \longrightarrow \ln \lambda.$$

The existence of the maximal non-negative (and simple) eigenvalue λ is due to the Perron-Frobenius Theorem for non-negative irreducible matrices (see Appendix 1). ∎

Remarks: 1. The crux of the proof, the identification of the matrix power, unveils the germ of **transfer matrices**. These were successfully used by E. Ising in the twenties in solving a slightly different problem and since then by many others especially in statistical mechanics. For a review of these in the dynamical systems context see [Bal]. We will discuss them again in the context of Gibbs measures.

2. A **periodic point** for the shift is a configuration satisfying $\sigma^p x = x$ for some $p > 0$. They have a number of interesting properties and although one might think that they are quite special and rare this is not really the case. In particular if N_p^{per} denotes the number of periodic orbits of period p then it can be shown under the assumptions of the Theorem that

$$\lim_{p \to \infty} \frac{1}{p} \ln N_p^{per} = \ln \lambda .$$

So there is exponentially as many periodic points as there are configurations! This property and the subsequent extension of the Theorem is left as an exercise. The result is not limited to just the shift space context. Moreover the periodic points are good representatives of the entire space X_A in the sense that they are nicely distributed. This can be formulated as follows: consider the empirical measures

$$\mu_m = \frac{1}{m} \sum_{p=0}^{m-1} \frac{1}{N_p^{per}} \sum_{\{x | \sigma^p x = x\}} \delta_x$$

where δ_x is a unit mass concentrated at the configuration x. Turns out that these probability measures μ_m converge to a limit that we have already encountered, the measure of maximal entropy!

Example 2.7.: Example 0.1. continued. The subshift in this example is defined by the matrix in Figure 2d. As its eigenvalues are $(1 \pm \sqrt{5})/2$ by the Theorem $h_{top} = \ln\left(\frac{1+\sqrt{5}}{2}\right) \in (0, \ln 2)$ and the system is occasionally called the one-dimensional golden mean.

We now briefly discuss certain invariant measures that a SOFT can have. To do this we first reformulate an ubiquitous probabilistic dynamical system.

Example 2.8.: First note that the Bernoulli shift corresponds to an independent process as one can interpret its dynamics as rolling a die with facet probabilities p_i. To generalize a bit suppose that we have a transition matrix P i.e. a $n \times n$ matrix the entries of which p_{ij} represent transitions (conditional probabilities of transitions) from state i to j. Because of this structure the row sums must be one and each element as a probability is non-negative. Assume furthermore that this matrix is irreducible and aperiodic. P and its transpose always have 1 as a maximal eigenvalue. Hence it fixes a vector p from the right: $pP = p$. That p is a positive probability vector ($p_i > 0$, $\sum p_i = 1$) is a consequence of the Perron-Frobenius Theory and normalization. To get a measure on the sequences of $X = \{1, \ldots, n\}^{\mathbf{Z}}$ we define on a cylinder set of length k

$$\mu_P\left(\{x \mid x_j = i_0, \ldots, x_{j+k-1} = i_{k-1}\}\right) = p_{i_0} p_{i_0 i_1} \cdots p_{i_{k-2} i_{k-1}} \qquad \forall j.$$

The measure on finite cylinder sets extends the usual way to all of the Borel algebra of X. Moreover the measure is clearly shift-invariant. The resulting dynamical system $(X, \mathcal{B}, \mu_P, \sigma)$ is called the (two-sided) (p, P)-**Markov shift**. Note that if we choose $p_{ij} \equiv p_j$ the Bernoulli case is recovered.

The measure entropy of this system is derived analogously to the Bernoulli case. We leave it as an exercise to the reader to verify the expression $-\sum_{i,j} p_i p_{ij} \ln p_{ij}$.

Turns out that there is a natural Markov shift associated to a SOFT. Given an irreducible matrix A by the Perron-Frobenius Theorem we can find strictly positive left and right eigenvectors u and v corresponding to the maximal eigenvalue λ: $uA = \lambda u$ and $Av = \lambda v$. Normalizing these by setting $\sum u_i v_i = 1$ we then define the (p, P)-Markov shift by setting

$$p_i = u_i v_i \qquad\qquad p_{ij} = \frac{a_{ij} v_j}{\lambda v_i}.$$

The left-invariance of p is immediate. By extending this to a measure on cylinder sets we obtain an invariant measure μ^* for the subshift (X_A, σ). Conversely one might think a SOFT as the skeleton of a Markov shift, it topologized version.

The star in the invariant measure refers to the fact that this measure, sometimes called the **Parry measure**, is a measure of maximal entropy:

$$\begin{aligned} h_{\mu^*} &= -\sum_{i,j} u_i v_i \frac{a_{ij} v_j}{\lambda v_i} \ln \frac{a_{ij} v_j}{\lambda v_i} \\ &= -\sum_{i,j} \frac{u_i a_{ij} v_j}{\lambda} (\ln a_{ij} + \ln v_j - \ln \lambda - \ln v_i) = \ln \lambda. \end{aligned}$$

Indeed it is *the* measure of maximal entropy – proving the uniqueness requires further work and can be found in standard references. The uniqueness, which fails in higher dimensional subshifts is related to the non-existence of phase transitions in one dimensional systems with finite range interaction. We will return to this in the next section where we show the existence of invariant measures in greater generality.

One might however point right away an immediate of generalization. Introducing weight a matrix W, $w_{ij} > 0$ we can define $[A_w]_{ij} = a_{ij} w_{ij}$, $\forall i, j$. A_w clearly inherits irreducibility, aperiodicity etc. from A. Hence imitating the construction above we can using Perron-Frobenius theory derive a Markov measure μ_{P_W}. $W \equiv 1$ obviously recovers the Parry measure. Indeed it is not hard to show that any (p, P)-Markov shift can be constructed via weights from a suitable A.

Utilizing the Perron-Frobenius Theory and the formulations above one can immediately derive further results. Given an irreducible and aperiodic SOFT $(X_A, \mathcal{B}, \sigma, \mu)$ and $\psi_i \in L^2(\mu)$, two complex-valued functions on X_A one can for example define the **one-point correlation function** as

$$C_{\psi_1, \psi_2}(n) = \int_{X_A} \psi_1(\sigma^n x)\, \psi_2(x) d\mu - \int_{X_A} \psi_1 d\mu \int_{X_A} \psi_2 d\mu\,.$$

The left shift is invertible so the expected symmetry $C_{\psi_1,\psi_2}(n) = C_{\psi_1,\psi_2}(-n)$ indeed holds. If $\psi_i = \chi_{A_i}, \ A_i \in \mathcal{B}$ then

$$C_{\psi_1,\psi_2}(n) = \mu\left(\sigma^{-n} A_1 \cap A_2\right) - \mu\left(A_1\right)\mu\left(A_2\right)$$

which then asymptotically vanishes if μ is mixing. For is the case if the SOFT is irreducible and aperiodic and for this there is a simple spectral criterium (see Exercises).

If ψ is locally constant i.e. $\psi(x) = \psi(x_0)$ then one can using the Perron-Frobenius theory show the exponential vanishing of one-point correlations (for a proof see e.g. [Bal]):

Theorem 2.9.: *Consider a (p, P)-Markov shift $(X_A, \mathcal{B}, \sigma, \mu_P)$ which is mixing. Given the spectrum $\sigma(A)$ let $a = sup_{\lambda_i} \{|\lambda_i| \mid \lambda_i \in \sigma(A) \setminus \{1\}\}$. Then for any $b > a$ ther is positive K such that for all locally constant ψ_i*

$$|C_{\psi_1,\psi_2}(n)| \leq K \sup|\psi_1| \sup|\psi_2| b^n .$$

The existence of a spectral gap in the transfer operator, exponentially vanishing correlations and asymptotic indepencence are phenomena that prevail in many dynamical systems.

SOFTs have certain key ingredients in their structure which makes their mathematical analysis feasible. But of course one might think of other ways of restricting the allowed sequences. A useful point of view to this is provided by a **graph formulation**. Suppose that we are given a finite directed graph with self loops allowed. Let us assign each node a unique symbol from an alphabet S requiring that each symbol appears only once in the graph. Transversing the graph along the arrows gives us a bi-infinite sequence of symbols.

Any SOFT clearly can be represented by such graph: the absence of an arrow from node i to node j corresponds to the exclusion i.e. the situation $a_{ij} = 0$. Starting from a complete graph with n nodes ($n \times n$ matrix s.t. $A \equiv 1$) we can by this removal procedure construct any adjacency matrix A hence any SOFT. Figure below, left, illustrates this in the case of Golden mean shift of Example 0.1. (the adjacency matrix of which was illustrated in Figure 2.1. d).

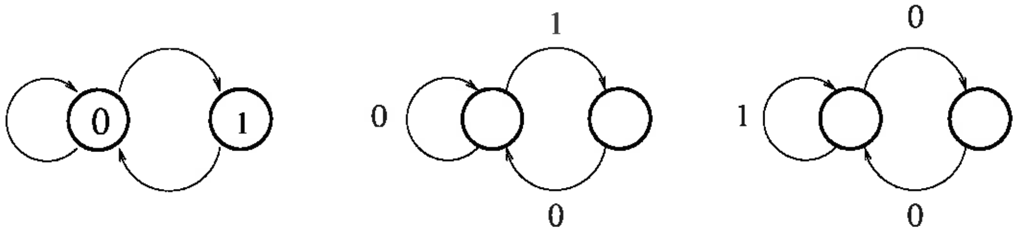

Figure 2.2. Graphs for shifts. a, b: Golden mean, c: Even.

Suppose that instead of labelling the nodes we label the arrows and moreover allow a symbol to appear multiple times in the graph. As an example of such graph see Figure 2.2. b. This is an other, non-unique, way of of representing the Golden mean subshift.

On the right there is another case where there are two nodes and the arrows connecting them are labelled by 0 and one of the nodes has a self-loop labelled by 1. Transversing this graph gives the set of subsequences of $\{0,1\}^{\mathbf{Z}}$ with the property that between any two consecutive 1's there is an even number of 0's. This is called the Even shift. As the length of the one-block does not have a bound this is not a local description and indeed this is not a SOFT (to convince of this try finding a graph description of the first type!).

The sequence space of the second type equipped with the left shift is called a **sofic shift** ("sofic"was coined by Benjamin Weiss from the Hebrew word for finite). They contain the SOFTs as a subset. Some of the SOFT theory carries over to them. For example the topological entropy can still be computed via a matrix formulation and is also given by the growth rate of number of periodc orbits. The reader might want to try proving that the even shift in fact has exactly the same entropy as the golden mean (can be done from first principles alone).

There is a hierarchy of sorts of subshifts. In some ways it resembles the **Chomsky hierarchy** of languages. Sofic shifts are actually equivalent to regular languages and **context free subshifts** correspond to context free languages. As an example of non-sofic context free shift consider $\{0,1,2\}^{\mathbf{Z}}$ with restriction that 01^i2^j0 may only occur if $i=j$. It is easy to see that in a graph corresponding to a sofic shift there is no mechanism to compare the lengths of two separate loops needed here. Hence the example is beyond the power of such description.

Another class in the hierarchy is **renewal systems** where the legal sequences are infinite concatenations from a finite dictionary of finite words. These again contains elements beyond those generated by sofic shifts.

References. Some basic introductory material is covered in Keane's article in [BKS] and more in [Pe] and [Wa]. SOFTs were introduced by Parry. Their popularity was greatly boosted by Smale in showing their usefulness in the context of a key class of smooth dynamics, the Axiom A flows. For an in depth review of transfer operators in dynamics see [Bal]. A wealth of results on Markov shifts can of course be directly translated from probability literature on Markov chains. An up to date treatise of 1-d symbolic dynamics including graph formulations, zeta-functions, digital coding applications etc. is [LM].

2.2. Thermodynamic formalism

The title is somewhat grandiose for what is coming. But there is a theory with this name and we present the basics of it. As one readily guesses this must be coming from physics, where even *theory of everything* is around the corner... But to give proper credit this important aspect of the subshift theory was initiated by mathematical physicists Dobrushin, Lanford, Ruelle and Sinai and later put into definitive form by (the mathematician) Bowen. We follow closely to the notes of the last one ([Bo]).

The physical intuition is as follows. Suppose that we have a mechanical system with some dynamics which is brought into contact with an other system of much larger size. Assume that the larger system is in thermal equilibrium (constant temperature) and the smaller will eventually reach an equilibrium with the bigger one i.e. it will be at the same temperature. The equilibrium distribution is given to a large variety of systems in statistical mechanics by a Gibbs distribution. The characteristic property of this distribution is that it minimizes the **free energy**

$$E - kTS.$$

E is typically potential or kinetic energy of the components of the system and S is the entropy usually in the form of thermal vibration. T is the absolute temperature and k is a constant. Loosely speaking free energy is the energy that could in principle be retrieved from the system. The fundamental law of thermodynamics tells that at equilibrium this quantity is minimized. Equivalently $S - E/(kT)$, the disorder, is maximized.

As we already have a measure of entropy the idea is to identify the interaction in the subshift context to provide the analog of energy and then reason what the equilibrium distribution/measure should be.

Suppose that we are given an irreducible and aperiodic SOFT on an alphabet of n symbols. We assume that on top of the "hard interactions" that the exclusions (in the form of the zeros in the adjacency matrix A) represent there is a "soft interaction" in the form of a **potential**. In its general form this is described with a function $\psi : X_A \to \mathbf{R}$. More specifically we can think ψ to contain contributions from self-interactions of a symbol (representing how much energy is involved in its presence) and pair-interactions of two symbols a given distance apart (representing spin-alignment -type interaction). So for the contribution at the origin we can write

$$\psi(x) = \psi_1(x_0) + \sum_{j \neq 0} \psi_2(j; x_0, x_j) \tag{2.2}$$

where ψ_1 and ψ_2 are the self- and pair-interaction potentials. Of course the summation should converge i.e. the distant contributions to be negligible. Also from purely physical point of view we should expect that the potential should decay in some reasonable fashion unless we model a situation where very complex interdependency between a large number of coordinates is expected.

Let us define on $C(X_A)$ a distance function as follows

$$\|\psi\|_k = \sup\{|\psi(x) - \psi(y)| \mid x_i = y_i \ \forall i \in \{-k, \ldots, k\}\}.$$

As X_A is compact elements of $C(X_A)$ are actually uniformly continuous therefore $\|\psi\|_k \to 0$.

We will restrict to the potentials in the class

$$H_A = \left\{\psi \in C(X_A) \mid \|\psi\|_k = \alpha\beta^k \ \forall k \text{ for some } 0 < \alpha, 0 < \beta < 1\right\}$$

i.e. functions with exponential tails. Turns out that these are the Hölder continuous functions on X_A (check).

The existence of an important type of equilibrium measures is established in the following key result (which is proved in e.g. [Bo]). Let $S_m\psi(x)$ denote the summation $\sum_{k=0}^{m-1} \psi(\sigma^k x)$.

Theorem 2.10.: *Given an irreducible and aperiodic SOFT and a Hölder potential ψ there is a unique σ-invariant probability measure μ on X_A, the* **Gibbs measure,** *satisfying for some constants $0 < c \leq C < \infty$*

$$c \leq \frac{\mu(\{y \in X_A \mid y_i = x_i \ \forall i \in \{0, \ldots, m-1\}\})}{e^{-Pm + S_m\psi(x)}} \leq C$$

for all $x \in X_A$ and all $m \geq 1$.

Remarks: 1. The proof of this result hinges on a highly non-trivial generalization of the the Perron-Frobenius theory to non-negative operators by Ruelle. We will not present it but refer to the standard sources i.e. work of Lanford, Ruelle and Bowen (see references in [Bo]).

2. The summation in the exponent is of course an consequence of the assumption that the potential contribution (2.2) is shift-invariant.

3. The number P appearing in the Theorem is called the **pressure**. Upto sign and scaling by temperature it turns out to be the same as free energy density. Note that it must be unique and finite as μ is to be a probability measure. P and μ of course depend on the potential ψ. However two different potentials may yield the same Gibbs measure (see exercises) and pressure.

4. Note the special case $\psi \equiv 0$ is just a plain SOFT. The Theorem implies that every m-cylinder $\{y \in X_A |\ y_i = x_i\ \forall i \in \{0, \ldots, m-1\}\}$ has upto a constant multiple the size $e^{-P(0)m}$. As these cylinders generate the Borel algebra in view of the Shannon-McMillan-Breiman Theorem we must have $P(0) = h_\mu(\sigma)$ if the Gibbs measure is ergodic. It turns out Gibbs measures are measures of maximal entropy so actually $P(0) = h_{top}$. Moreover as a fall-out of their construction the mixing property is immediate (and we know that mixing implies ergodicity).

This observation is actually not as limited as one might think as it is quickly extended. Consider the exponent in the denominator when $-m$ has been pulled out. It is $P(\psi) - (1/m)\sum_0^{m-1} \psi\left(\sigma^j x\right)$. If μ were ergodic we would by the Ergodic Theorem know that this expression gives in the limit $P(\psi) - \int \psi d\mu$ which by again quoting the SMB Theorem equates to h_μ. Turns out that this equality is actually true and our next task is to really show it!

To show the special status the Gibbs measures enjoy among σ-invariant probability measures we proceed to establish a variational principle. A couple of preliminary result are first in order. Aside from being technically useful the first one also indicates the naturalness of the Gibbs distribution.

Lemma 2.11.: *Given $\{E_i\}_1^n$ the problem*

$$\begin{aligned} &\text{MAX} \quad -\sum_1^n p_i \ln p_i + \sum_1^n p_i E_i \\ &\text{s.t.} \qquad p_i \geq 0,\ \sum_1^n p_i = 1 \end{aligned}$$

has a unique solution

$$p_i = \frac{e^{E_i}}{\sum_1^n e^{E_j}}.$$

The proof is calculus (check). In terms of our earlier thermodynamic description the numbers E_i are the different energy levels available hence the second summation can be viewed as the total mechanical energy of a system.

Given the natural generating partition $\mathcal{P}$ and its m-fold refinement $\mathcal{P}_0^{m-1}$, recall that an atom in the latter, $X_{\mathbf{a}} = \{x \in X_A | \ x_0 = a_0, \ldots, x_{m-1} = a_{m-1}\}$, is determined by its m-name $\mathbf{a} = (a_0, \ldots, a_{m-1})$. Using this shorthand define

$$\sup_{\mathbf{a}} S_m(\psi) = \sup_{x \in X_{\mathbf{a}}} \sum_{k=0}^{m-1} \psi\left(\sigma^k x\right)$$

and

$$Z_m(\psi) = \sum_{\mathbf{a}} e^{\sup_{\mathbf{a}} S_m(\psi)}.$$

Those familiar with statistical mechanics will recognize the last expression: it is a partition function. It is informationwise an extremely densely packed quantity. In particular it provides an alternate definition of the pressure. This is shown in the following result which also proves the earlier claim that this quantity is the free energy density.

Lemma 2.12.: *The expression $P(\psi) = \lim_{m \to \infty}(1/m) \ln Z_m(\psi)$ exists for all continuous potentials. If the potential is in H_A, A irreducible and aperiodic, then $P(\psi)$ equals to the pressure in the definition of the Gibbs measure.*

Proof: The existence of the limit is a simple application of subadditivity of $\ln Z_m(\psi)$ and Lemma 1.18.

For the second part recall that if μ_ψ is a Gibbs measure then

$$\frac{\mu_\psi\left(X_{\mathbf{a}}\right)}{e^{-Pm+S_m\psi(x)}} \in [c, C].$$

Summing over all allowed vectors $\mathbf{a}$ and taking into account that μ_ψ is a probability measure gives $ce^{-Pm}Z_m(\psi) \leq 1 \leq Ce^{-Pm}Z_m(\psi)$ or equivalently

$$\frac{Z_m(\psi)}{e^{Pm}} \in \left[\frac{1}{C}, \frac{1}{c}\right].$$

Taking logs and the limit shows the equivalence of the two definitions. ∎

Proposition 2.13.: *For a continuous potential ψ and a σ-invariant probability measure μ we have*

$$h_\mu + \int \psi d\mu \leq P(\psi) \ . \tag{2.3}$$

Proof: Since μ is σ-invariant we have for any $m \geq 1$

$$\frac{1}{m} \int S_m \psi d\mu = \int \psi d\mu.$$

So

$$h_\mu + \int \psi d\mu = \lim_{m\to\infty} \frac{1}{m}\left(H\left(\mathcal{P}_0^{m-1}\right) + \int S_m\psi d\mu\right)$$

where $\mathcal{P}$ is the natural generating partition. Using our new notation and recalling the definition of entropy with respect to a partition we can bound the expression inside the parenthesis from above by

$$\sum_{\mathbf{a}} \mu(X_{\mathbf{a}})\left(-\ln \mu(X_{\mathbf{a}}) + \sup_{\mathbf{a}} S_m\psi\right).$$

But by the Lemma 2.11. the maximum of this expression is attained at the Gibbs distribution. Substituting $E_{\mathbf{a}} = \sup_{\mathbf{a}} S_m\psi$ into the expression for maximum indicated in the Lemma 2.11. we get

$$\ln \sum_{\mathbf{a}} e^{\sup_{\mathbf{a}} S_m\psi} = \ln Z_m(\psi)$$

which after the limit gives the result. ∎

The measures for which we attain the equality in (2.3) are called **equilibrium states or equilibrium measures**. After this groundwork we are ready to identify them.

Theorem 2.14.: *Given a potential $\psi \in H_A$ and an irreducible and aperiodic SOFT the Gibbs measure μ_ψ satisfies*

$$h_{\mu_\psi} + \int \psi d\mu_\psi = P(\psi).$$

Proof: Given $x, y \in X_{\mathbf{a}}$ note that

$$\begin{aligned}
|S_m\psi(y) - S_m\psi(x)| &\le \sum_{k=0}^{m-1} \left|\psi\left(\sigma^k y\right) - \psi\left(\sigma^k x\right)\right| \\
&\le \|\psi\|_0 + \|\psi\|_1 + \cdots + \|\psi\|_{[m/2]} + \|\psi\|_{m-[m/2]} + \cdots + \|\psi\|_0 \\
&\le 2\alpha \sum_{k=0}^{[m/2]} \beta^k \le \frac{2\alpha}{1-\beta} = d.
\end{aligned}$$

So for any $x \in X_{\mathbf{a}}$

$$\begin{aligned}
&-\mu\left(X_{\mathbf{a}}\right)\ln \mu\left(X_{\mathbf{a}}\right) + \int_{X_{\mathbf{a}}} S_m\psi d\mu \\
&\ge -\mu\left(X_{\mathbf{a}}\right)\left(\ln \mu\left(X_{\mathbf{a}}\right) - S_m\psi(x) + d\right) \\
&\ge -\mu\left(X_{\mathbf{a}}\right)\left(\ln\left(Ce^{-Pm+S_m\psi(x)}\right) - S_m\psi(x) + d\right) \\
&\ge \mu\left(X_{\mathbf{a}}\right)\left(Pm - \ln C - d\right)
\end{aligned}$$

and therefore

$$
\begin{aligned}
&H\left(\mathcal{P}_0^{m-1}\right)+\int S_m\psi d\mu \\
&\quad=\sum_{\mathbf{a}}\left(-\mu\left(X_{\mathbf{a}}\right)\ln\mu\left(X_{\mathbf{a}}\right)+\int_{X_{\mathbf{a}}} S_m\psi d\mu\right) \\
&\quad\geq\sum_{\mathbf{a}}\mu\left(X_{\mathbf{a}}\right)(Pm-\ln C-d)=Pm-\ln C-d\,.
\end{aligned}
$$

The result follows from

$$
\begin{aligned}
h_\mu+\int\psi d\mu &=\lim_{m\to\infty}\frac{1}{m}\left(H\left(\mathcal{P}_0^{m-1}\right)+\int S_m\psi d\mu\right) \\
&\geq\lim_{m\to\infty}\frac{1}{m}(Pm-\ln C-d)=P(\psi)
\end{aligned}
$$

as the reverse inequality was established in Proposition 2.13. ∎

Remark: With some further work this could be sharpened. Turns out that the Gibbs measure is actually the unique translation invariant probability measure on X_A which is an equilibrium measure i.e. gives the equality in the variational principle. The proof can be found in [Bo].

The class H_A is just a gentlemen's agreement on what is a nice class of potentials to deal with. Note that all potentials with finite range are of course in this class. The uniqueness of the Gibbs measure can hold for potentials with tails decaying at a sub-exponential rate. The exact characterization of the potentials yielding uniqueness is an open problem.

However the more interesting question is really the complementary one i.e. for which potentials the uniqueness fails and how seriously. By this we mean questions like how many extreme points the set of equilibrium measures has and how do the correlations between the coordinates behave in generic sequences (power law versus exponential etc.).

The existence of multiple equilibria corresponds physically to phase transitions the different phases being the different equilibria. In one dimension the rule is that fast decaying potentials do not allow non-uniqueness but this is not true for two and higher dimensions. One of the first indications of this dichotomy was found in the context of the Ising model. It holds to great generality as we will see already in the context of one-dimensional cellular automata which are a step towards two-dimensional models. Part of the fascination of these models is that non-uniqueness of the equilibria is deceptively simple to generate but usually quite difficult to analyze.

References. Some basic introductory material is covered in Keane's article in [BKS] and more in [Pe] and [Wa]. For the more ambitious reader the standard pointer is [Bo]. This reference develops the theory towards application to a class of hyperbolic dynamical systems called Axiom A. Together with one-dimensional maps of the interval they are perhaps the best understood of all (chaotic) smooth dynamical systems. The definitive (and extremely dense) account on the thermodynamic formalism is [Rue]. The classic treatise on related statistical mechanics models is [Ba]. The first application of the thermodynamic thinking to smooth dynamics can be found in the papers by Sinai.

3. Cellular automata

3.1. Basics

We will now proceed to investigate a class of deterministic dynamical systems that introduce us to higher dimensional symbolic dynamics. They are defined on one dimensional subshifts of finite type of the previous section but their space-time evolutions illustrate many of the phenomena encountered in the general two and higher dimensional set-up.

Consider the space of configurations $X = S^{\mathbf{Z}}$ where $S = \{0, 1, \ldots, n-1\}$ is the the set of symbols. As before σ will be the left shift on the configurations. From mathematical point of view it is quite natural to start asking questions on properties of mappings of the configuration space i.e. $F : X \to X$. As X has the product topology described in Section 3.1. it is in particular meaningful to ask what kind of characterizations there are for F's that are continuous?

One might also view this question concerning maps F as coding problems: a code translates a sequence of symbols to another sequence. If the symbols used in both the source and the target languages are the same then this coding corresponds to some map F.

Yet another way of looking at this set-up is the physical one: suppose that the symbols lying on the integers represent the output of some physical process. If the neighboring site values interact according to a physical law then the symbol values should be up-dated according to a discrete version of this law. As a first assumption it might be useful to declare that the physical laws are everywhere the same i.e. the local update rule should be the same no matter where it is applied to the symbol sequence.

Motivated by these considerations let us now define the basic object of study.

Definition 3.1.: *Given integers $r_-, r_+ > 0$ let $f : S^{r_-+r_++1} \to S$ be a* **block map** *on $r_- + r_+ + 1$ consecutive symbols. Applied to a configuration $x \in X$ at site j the block map is a* **cellular automaton rule** *assigning to the site x_j the new value $f\left(x_{j-r_-}, \ldots, x_j, \ldots, x_{j+r_+}\right)$. Applying the rule to each contiguous r_-+r_++1-block in a configuration defines a global map F called a* **cellular automaton (CA).**

As every asymmetric block map i.e. $r_- \neq r_+$ can be extended to a symmetric one we will assume this from this on (exercise). So let $r = r_- = r_+$ and call it the **radius** of the rule. One should think of the block map updating the cell/site value at the center of the block of length $2r+1$ if r is an integer. If r is a half-integer then the midpoint is not a lattice point and one should think of the update being of the cell next to the left (say) or on the dual lattice $(\mathbf{Z}+1/2)$ on odd times. The radius is the range at which the site exerts its influence in one time unit (iterate).

A block map like the one in the Definition is in the coding theory community called a (finite) **sliding block code.**

Note that by definition the global map F commutes with the shift: $F \circ \sigma = \sigma \circ F$. So it is a translation invariant action on the bi-infinite symbol sequences. Does it have other nice properties? The following answer which is due to Curtis, Lyndon and Hedlund can and also should be viewed as an alternative, equivalent definition of a CA.

Recall from Section 2.1. that the metric of our choice on X is $d(x,x) = 0$, $d(x,y) = 2^{-\min\{|i|| \ x_i \neq y_i\}}$.

Theorem 3.2.: *A global CA map F is continuous and every continuous and shift-invariant map on X is given by a block map.*

Proof: The first part of the theorem is easy. Suppose that the rule is of radius r. Given $\epsilon > 0$ pick $n > 0$ such that $2^{-n} < \epsilon$. Two sequences x and y which are at most distance 2^{-n-r} apart agree on the interval $\{-n-r, \dots, n+r\}$ so $F(x)$ and $F(y)$ surely agree on the interval $\{-n, \dots, n\}$.

For the converse let $C_i = \{x \in X | \ x_0 = i\}$ be the usual partition of X to compact sets. Given a continuous map $F : X \to X$ by the following Lemma $F^{-1}(C_i)$ are disjoint and compact and there is $\delta > 0$ such that if $x \in F^{-1}(C_i)$ and $y \in F^{-1}(C_{i'})$ for distinct indices then x and y are at least δ apart. So choosing n such that $2^{-n} < \delta$ guarantees that any two sequences x and y agreeing on $\{-n, \dots, n\}$ are actually in the same set $F^{-1}(C_i)$. Consequently $F(x)_0 = i = F(y)_0$ i.e. the zeroth coordinate depends only on a $2n+1$-block. This together with the fact that F is shift-invariant implies that it is given by a block map. ∎

Lemma 3.3.: *If M is a compact set, $g : M \to N$ is a continuous map between metric spaces and E is compact in N then $g^{-1}(E)$ is compact in M.*
If E and F are compact and disjoint subsets of a metric space (M,d) then there exists $\delta > 0$ such that $d(x,y) \geq \delta$ for $x \in E$ and $y \in F$.

The proof is left as an exercise.

Theorem 3.2. identifies the CA exactly as the set of continuous maps of the sequence space to itself which are translation invariant. The last requirement is clearly necessary – without it the (block) map could be different from point to point.

As the result indicates a nice closure property let us denote by $E(X)$ the set of all continuous maps on the sequence space X which commute with the shift. The notation refers to the fact that this is the semigroup of endomorphisms of the full shift $(X, \mathcal{B}, \sigma)$.

Inside $E(X)$ we can distinguish finer classes. The set of surjective CA are the ones for which $F^{-1}(x)$ is non-empty for all $x \in X$. Turns out that the set of injective CA $A(X)$ are contained in the set of surjective ones. Therefore each of these CA is invertible and their set is called the automorphisms of the shift. $A(X)$ is group and indeed it is not too difficult to show that every finite group is isomorphic to one of its subgroups ([He]). Although this set still seems to be large, for the purposes of finding interesting dynamics it is a bit limited. Subsequently we will sometimes have invertibility but do not impose it.

Let us however remark that invertible or **reversible CA** are sometimes important for physical reasons. If the CA models a process at a level where information is preserved then the CA should have this property, too. Also, deleting information has an energy cost/equivalence, which at high densities might affeect the actual circuit implementation of a CA.

On a more theoretical level it is known that finding the inverse of a reversible CA is a decidable problem in 1-d, but not in higher dimensions. There are also tight bounds for the radius of the inverse block map in 1-d (see e.g. [Kar]). For some CA the inverse is immediately available as we will later see in Example 4.8.

Note that the set of surjective CA excludes any CA which has a nontrivial **attractor**. Such attractor is a strict subset of the space towards which the orbits tend asymptotically. So the asymptotic action in a system which has an attractor is confined to a subset of the space X. Indeed one can prove that if a CA is not surjective then asymptotically its action is restricted to a set (attractor) of Bernoulli measure zero. The sense in which we should expect the convergence to take place is different from that in the context of e.g. differential equations. We will return to this in detail in the subsequent sections.

3.2. Phenomenology

After the formal prelude let us take a step back and look at a few CA with the purpose of trying to see the spectrum of qualitative behavior possible. It should be emphasized that we are not proposing a classification scheme – there is no such thing for CA in the rigorous sense as of now. The purpose here is to provide some insight and screen for CA that the concepts developed so far might be applicable.

Example 3.4.: Figure 3.1. shows a few evolutions on a toral universe of perimeter 100 cells i.e. the left end wraps around to the right one. The time runs downwards, in these samples for 100 iterates. The initial state in each run a disordered configuration where zeros and ones follow independently and with equal probability.

The rules shown here are the so called **elementary cellular automata (ECA)** which means that the alphabet is $\{0, 1\}$ (corresponding to white and black) and the radius is 1. There are $2^{2^3} = 256$ such rules (as a rule rule is uniquely defined after telling to which symbol any of the eight possible binary triples maps). These CA go by the **rule number** defined as follows. Consider the set of binary triples that map to one under the rule. Each of these triples is the binary representation of an integer between 0 and 7. Let these integers be $\{a_i\}$. The rule number is then simply $\sum 2^{a_i}$ (convince yourself that two different rules cannot have the same number).

The top left evolution in Figure 3.1. is that of ECA Rule 40. $40 = 2^5 + 2^3$ and as $5_{10} = 101_2$ and $3_{10} = 011_2$ the rule is in terms of the **rule table**: $\{\{101,\ 011\} \mapsto 1\}$ whereas the other binary triples map to zero. The rule shows quite boring transient behavior: it is easy show that the evolution for this rule from a generic disordered state dies off quickly. Fading into the homogeneous state (here all zeros but could be else) is perhaps the simplest class of CA behavior. But note that already in here there is an exception: from the initial state $\underline{01}$ (periodic extension of the block 01) results an evolution of temporal period 2, a checkerboard pattern. Are there other configurations that do not die off?

Top right CA should look familiar as it is the CA implementation of the left-shift. A second of thought reveals that for a CA to be the left-shift we must have $001, 011, 101$ and $111 \mapsto 1$ while the other triples map to zero. Summing these up tells us that the rule is 170. Similarly there is among the elementary CA a right-shift, an identity and a zero-map as well as some other pretty trivial rules.

The ECA Rule 73 is illustrated on the bottom left in Figure 3.1. The rule is a canonical representative of a another class of CA namely those whose evolution results in a temporally periodic pattern. Note in the context of CA one has to distinguish between the spatial and temporal periodicity. A configuration $x \in X$ is spatially periodic if $\sigma^p x = x$ for some positive p and temporally periodic if $F^q x = x$ for some positive q. The picture indicates that several different spatial and temporal periods are allowed by Rule 73.

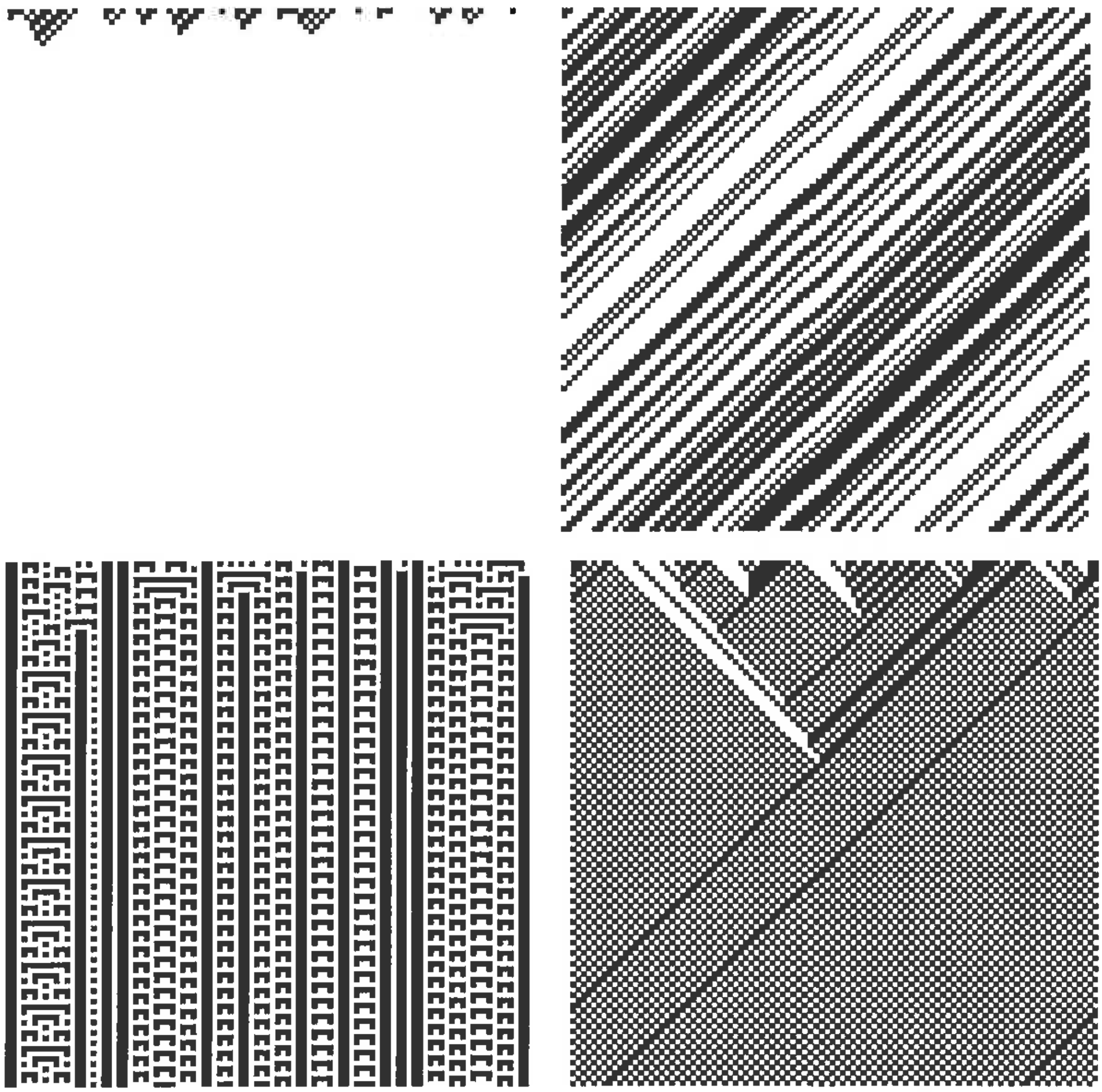

Figure 3.1. Evolutions under ECA Rules 40, 170, 73 and 184.

The last rule in Figure 3.1. is number 184 ($\{\{011, 100, 101, 111\} \mapsto 1\}$). The notable feature here is the appearance of particle-like structures. They move either to the left or right with maximal velocity ± 1 cell per iterate. Particles with the same color cannot meet and for the black and white collisions there is a simple rule: the bigger, that is wider, particle wins. It is left as an exercise to the reader to find out the exact mechanics of this collision rule. Aside the ballistic annihilation above, other interpretations are also possible: this rule has attracted quite a bit attention within non-equilibrium statistical mechanics as a minimal deterministic model for various phenomena from crystal growth to traffic jams.

Already from this Example we see a phenomenon that is of critical importance. The rule 40 looks transient in the sense that a generic initial state seems to die quickly off. But as we saw it quite isn't hence the classification of rules requires care. What does transient for most initial states really mean? In simulations CA seem to fit a handful of distinct behavioral categories but to really classify them is a different matter. None of the proposed broad classification theories has a firm foundation and it is not clear that there ever will be such theory. This is a consequence of the model class at hand being too big i.e. containing elements that lead to undecidability problems. This is useful to keep in mind as we consider some other rule and also try to find subclasses of CA that might yield to rigorious methods.

Example 3.5.: The top left evolution in Figure 3.2. is that of a radius $1/2$ binary rule. There are 16 such rules and they can be numbered the same way as the ECA (actually these are even more elementary, call them Fundamental CA). The rule in question simply sends 01 and 10 to 1 and the other two binary doubles to 0. We denote it by the code FCA Rule 6. The evolution is drawn to alternate between the lattice $\mathbf{Z}$ and its dual lattice $\mathbf{Z}+1/2$.

This rule is a close relative of the ECA Rule 90 of Example 0.3. As we will analyze these later we just note here that these are as "chaotic" CA as they come. The exact degree of mixing will be decided later.

The evolution to the right at top of Figure 3.2. is that of ECA 22 ($\{\{001, 010, 100\} \mapsto 1\}$). It is a strange rule of which very little is known rigorously. For some reason it seems to support long range order. The initial state here is engineered. Off-center is generated from blocks $S_4 = \{0000, 0010\}$ with equal probabilities and at the center there is the block 11. A careful inspection reveals that this initial state generates two distinct **topological defects** in the following configurations. These are dislocations in the ordered phase generated by the block alphabet S_4. The defects perform unbiased random walks (can be proved). When the two defects meet near the center of the figure they give rise to a "complex" phase with distinctly different density of ones in it (about .35 instead of the $1/4$ on the outside).

The bottom left evolution is that of a radius $1/2$ CA with six states. The states are in three groups, each containing two elements. For clarity the groups/phases are coded to white, gray and black. It can be proved that the phase boundaries or particles perform random walks. Both annihilations and coalescings are visible. Another variant of the theme is on the right: when also branchings of phase boundaries are allowed and they occur with suitable intensity avalanche-like particle creation can result. The initial state on the left is uniform Bernoulli on the symbols, on the right uniformly Bernoulli on subalphabets, which are set in two contiguous blocks.

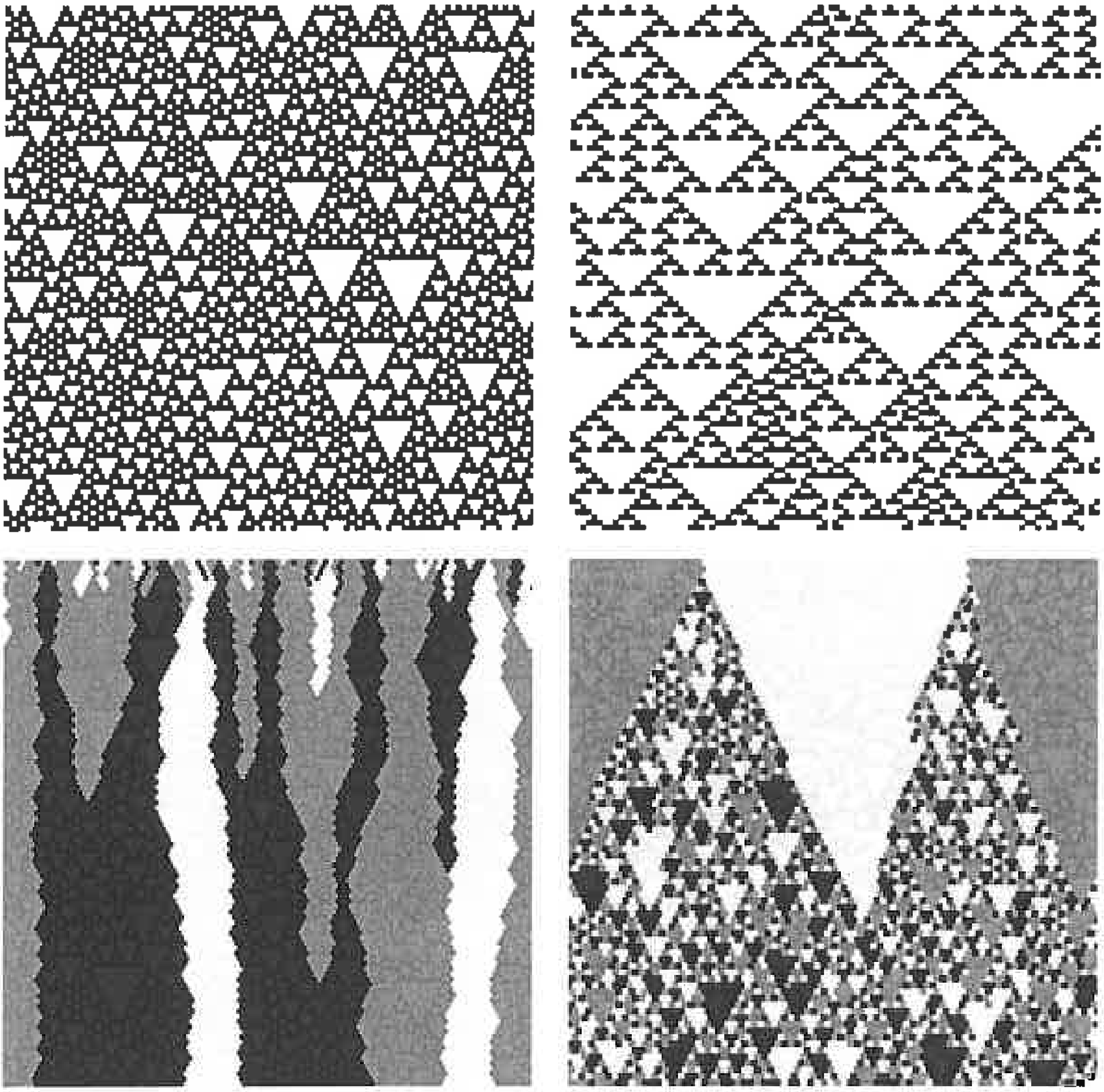

Figure 3.2. Evolutions under FCA 6, ECA 22 and two three-species rules.

The samples in Figure 3.2. are also to illustrate another aspect of CA: rules, when their principles are properly understood, can be synthesized to perform various (physics-like) phenomena. As completely discrete systems CA are extremely efficient to run on computer (especially parallel such) and thereby provide a way of simulating other systems.

The rules in Example 3.5. seem to have a chaotic component in them. The characteristic property of a chaotic CA is that from most initial states the evolution tends to a disordered state. As such states are naturally described in terms of measures these CA are the likely candidates for some of the ergodic theoretic principles to apply. Before proceeding to that we take a look at yet another, qualitatively different class.

Example 3.6.: Instead of drifting towards a homogeneous disorder a few CA seem to generate from almost all initial state localized structures travelling in time like particles. Furthermore these particles can come in great variety and be stable objects to the extent that their interaction properties with other similar structures can be laid out.

A case in point is the ECA Rule 110 in Figure 3.3. showing a menagerie of particles. Here $001, 010, 011, 101$ and 110 map to 1. The initial state is again disordered but note that it is now wider that before, 400 cells. This is necessary to see the key action, the particle formation and interaction. The length-scale of the interaction is distinctly larger than in our other examples. The rule seems to have an attractor which is the ordered background phase (spatial period 14, temporal period 7) in which the particles are cracks. Note that several of the particles are capable to soliton-dynamics: they pass each other without changing their type or direction but experiencing a slight delay. This rule is known to support universal computation: by arranging the initial state the particles can be made to perform any Boolean logic operation and thereby implement any recursive function.

Suppose that we start the CA action on the full shift $X = S^{\mathbf{Z}}$ as in most examples above. If the CA is not onto then the set $X^i = F^i(X)$ is a shift invariant closed subset of X. In some cases it should have a fairly simple description like in the context of the ECA Rule 40 but it is not in general a subshift of finite type. But for all finite i it is a sofic system i.e. a regular language. So any finite time set has a description in terms of a finite graph and one can further characterize the complexity of the CA evolution by measuring the size of this graph. The infinite limit of X^i can be an extremely complicated object. For some (exclude onto rules for non-triviality) rules it is known to still be a regular language, but in general it can be any language in the Chomsky hierarchy, i.e. also context-free, context-sensitive or unrestricted. To get an idea what it might be a given automaton one could e.g. look at the exponential growth rate of the number of nodes in the minimal describing graph (a topological entropy like measure), but there are striking counterexamples to this, too (i.e. a fast growing graph sequence has a trivial limit).

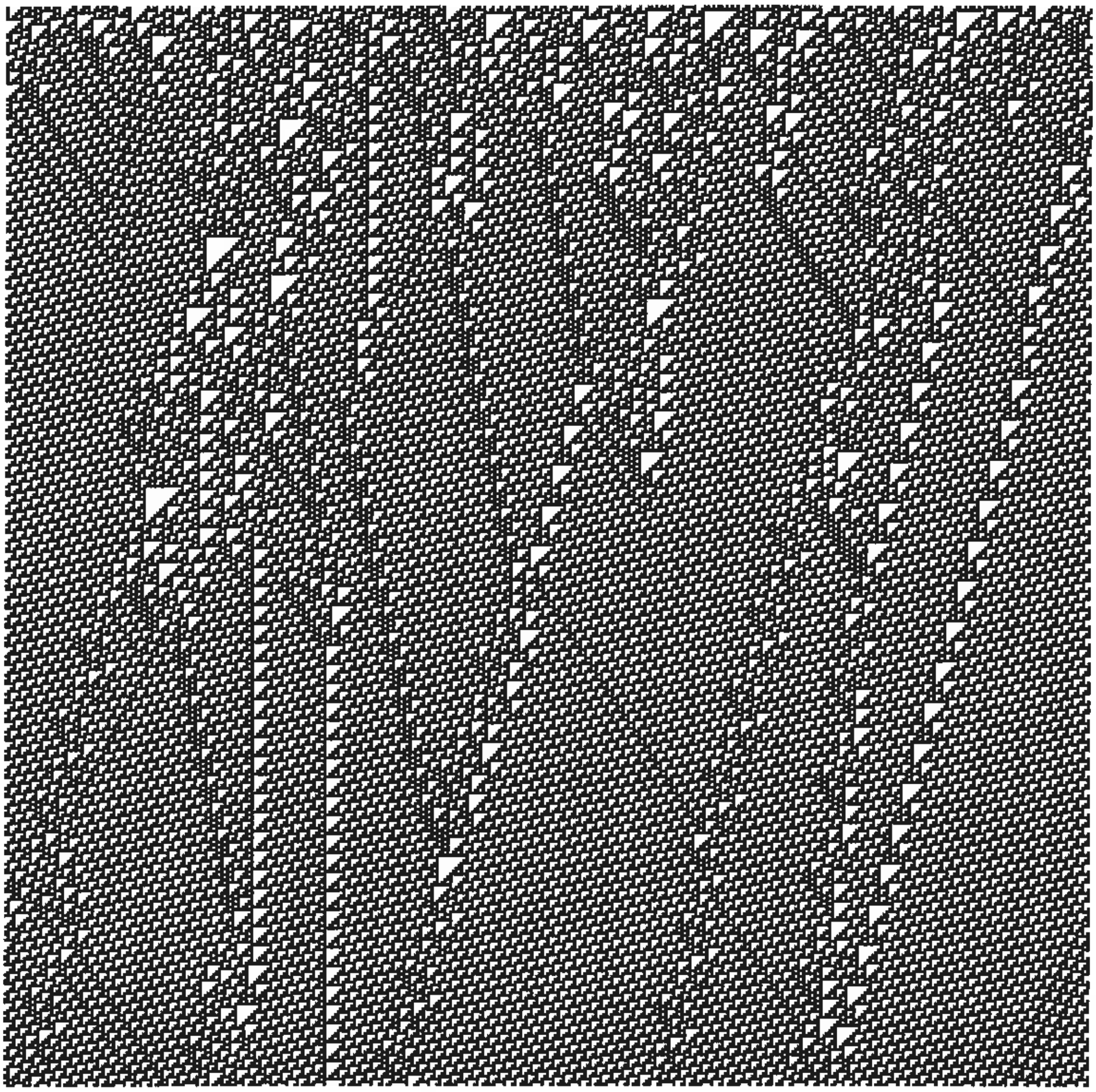

Figure 3.3. ECA Rule 110 evolution from Bernoulli(1/2) initial state.

The problem of classifying CA is related to this phenomenon. If the classification is based on the long-term behavior then one has to be able to sort out the structure of the set X^{∞} for the set of CA under consideration. But if the CA set is not carefully chosen it may contain elements whose asymptotic properties are undecidable whence a statement cannot be made on the behavior of the CA in that set.

References. Cellular automata were invented by the mathematicians John von Neuman and Stanislav Ulam in the early fifties. They didn't get very far with the theory as there were no suitable computers yet around to simulate the rules to guide their intuition. Von Neuman was among many other things interested in defining life by instilling its essential ingredients lying underneath the wrappings of chemistry and biology. To this end he was able to construct a (very complicated) self-reproducing CA. A long latent period followed until John Horton Conway in the early seventies published the rules for the Game of Life (see Example 1.4.). It caused an explosion in activity, mostly recreational, but also a few notable scientific results the prime being perhaps the implementation of universal computer in the GoL. This stimulus was likely to be still felt when the current wave of scientific assault started in early eighties. Among the key figures in the beginning were Margolus, Toffoli and Wolfram who together with many others mostly in the physics community systematically compiled a bulk of data on CA as well as build hardware and software to simulate a large variety of rules. An effort was made to fit the data to existing statistical mechanics theory but the subject provided problems that stretched the classical notions. A bit skew but still comprehensive summary of the non-rigorous work by mid-eighties is [Wo]. Theoretical results stated emerging after that time and these days there is a steady stream of even rigorous results surfacing. Notable recent results include e.g. the universal computation in the Rule 110 proved by Cook. The subject matter is acknowledged to be rather deep, its analysis requiring new notions to be formulated and seems that many CA rules are still beyond mathematics.

There are a number of ways modifying the set-up from the deterministic CA we are considering. If the up-date rules have a random component in them the systems are probabilistic CA. These in turn are close to interacting particle systems studied in probability theory and motivated by the Ising model and other such statistical mechanical models. Note that is these models the up-dates are asynchronous, a significant demarcation from the standard CA An altogether different approach to the subject are the lattice gas CA In them a lattice is fixed and so is usually the number of particles living on the lattice (or their density if the lattice is infinite). The particles move from one lattice site to a neighboring one according to where their discrete valued momentum vector points. The CA is defined by giving the collision rules i.e. what is going to happen if two or more particles are entering the same site simultaneously. These CA have turned out to be quite useful in simulating e.g. fluid flow and also some rigorous results exist for them. For more details see [TM].

3.3. Permutivity

In this section we single out one subset of CA that have properties that make them amenable for mathematical analysis. The modest aim in here is to present the first statements together with their proofs and to motivate and state a few more involved results.

We start with the a definition which should be thought as describing the information transmission properties of a rule.

Definition 3.7.: *The block map $f(x_{-r},\ldots,x_r) = x_0'$ is* **right permutive** *if for any fixed block of length $2r$, $(x_{-r},\ldots,x_{r-1})$, the map from x_r to x_0' is a permutation.* **Left permutivity** *is defined in a symmetric fashion. If a block map is both left and right permutive it is called* **permutive**.

The left-shift CA of Example 3.4. is right permutive but not left permutive and identity block map is neither. A symmetric CA rule that is left permutive is of course right permutive. The usefulness of this concept is indicated by a simple consequence:

Proposition 3.8.: *A left/right permutive CA is onto.*

Proof: Suppose that the rule is right permutive. Given $y \in X$ and a block $(x_{-r},\ldots,x_{r-1})$ by the right permutivity we can choose x_r so that $f(x_{-r},\ldots,x_r) = y_0$. Proceed inductively to find out all $x_i,\ i > r$ from $y_j,\ j > 0$. But there is nothing special about starting at the origin i.e. we can start the induction from arbitrarily far to the left. Hence any y has a preimage. ∎

Note that not every surjective CA is permutive. The trivial counterexample is the identity CA Can you think of others?

As we are gearing towards ergodic theory a useful consequence of the permutivity (surjectivity) is the measure preservation. Recall that the uniform Bernoulli measure on the sequences from an alphabet S, μ_S, assigns the probability $1/|S|$ to each symbol independently.

Proposition 3.9.: *A left/right permutive CA on the alphabet S preserves the uniform Bernoulli measure μ_S on the configurations.*

Proof: Again suppose right permutivity. Given the $(x_{-r}, \ldots, x_{r-1})$ block we see that if the entry x_r is uniformly Bernoulli distributed then by the permutivity so is the image of the block, y_0. y_1 is similarly determined by the next (overlapping) block of length $2r$ on the right and the entry x_{r+1}. Moreover y_0 and y_1 are independent because x_r and x_{r+1} are. Inductively one extends this to all of the coordinates. By the construction any k-tuple of y_i's is independent so the measure on the image is again uniform Bernoulli. ∎

Given an infinite **CA evolution** $x^{(2)} = \{F^i(x)\}_{i=0}^{\infty}$ from the initial state $x \in X$ we can define a second shift, the **temporal shift**, σ_t, simply as $\sigma_t x = F(x)$ (or on $x^{(2)}$ as $\left(\sigma_t x^{(2)}\right)_i = \left(x^{(2)}\right)_{i+1}$. As the measure μ_S is preserved under the global (one-sided permutive) CA map F there is an induced measure on the evolutions of the CA This measure, call it $\mu^{(2)}$, is by the Proposition invariant under the time-like shift σ_t. This property is called **stationarity** or time-independence. This is a fundamental property of any physical system *in equilibrium* and it could be vaguely interpreted as no time-instance having a particular position in history (to make this really true extend the evolution to be also timewise bi-infinite i.e. starting from $-\infty$). It is essential for any ergodic properties to hold for a system. On this background the Proposition could be interpreted as how stationarity enters to CA evolutions.

3.3.1. A case study: Rule 90

In this section we will analyze the mixing properties of the simplest permutive ECA It was introduced in Example 0.3. and goes by the code number 90. Turns out that this CA is actually as a factor hidden in a number of other cellular automata and provides an explanation to their chaotic-like behavior. Moreover the here analysis can be readily extended to a number of other permutive CA

Recall that Rule 90 maps $001, 011, 100$ and 110 to 1 whereas the other two binary triples map to 0. The first observation is that the rule table i.e. update to x_0 from time i to time $i+1$ can be compactly described as

$$x_0^{i+1} \leftarrow x_{-1}^i + x_1^i \pmod 2. \tag{4.1}$$

From this it is plain that the rule is symmetric and permutive and hence onto (but the global CA map F is 4-1). Moreover it has a very strong property formulated as follows.

Definition 3.10.: *A CA F has* **additive superposition property** *if for any two configurations x and x' in $S^{\mathbf{Z}}$ it holds that*

$$F(x+x') = F(x) + F(x') \pmod{|S|}.$$

To appreciate this property we should view the space of configurations $X = S^{\mathbf{Z}}$ as a compact group the operation being coordinatewise addition modulo $|S|$. So if $x, x' \in X$ then $(x+x')_j = x_j + x'_j \pmod{|S|}$ $\forall j$. The property above tells us that the map F is compatible with the underlying group structure; it commutes with the group operation. We say that F is a **group homomorphism**.

Rule 90 clearly has the additive superposition property. It enables us to decompose its evolutions to the simplest building blocks. To see how this works let us consider the rule acting on a configuration with a single 1 at origin and the rest of the coordinates 0's. The resulting evolution is Pascal's triangle modulo 2 illustrated in Figure 3.4. (64 steps).

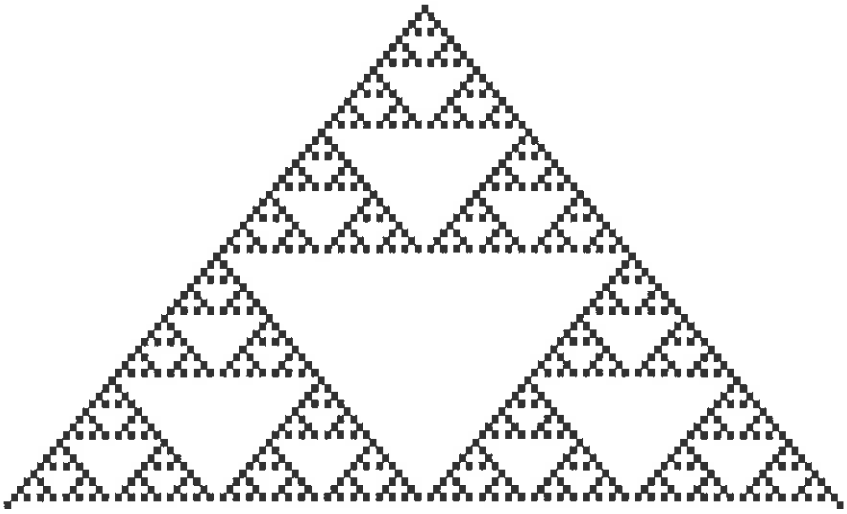

Figure 3.4. Pascal's triangle modulo 2.

Given a general initial configuration x one should think of a copy of the infinite Pascal's triangle to be attached at every site j where $x_j = 1$. The state at any point (j, i) in the evolution $\{F^i(x)\}_0^\infty$ is then obtained by superimposing modulo 2 the values in the triangles. A particular consequence of this is that changing the value of a single site in the initial configuration flips the values in the Pascal's triangle attached at the site. Hence we see an extreme sensitivity to initial condition, a characteristic feature of chaotic evolutions.

To make a more exact statement on mixing properties we need some theoretical tools. For more details of these see [Kat] or [HR].

Definition 3.11.: *The* **character** *of a compact group H is a complex-valued continuous homomorphism $g : H \to H$ with unit modulus i.e.*

$$|g(h)| = 1 \qquad g(h+h') = g(h)g(h').$$

Characters form a group called the **dual group** *of H denoted by $\hat{H}$. $\hat{H}$ is a commutative group pointwise multiplication being the operation.*

The character group of $X = \{0,1\}^{\mathbf{Z}}$ is a countable group G whose elements are sequences of the form $g = \{g_j\}_{j=-\infty}^{\infty}$, $g_j = 0$ except for a finite number of j's for which $g_j = 1$ (exercise). They evaluate to ± 1 by

$$g(x) = \prod_{-\infty}^{\infty}(1-2x_j)^{g_j} = (-1)^{\sum_{-\infty}^{\infty} g_j x_j} \quad .$$

On the character group we have an induced homomorphism. Given a character g we say that the **dual homomorphism** of F is $\hat{F}$: $\hat{F}(g) = g \circ F$ (the definition is as that of any adjoint operation). So $\hat{F}$ is again a map from $\{0,1\}^{\mathbf{Z}}$ to itself and as it quickly turns out that $\hat{F}$ is given by the same rule (4.1) so we drop the hat.

From the observation that Pascal's triangles are infinite and cannot completely cancel each other if rooted to different coordinates we see that for any non-trivial g the orbit $\{F^i g\}$ is infinite. Furthermore by Proposition 3.9 the map F preserves the uniform Bernoulli measure μ on G. Then by the result in Appendix 2. the action of the map F is ergodic with respect to μ. Furthermore as the action with respect to the spatial shift is independent one can show that the joint action $(j,i) \mapsto \sigma^j F^i$ is ergodic.

To make use of the result consider the frequency of a pattern in a generic evolution from μ. To formulate this let E be a finite subset of $\mathbf{Z} \times \mathbf{N_0}$ and $P : E \to \{0,1\}$ the pattern of zeros and ones on it. Let $p(x) = 1$ if $\left(\sigma^j F^i x\right)_0 = P(j,i)$ for all $(j,i) \in E$. So p indicates the pattern P. Using the the ergodicity and the fact that the actions σ and F commute one deduces an ergodic theorem for this CA:

Theorem 3.12.: *For μ almost every $x \in X$ the asymptotic frequency of the pattern P is given by*

$$\lim_{n\to\infty} \frac{1}{(2n+1)n} \sum_{j=-n}^{n} \sum_{i=0}^{n-1} p\left(\sigma^j F^i x\right) = \int_X p d\mu.$$

Remark: One should pay attention to the following detail. Just by knowing that the uniform Bernoulli measure μ is preserved under σ and that the σ-action is ergodic we can apply the ergodic theorem to any finite block of symbols e.g. compute the frequency of [010101] at time zero. Furthermore knowing that μ is preserved by F we can do the same at any fixed iterate. But to find the frequency of any genuinely two-dimensional pattern we need the joint ergodicity of the actions. In particular using the result the most visible feature of the evolution of this CA can be formulated. One can compute the frequency of the inverted triangles as a function of their size (exercise).

The Rule 90 has multiple invariant measures, most of them concentrated on periodic points (e.g. unit mass on the all-zero configuration, δ_0). Among the Bernoulli measures μ_p, where p is the density of ones, $p = 1/2$ is the only invariant one (suppose $0 < p < 1$ for non-triviality). We now indicate how the CA evolves from the non-invariant μ_p's.

The weak convergence of measures $\{\nu_n\}$ means that $\int f d\nu_n \to \int f \nu$ for all continuous functions f. From real analysis (e.g. [Ka]) we know that weak convergence is equivalent to the Fourier transforms converging pointwise to the appropriate limit: $\hat{\nu}_n(g) \to \hat{\nu}(g)\ \forall g$. The Fourier transform of a measure ν on X is in turn defined as

$$\hat{\nu}(g) = \int_X g(x) d\nu(x) \qquad g \in G.$$

(the symbol g is no accident: e.g. on the compact group [0,1] the characters are the functions $t \mapsto e^{int}$ and in general the abstract Fourier transform is equivalent to evaluating a character). Since F is a homomorphism on X we have $\widehat{F^i\mu}(g) = \hat{\mu}\left(F^i g\right)$. So to determine whether the sequence of measures $F^i\mu$ converges it suffices to consider the pointwise convergence of the functions $\hat{\mu}\left(F^i g\right)$. This fact is useful for us here since the Fourier transforms of μ_p can be computed (for the details see [Li]) and one gets

Theorem 3.13.: *For $p \neq 0, 1/2, 1$ the sequence $\left\{F^i\mu_p\right\}$ does not converge but the Cesaro average $\frac{1}{n}\sum_{i=0}^{n-1} F^i\mu_p$ converges weakly to the uniform Bernoulli measure μ.*

The ECA Rule 90 evolution from a disordered state with density of ones exceeding $1/2$ results in density fluctuations upto arbitrarily high iterates. But their spacing is so thin that the Cesaro average isn't affected.

After these technicalities let us step back and make a couple of general observations which also motivate the following section.

The Rule 90 is a bit redundant. By inspecting it one notices that every evolution splits into two completely independent parts which together form a checkerboard. If the initial state is x the entries $\{x_{2j}\}$ completely determine the sites $\{F(x)_{2j+1}\}$ which determine $\left\{F^2(x)_{2j}\right\}$ and so on. Same with odd initial coordinates. The action of the Rule 90 on the components is upto spatial scaling by factor

2 identical to the radius 1/2 FCA Rule 6 introduced in Example 3.5. Indeed the analysis above could be done for this rule with only slight extra bookkeeping to account the appearance of the dual lattice.

There is actually a rather general result in the lattice action set-up which tells that an ergodic action is immediately mixing, too. So in particular the joint action of σ and FCA 6 is mixing. This result by the previous observation carries over to Rule 90.

Finally it should be noted that the machinery used in this section is available for additive rules on any alphabet with $|S|$ being prime. Non-primes cause some technical problems but similar results can be proved for them with other methods.

References. Permutivity was introduced by Hedlund in his seminal study [He]. The measure theoretic ramifications discussed were first formulated by Coven and Paul [CP]. The analysis of Rule 90 follows the ideas of Lind [Li].

3.4. Subpermutivity

The definition of permutivity cannot be effectively extended to other entries than the extreme left and right ones of the block map. By this we mean that the key surjectivity result (Proposition 3.8.) is obtained only when a distal entry permutes. This motivates us to redefine the CA in a way that makes permutivity a more natural concept and moreover immediately suggests ways of generalizing it.

Definition 3.14.: *Given the set of $4r$-tuples of symbols in S a block map f of radius r defines a mapping from it to $2r$-tuples. On the new alphabet $T = \{0, 1, \ldots, |S|^{2r} - 1\}$ we thus obtain an* **induced cellular automaton map** *$\tilde{f}$ on two-blocks: $\tilde{f} : T \times T \to T$.*

Examples of two-block maps induced by rules of radii 1 and 3/2 are shown in Figure 3.5. Note that the blocking introduced here does not need to preserve the symmetry of the rule unless a permutation in the $2r$-blocks is also accounted. In the subsequent analysis we primarily deal with the two-block representation and it secondary whether it was obtained via a tiling or not. Therefore the radius of the induced two-block rule is usually taken to be 1/2 and it is then understood that the global CA map alternates between configurations on integers and half-integers.

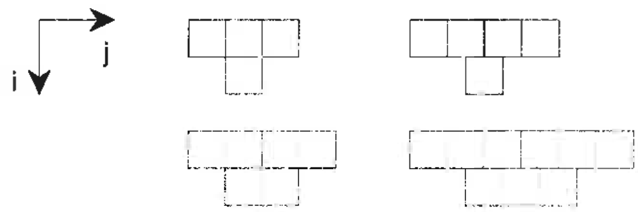

Figure 3.5.

A graphic consequence of this definition is that the space-time evolution of the rule is converted into a **tiling**. The tiles are the $(2r \times 1)$-rectangles that are piled in the fashion indicated in the Figure.

Proposition 3.15.: *If the original block map was (right/left-) permutive (in the sense of Hedlund) so is the induced two-block map and conversely.*

Proof: Pick $t^{(i)} = (t_1^{(i)}, \dots, t_{2r}^{(i)}) \in T$, $i = 1, 2, 3$, let $t^{(2)} \neq t^{(3)}$ and suppose that they agree upto the k^{th}, but disagree at the $(k+1)^{\text{st}}$ coordinate, some $k < 2r$. If the $(2r+1)$-block map is right-permutive then f applied to $t^{(1)}t^{(2)}$ and $t^{(1)}t^{(3)}$ cannot match beyond the k^{th} entry. So $\tilde{f}(t^{(1)}, \cdot)$ is injective (and surjective).

Conversely if f in not right-permutive there is a $2r$-vector and distinct $s_1, s_2 \in S$ such that $f(t_1, \dots, t_{2r}, s_i)$ agree. But then augment from the left the vectors $(t_1, \dots, t_{2r}, s_i)$ to two $4r$-vectors. $\tilde{f}$ on these agree so its permutivity fails. ∎

The two-block map allows a simple algebraic formulation of CA since we are given a finite set of symbols T closed under a binary operation $\tilde{f}$ between any two of them. If in the equation $\tilde{f}\,(t_1, t_2) = t_3$ between symbols from T any two determine uniquely the third one the system $(T, \tilde{f})$ is called a **quasigroup**. If it moreover has a identity element it is called a **loop**. Hence an alphabet with a permutive two-block CA map is a quasigroup or a loop. Note that these are not necessarily groups since associativity is not assumed and this would indeed be unusual for a CA action. They are nevertheless rare since they are the latin squares on the given alphabet. For a strictly subpermutive CA the entire alphabet together with the two-block rule is not a quasigroup but it has subsets which are (or are even loops). Note also that the identity, if it exists, is unique and forms a subalphabet by itself but it does not need to be contained in every permutive subalphabet. If a rule has a **quiescent state** i.e. a symbol $s \in S$ such that $f(s, s, \dots, s) = s$ then of course the two-block rule fixes the appropriate symbol $t = (s, \dots, s)$ and $\tilde{f}\,(t, t) = t$ holds. The symbol t is a natural candidate for the identity.

The **Cayley table** will provide a compact representation of a rule. As an example we have included in Figure 4.6. the table of the Klein four group and a sample of the CA evolution corresponding to it. On the right we have the multiplication table of the elementary cellular automaton Rule 18.

	0	1	2	3
0	0	1	2	3
1	1	0	3	2
2	2	3	0	1
3	3	2	1	0

0 1 1 2 2 0 3 3 3 1 0 2 3 2
1 0 3 0 2 3 0 0 2 1 2 1 1
1 3 3 2 1 3 0 2 3 3 3 0
2 0 1 3 2 3 2 1 0 0 3

	0	1	2	3
0	0	1	2	2
1	1	0	0	0
2	2	3	0	0
3	1	0	0	0

Figure 3.6.

The Klein four group is actually the Cayley table of one of the four permutive ECA, the Rule 90. By inspecting the Cayley tables of the others one immediately notices that the rules 90 and 165 are groups and 105 and 150 are asymmetric quasigroups.

Moreover the first two are a conjugate pair i.e. $(p \circ f_{90})(s, s') \equiv f_{165}(p(s), p(s'))$ for some permutation p on the four symbols and so are the last two. But by the stated asymmetry e.g. 105 cannot be conjugate to either 90 or 165.

Permutive cellular automata even in their one sided form are quite rare. For example out of the 16 rules on binary doubles two are permutive and of the 256 elementary cellular automata there are 4 permutive and 24 left- or right-permutive ones. However permutivity prevails in a partial form in many rules and can still dominate the behavior of the rule as we will see. We now proceed to formulate partial permutivity and its measures for two-block rules.

Let f be a two-block map on an alphabet T. Suppose that T_l and T_r are subsets of T such that $f(l, \cdot)$ and $f(\cdot, r)$ are permutations for each $l \in T_l$ and $r \in T_r$ respectively. If $p_l = |T_l|/|T|$ and $p_r = |T_r|/|T|$ then the rule f is called (p_l, p_r)-**permutive.** If $p_l = p_r = p$ it is simply called p-permutive. This definition however does not have useful closure properties and we therefore proceed to refine. By a maximal subalphabet with respect to P we mean a subset of the full alphabet with a property P such that if this subalphabet is augmented with any element from its complement it loses the property.

Definition 3.16.: *A subset T_r is a right-invariant subalphabet of T if $f(r, T_r) = T_r$, $\forall r \in T_r$ i.e. $f(r, \cdot)$ is right permuting on T_r for each $r \in T_r$. Left-invariant subalphabets are defined in a symmetric way. If T_l and T_r are maximal such subalphabets and $p_l = |T_l|/|T|$ and $p_r = |T_r|/|T|$ the rule is called (p_l, p_r)-***subpermutive***. If these maximal subalphabets coincide then $p_l = p_r = p$ and the rule is called p-***subpermutive***.

Remarks: 1. A left-invariant subalphabet does not need to be right-invariant and even if T_l and T_r are unique they are in general different. For a symmetric rule they of course agree.
2. Only rules that are 1-permutive can be additive. The strictly subpermutive rules have a nonlinear character when not restricted to one of the invariant subalphabets. In particular this implies that the Fourier analytic techniques of the previous Section will not be of much help and one has to argue with different methods.

Given a subalphabet $T' \subset T$ the subset $X' = (T')^{\mathbf{Z}} \subset X = T^{\mathbf{Z}}$ consisting of sequences from this subalphabet is said to be **generated** by T' or tiled by T'-blocks if these are induced from a wider block map. By the surjectivity if the subalphabet T' is either left or right invariant then the image of X' under the CA is also generated by T'.

Example 3.17.: The ECA Rule 18 is defined by requiring that 001 and $100 \mapsto 1$ while the other binary triples map to 1. It induces a two-block map on symbols $\{0, 1, 2, 3\}$ which has the permutive subalphabets are $T_1 = \{0, 1\}$ and $T_2 = \{0, 2\}$. This can be readily seen from the Cayley table on the right in Figure 3.6. The action of the induced CA map on the subalphabets is that of FCA 6 on binary doubles (see Example 3.5.).

Example 3.18.: The ECA Rule 22 is defined by requiring that $001, 010$ and $100 \mapsto 1$ (call the block map τ_{22}). The induced two-block map only has one invariant symbol (00) which generates the quiescent state. The square of the rule, τ_{22}^2, is of radius two and the two-block map it induces on the alphabet $\{0, 1, \cdots, 15\}$ is 1/8-permutive. The six maximal (left- and right-) invariant subalphabets are $\{0, 1\}$ and $\{0, 7\}$ i.e. tile sets $\{0000, 0001\}$ and $\{0000, 0111\}$ together with their translates. The action on the subalphabets is again radius 1/2 Rule 6/16. Higher powers of τ_{22} yield families of larger tiles which seem however to be just piles of these.

3.5. Interaction of subalphabets

The interaction of two subalphabets/tilings can lead to different phenomena. We present here only the rudiments of the theory: the basic definition and a few special cases indicating the possibilities.

Definition 3.19.: *Given two subalphabets S_1 and S_2 let $A = S_1 \cap S_2$ be the set of* **ambiguous symbols**. *If it is nonempty it is by itself an invariant subalphabet. Ambiguous symbols are* **receding** *i.e. $f(s, a) \in S_i \setminus A$ for all $s \in S_i \setminus A$, $a \in A$ and $i = 1, 2$. The configurations in the set*

$$\underline{S_1 S_2}(j) = \{\{s_k\} \mid s_k \in S_1 \ \forall k \leq j, \ s_k \in S_2 \ \forall k > j \text{ and } s_j, s_{j+1} \notin A\}$$

are said to have a **boundary point** *at $j + 1/2$.*

The set of ambiguous symbols is usually empty or consists of one element which is generated from the quiescent state. Note that if the subalphabets are disjoint ($A = \emptyset$) then every configuration in $\underline{S_1 S_2}$ (i.e. left half generated from S_1 and right from S_2) is in $\underline{S_1 S_2}(j)$ for a unique j. If $A \neq \emptyset$ then any configuration of the form $\underline{S_1 A S_2}$ where $\underline{A}$ is a finite block of symbols from A is eventually reduced to the form $\underline{S_1 S_2}$ (the $\underline{A}$-block is shortened by one at each iterate of the automata hence the number of iterates needed to this reduction is $|\underline{A}|$). Therefore the definition above applies again and we define the location of the boundary point in between these instances by interpolating.

In physics terminology one might call the configurations generated from an invariant subalphabet as phases or ground states and the just identified boundary point as the phase boundary/topological defect/kink/dislocation (all these names and others are actually used in physics literature...).

Example 3.20.: In the Example 3.17. we identified the subalphabets $T_1 = \{0,1\}$ and $T_2 = \{0,2\}$. The ambiguous symbol is $T_1 \cap T_2 = 0$ which generates the quiescent state. In Example 3.18. the square of the rule has the ambiguous symbol 0000 again generated by the quiescent state. However it is possible for a rule to have an ambiguous symbol which is not generated by the quiescent state (exercise).

The basic case is the one where the collection of sets $\underline{S_1 S_2}(j)$ is closed under a CA map. In this **inert** case the boundary will prevail at all times and no new ones are created.

Example 3.21.: In the Example 3.17. the boundary between the subalphabets defines a defect and the action of the FCA Rule 6 (induced by the ECA Rule 18) is inert on $\underline{T_1 T_2}$. Ordered in the reverse direction the action is not inert anymore since symbol $3 = 11_2$ is produced but the multiplication table (Figure 3.6., right) is of course closed. Here the ordering is not essential in characterizing the motion of the boundary point. This case was analyzed in [EN] where it was proved that the phase boundary performs an unbiased random walk.

If in the table the shaded element is changed into 0 we have a perfectly symmetric (and closed) CA action on a subalphabet and a unbiased random walk will prevail under iteration. However if instead we have 1 or 2 in this entry then the configurations $\underline{T_2 T_1}$ show T_1 or T_2 respectively dominating i.e. winning all the interactions not involving the neutral element 0. As a result the phase boundary moves in a monotone fashion at maximal speed (1/2) either to the left or right.

The two distinctly different types of motion that a boundary point can have are those of a **signal** and a **random walk**. By a signal we mean rectilinear trace in the space-time-diagram i.e. a motion with constant speed on he background. Upon creation this motion is straight forward. Upto a possible collision with another structure it is only dependent on its the creation i.e. a localized part of the initial configuration. On the other hand random walks are motions which exhibit strong dependency to the initial configuration. If this is distributed according to the appropriate product measure the boundary motion will perform a non-deterministic, stationary motion and have positive variance etc. It can have independent increments and hence be Markovian but this is not the case in general. Signals can be viewed as degenerate forms of random walks (zero variance). See [E2] for the exact definitions and the dichotomy between these types.

When the motion of the boundary point is considered for a CA with a random initial configuration random walks prevail under fairly general circumstances if neither of the subalphabets is dominating with respect to the other one. This and some other features of the interaction can be readily read from the Cayley table. Since we do not want to get too deep into the subject and we already gave in Example of a Cayley table generating a signal we cap things off by showing a basic argument in the diffusive case. The general theory is in [E1&2].

Consider the simplest case i.e. where the alphabet S partitions into two subalphabets S_1 and S_2 each consisting of two symbols. So there are no ambiguous symbols. Suppose that the radius $1/2$ CA F acting on $X = S^{\mathbf{Z}}$ (and the dual) has a Cayley table where the off-diagonal 2×2 matrices are such that each row and column has one symbol from each subalphabet. Finally let ν be the measure on X formed by joining the uniform Bernoulli measures on $X = S_a^{\mathbf{Z}_- \cup \{0\}}$ and $X = S_{3-a}^{\mathbf{Z}_+}$, $a = 1$ or 2. The measure is supported by either of the subsets of configurations $\underline{S_a S_{3-a}}$ i.e. the initial configuration has exactly one defect, located at $1/2$.

Theorem 3.22.: *Under the iteration of F the boundary point performs an unbiased nearest neighbor random walk. It has independent and identically distributes increments $\pm 1/2$ and unit variance $1/4$ from ν-almost every initial state.*

Proof: Without loss of generality suppose that $a = 1$ and that the defect is at time i at j_i. Consider the triangle T_i with vertices at $(j_i, i+1)$ and $(j_i \pm (i+1)/2,\ 0)$. Define the backward cone of the boundary pair centered at j_i at time i to be the set $T_i \setminus \{(j_i, i+1)\}$. The past of the walk at time i is clearly contained in this backward cone and the cone determines the next jump i.e. value of the cell at $(j_i, i+1)$. Suppose that the walk jumps to the right i.e. the cell at $(j_i, i+1)$ is in S_1. We claim that given the backward cone at time i the value of the neighbor at $(j_i + 1, i+1)$ is determined permutively by the entry at $(j_i + (i+3)/2,\ 0)$. This follows by noting that as $(j_i + (i+1)/2,\ 0)$ is now fixed $(j_i + (i+3)/2,\ 0)$ permutes $(j_i + (i+2)/2,\ 1)$ and then iterating this argument i times. So the next jump is independent of all the previous ones. Moreover as the symbol at $(j_i + (i+3)/2,\ 0)$ is uniformly distributed in S_2 so is the symbol at $(j_i+1, i+1)$. By the column structure of the off-diagonal 2×2 matrices the jumps to both directions take place with probability $1/2$. Therefore the unit variance is simply $1/2(-1/2)^2 + 1/2(1/2)^2 = 1/4$. ∎

Remark: The case here is the cleanest and in particular the walk is Markovian. In general the walks are non-markovian but that does not prevent from analyzing them. Also the biassed case can be treated and the exact drift etc. computed.

Example 3.23.: An example satisfying the assumptions of the Theorem is given in Figure 3.7. The subalphabets are $S_1 = \{1, 2\}$ and $S_2 = \{3, 4\}$. The 2×2 matrices on the main diagonal show that the CA action on invariant subalphabets is that of FCA 6 on binary doubles (see Example 3.5., just the coding is here different). The framed off-diagonal submatrices determine the boundary point motion.

	1	2	3	4
1	1	2	1	4
2	2	1	3	2
3	1	3	3	4
4	4	2	4	3

Figure 3.7.

The next step after pinning down the motion of individual boundary points is to characterize their interaction. As one might guess after seeing Figures 3.2. and 3.3. that in general this is a rather tall order. But it turns out that even the diffusive case i.e. random walk ensembles can be analyzed to a considerable extent (but not in this course).

3.6. Asymptotics

We will now take a step back from the partial permutivity formulation and consider general principles in the convergence of iterates of non-surjective CA. Our particular aim is to distill the appropriate concepts to use in this context. The fact alone that our dynamics is on an infinite sequence space forces some special consideration in comparison to e.g. flows on manifolds.

Given a map $T : X \to X$ we have already encountered invariant sets. Any limit set of a converging iteration clearly has to be invariant but to proceed a bit beyond this we first review some classical notions from the dynamical systems theory.

Definition 3.24.: *A point $x \in X$ is* **non-wandering** *for T if for any open neighborhood O of x there are arbitrarily large i such that $T^i(O) \cap O \neq \emptyset$. Let the collection of all such points be the* **non-wandering set** *$NW(T)$.*
*The ω-***limit set** *is the set of all possible limits: $\omega(x) = \{x' \in X | T^{i_k}(x) \to x'\}$. Its elements are called the ω-***limit points of** *x. If $x \in \omega(x)$ we call x* **positively recurrent**.
A closed invariant set $A \subset X$ is an **attracting set** *if there is a neighborhood U of A such that $T^i(x) \in U$ for all $i \geq 1$ and $T^i(x) \to A$ for all $x \in U$. The set $\cup_{i \leq 0} T^i(U)$ is the* **basin of attraction** *of A.*

Remarks: 1. A non-wandering point is on or near an orbit that comes back close to it. So fixed and periodic points are clearly non-wandering. $NW(T)$ is a closed and invariant set, hence necessarily contains the closure of all fixed and periodic points of the system. It also contains all ω-limit points. For a compact space X it is always nonempty. Wandering points are not relevant to the asymptotics whereas non-wandering points are.
2. An attracting set ultimately catches all orbits starting from its basin of attraction. The basins of attraction of distinct attractors are non-intersecting.

As we have seen in non-surjective CAs there is a fairly general phenomenon for evolutions from a random initial state to generate ever larger tiling-like patches. These structures correspond to bi-infinite sequences that are both spatially and temporally periodic. It is of major interest to characterize the convergence towards them. This is a delicate problem - sometimes results can be achieved, sometimes they are impossible, but for varying reasons.

Example 3.25. Recall the ECA Rule 184 of Example 3.4. It is fairly straightforward to see that among the invariant sets there are at least 0^∞, 1^∞, $0^\infty 1^\infty$ and $\{(01)^\infty, (10)^\infty\}$ (the third means that left and right tails are all-0 and all-1 respectively). The last set of course generates the prominent spatial/temporal period 2 checkerboard pattern.

Some limits sets are clear because of the dominance of certain 2-blocks (see also Exercises). For example if the initial configuration is $x = 0^\infty B 1^\infty$, where B is any finite block of 0's and 1's, then the $\omega(x) = 0^\infty 1^\infty$. If the infinite 0- and 1-tails are swapped, then $\omega(x') = \{(01)^\infty, (10)^\infty\}$. This is a simple but important observation: the given invariant set takes over since all particles annihilate by in finite time (by Lenght$(B)/2$ in this case).

If the initial configuration is distributed e.g. $B(1/2)$ then almost surely there is no finite time when every left- and right-going particle has forever cleared out a neighborhood of origin. Particles of each type will forever pass origin so there cannot be an attractor in the sense of Definition 3.24. for a set of full measure.

Indeed the situation is even more frustrating: by the same argument none of the given invariant sets can be attractors in the sense of the Definition above.

The Rule 184 is by various empirical studies a rather simple one in comparison to e.g. the Rule 110. In the latter rule the generic behavior nevertheless seems similar: ever larger patches of a spatial/temporal period 14/7 configuration form. However knowing that the rule is capable of universal computation we know that this certainly cannot prevail for all initial configurations. We can (at least in principle) program particles of various kind to arrive at origin at arbitrary large times. Characterizing a large nontrivial set of "dummy initial states" for which particles clear out from arbitrarily large neighborhoods of the origin seems very hard.

In the examples above we have already augmented the topological considerations by introducing a measure. Here is one attempt to proceed:

Definition 3.26.: *A periodic orbit P is a μ-***attractor** *if there is a set B of positive μ-measure with the property that $\omega(x) = P$ for all $x \in B$.*

Remarks 1.: The concept of μ-attractor was introduced by J. Milnor in 1985. Later M. Hurley showed that if a CA has a periodic μ-attractor P then the points of P must all be fixed by every shift of the lattice (for now just σ). Hence the points are constant sequences and there are at most $|S|$ of them. Moreover P is the only μ-attractor of the CA and $\omega(x) = P$ for μ-almost all $x \in X$.
2. It is worth noting that a periodic point being μ-attractor does not imply stability in the classical sense but allows orbit "detours". In smooth dynamics like that defined by differential equations the stability of a fixed point is usually due to contraction (which one can establish via linearization) and stability follows. CA being block maps on a Cantor set X are essentially never contractions.

The result above by Hurley applied to Rule 184 means that the only possible μ-attractors could be 0^∞ and 1^∞. But neither is a ω-limit set to a nontrivial set, so they are e.g. not $B(1/2)$-attractors. The Definition above seems like a step towards the right direction though since introducing ω-limit sets frees us from a predetermined iterate sequence by which the periodic state is reached.

To test the notions further let us also consider the second type of topological defects, the diffusive ones.

Example 3.27. Recall ECA Rule 18 of Example 3.17 ($001, 100 \mapsto 1$, else to 0, Cayley table in Figure 3.6., right). We have found its smallest invariant sets to be

0^∞ and $\{\{00,01\}^\infty, \{00,10\}^\infty\}$ We might want to call the latter the support of the **ground state**.

Using the principles of Section 3.5. we know that configurations of the form $\{00,01\}^\infty\{00,10\}^\infty$, both tails nontrivial, have exactly one topological defect in them. If we put $B(1/2)$ on the double symbols the usual way, the defect will persist for all times and perform a symmetric random walk. Hence by standard probability theory it is recurrent to origin (see e.g. [KT]). Consequently we cannot have the ground state as an attractor in the sense of Definition 3.24. although *most of the time* the neighborhood of origin looks exactly like the ground state.

If we use an initial measure introducing an ensemble of defects, various complications arise. If there were initially a finite number of defects, the measure again $B(1/2)$ the appropriate way and the defects would move jointly independently (which they don't), then almost surely they all would be annihilated in finite time except perhaps one (depending on the initial parity). In fact this can be proved for e.g. Rule 18 without any extra independence assumptions on the joint motion. So in this case the ω-limit would again be the ground state.

If the initial configuration is given from $B(p)$, $p \neq 0, 1$ on $\{0,1\}$ then there is an infinite number of defects at all times. However due to the diffusive behavior their density is expected to be decaying in time like $i^{-1/2}$. The same conclusion with respect to the ω-limit set should therefore hold.

The examples above indicate that perhaps the right mode of convergence to aim at is Cesaro-type. In the CA that we have considered for any $B(1/2)$-large set of initial values the defects will prevail for all times. But whether rectilinear or random motions, their density in configurations will decay as a consequence of annihilations and coalesings. Their passes of the origin, a phenomenon that sinks the Definition 3.24. for a general (non-periodic) attractor, are asymptotically negligible events if time averages are considered.

The following formulation, here converted to CA, is in the spirit of H. Hilmy, who proposed the idea already in the 1930's ([Hi]).

Definition 3.28.: *Given a cellular automaton F and $x \in X = S^{\mathbf{Z}}$, define* $\mathrm{Cent}(x)$ *to be the smallest closed subset $C \in X$ with the property that if U is a neighborhood of C then the proportion of the orbit $\{F^i(x)\}_{i=0}^{n-1}$ that is in U tends to one as $n \to \infty$. A set $A \subset X$ is a μ-**minimal center of attraction**, μMCA, if there is a set $B(A)$ with positive μ-measure with the property that* $\mathrm{Cent}(x) = A$ *for μ-almost every $x \in B(A)$.*

Using this Hurley proved that if a CA has a μ-minimal center of attraction A, then A is σ-invariant and unique. Furthermore $\mathrm{Cent}(x) = A$ for μ-a.e. $x \in X$. At a general level this clarifies the situation somewhat, but does not resolve any of the particular cases considered in Examples above.

References. Topological defects abound in CA and their understanding is essential for sorting out the statistical mechanics of CA. The reason for this is the same as in the context of the Ising model. The key objects are the kinks or the contours i.e. one- or two-dimensional defects, which determine the energy of the system ([Ba]). Defects in CA were first studied by Grassberger (see his article in the same special CA issue as [Li]). The notions of partial permutivity and invariant subalphabet were introduced in [EN], [E1] and [E2] to resolve a number of open questions concerning topological defects and their ensembles in one-dimensional CA. For more detailed takes on both classical and modern versions of concepts like attractor see e.g. [GH], [M2], [Hi] and in particular in CA context the paper [Hu].

The notion of partial permutivity generalizes naturally to higher dimensions. Deterministic Ising model (including voter models as zero temperature cases) and growth models have remarkably similar behavior to their classical counterparts which are utilizing various independence properties ([E3] and references therein).

4. Higher dimensional actions

4.1. Basic set-up

The subshifts considered in Chapter 2 involved one transformation, the left shift. Its integral powers were considered i.e. the action of the group $\mathbf{Z}$ on the space of allowed configurations or the action of the semigroup of non-negative integers in case of one-sided sequences. To restrict just to one action suffices if e.g. the interpretation of the action is that of the passing of time (one-dimensional event). However in many instances one might want to consider multiple actions defined on the same space. In this chapter we present the basic formulation of higher dimensional symbolic dynamics and investigate certain new phenomena that appear through a few examples.

The basic object of study is the following sequence space.

Definition 4.1.: *Let the set $S = \{0, 1, \ldots, n-1\}$ be the finite set of* **symbols** *i.e. an* **alphabet**. *If we assign one of the symbols to each* **site** *of the d-***dimensional integer lattice** $\mathbf{Z}^d$, *$d \geq 1$ we obtain the space $X = S^{\mathbf{Z}^d}$ of* **configurations**. *The cylinder sets generate the standard topology and we denote the σ-algebra of measurable sets again by $\mathcal{B}$.*

Remark: In most of the subsequent theory it suffices to consider the two-dimensional case as the essential differences to the one-dimensional case arise already there. We define the actions assuming the underlying lattice to be the square lattice but this is not essential. We could as well study actions on triangular lattice etc. Many of the models that one encounters can be defined on different lattices but only for a few the properties are distinctly different on different lattices of the same dimension.

The notation $x_{\mathbf{j}} = x_{(j_1, j_2, \ldots, j_d)}$ will be used to denote the symbol at the site $\mathbf{j} = (j_1, j_2, \ldots, j_d)$ of a given configuration $x \in X$. The set of configurations is a compact metric space for all $|S|$ and any dimension d. It can be metrized in many equivalent ways. An intuitively appealing choice is

$$d(x,x) = 0 \quad \text{and} \quad d(x,y) = 2^{-\min_{\mathbf{j}}\{\|\mathbf{j}\| \ | \ x_{\mathbf{j}} \neq y_{\mathbf{j}}\}} \quad \text{for} \quad x \neq y, \tag{4.1}$$

where $\| \ \|$ denotes the norm on the lattice. The geometric intuition behind this will become clear by thinking about $\mathbf{Z}^2$-lattice.

The basic dynamical operation on the configurations is again the (left, down, etc.) shift.

Definition 4.2.: *The* k^{th} **(coordinate) shift** σ_k *on a configuration* $x \in X$ *is defined as*

$$(\sigma_k x)_{(j_1, j_2, \ldots, j_d)} = x_{(j_1, j_2, \ldots j_k+1, \ldots, j_d)}.$$

Multi-indices provide a useful piece of notation: let $\sigma^{\mathbf{k}}$ denote the shift $\sigma_1^{k_1} \cdots \sigma_d^{k_d}$.

The shifts are homeomorphisms of the space X and they commute. This simple point is important to realize since non-commutative actions are in general much harder to analyze.

Analogously to one dimension we call

Definition 4.3.: *The dynamical system* $(X, \mathcal{B}, \sigma_1, \ldots, \sigma_d)$ *is the* d**-dimensional full shift.**

The full shift is again topologically transitive, it has unique measure of maximal entropy and $h_{top} = \log |S|$. It is indeed nearly as boring object as before but not quite as can be seen from the following factoid. From permutive cellular automata we know that there are many continuous shift-commuting onto maps F on $\{0,1\}^{\mathbf{Z}}$ such that $\left|F^{-1}(x)\right| = 2$ for all $x \in X$. In two dimensions there are none!

Let $W \subset \mathbf{Z}^d$ be a finite set called the **window** and $\pi_W : S^{\mathbf{Z}^d} \to S^W$ the natural projection i.e. $\pi_W(x) = x|_W$. Let $P \subset S^W$ be a set of **patterns** or **sceneries.**

Definition 4.4.: *Let*

$$X^{(W,P)} = \left\{ x \in X \;\middle|\; \pi_W\left(\sigma^{\mathbf{k}} x\right) \in P \ \forall \mathbf{k} \in \mathbf{Z}^d \right\}.$$

The dynamical system $\left(X^{(W,P)}, \mathcal{B}, \sigma_1, \ldots, \sigma_d\right)$ *is a* d**-dimensional topological Markov shift.**

Note that this of course subsumes the Definition 2.5. of a one-dimensional SOFT. If W is non-empty and $P = S^W$ then $X^{(W,P)}$ is the full shift whereas $P = \emptyset$ gives the empty shift. $X^{(W,P)}$ is always closed and shift-invariant subset of X i.e. $\sigma^{\mathbf{k}} X^{(W,P)} = X^{(W,P)}$ but for $d > 1$ usually much less trivial than the two extremes. For example there is a novel problem immediately at hand: the question whether $X^{(W,P)}$ is non-empty is *undecidable* i.e. there is no general finite procedure to check whether a given set of patterns allows global configurations! We'll return to this later. Meanwhile the reader might want to refresh details of the one-dimensional set-up and think why is there no such problem there.

Some unexpected problems withstood how does our scheme fit to the framework of ergodic theory discussed so far? Turns out that many of the definitions given in the case of a single transformation readily generalize (and the transformations do not need to be just shifts). The notion of invariant measure works in the obvious way to multiple actions: if the measure is preserved under all of the individual actions σ_k on $X^{(W,P)}$ then the measure is preserved under the **joint action** $\sigma^{\mathbf{k}}$ of the transformations (check). As before we can consider the ergodicity of individual actions but in general we would like to characterize the ergodicity/mixing etc. of the joint actions (recall the analysis of Rule 90 – there we had for the first time two commuting actions and it made a difference to know the ergodicity of joint actions). Note that the joint ergodicity of the two shifts on $\mathbf{Z}^2$ implies the ergodicity of the action to any *rational direction* k_2/k_1 (given by $\sigma_1^{k_1}\sigma_2^{k_2}$) which thereby implies the ergodicity of all one-dimensional actions.

Definition 4.5.: *A point* $x \in X^{(W,P)}$ *is* **periodic** *if it has a finite orbit i.e. the set* $\{\sigma^{\mathbf{k}}x|\ \mathbf{k} \in \mathbf{Z}^d\}$ *is finite.*

Remarks: 1. Note that the d-dimensional lattice $\mathbf{Z}^d$ is also a group, a discrete infinite abelian group addition being the operation (or rather the lattice is invariant under the action of this group). It has lot of subgroups and one should view them as groups of different types of isometries of the lattice. Another way of declaring periodicity of a point is to say that it is fixed under the action of a nontrivial subgroup: $\sigma^g x = x\ \forall g \in G$ (interpret the elements of G as multi-indices). The non-triviality just serves to dispose the cases $G = \{0\}$ fixing everything and $G = \mathbf{Z}^d$ fixing the configuration of period 1 to all coordinate directions i.e. a constant configuration. But one needs to add a bit: if G is a free group on d generators then x is periodic, otherwise it is periodic to just some rational directions.
2. Periodic points in two (and higher) dimensions also point to a new connection. Doubly periodic arrangements in the plane can of course be viewed as **tilings** the period rectangle/rhombus being the tile. Indeed these are just the simplest of tilings appearing in the context of $\mathbf{Z}^2$-actions. More complicated tilings do not correspond to periodic points but to other types of highly ordered configurations (think e.g. Penrose tiling).

Topological entropy is most straightforward to define using the counting argument. Consider a d-dimensional box centered at the origin containing N lattice points.

Count the number $N(C)$ of allowed configurations in it (the maximum is $|S|^N$). The topological entropy is the exponential growth rate of this quantity:

$$h_{top} = \lim_{N \to \infty} \frac{1}{N} \log N(C). \tag{4.2}$$

It is actually not necessary to take d-dimensional boxes but it is essential to retain a d-dimensional shape as otherwise one might asymptotically only count in a lower dimensional subspace and thereby only get some a "directional entropy"(for developments along this line see [M1]).

The transfer matrix formalism is available for higher dimensional subshifts only in special cases. Consequently there is no straightforward procedure to compute e.g. the topological entropy.

4.2. A few examples

There is no general theory of higher dimensional subshifts as of now. However a number of different types of subshifts that have been worked out. The interesting thing is that some of them have originally been statistical mechanics models, some discovered as tilings whereas some surface from abstract corners of number theory. In this section we indicate a few of them and point the problems involved. In order not to lose the picture we restrict here to examples that are two-dimensional.

Example 4.6.: Let $d = 2$, $S = \{0, 1\}$ and W be the window containing the coordinate points $(0,0), (0,1)$ and $(1,0)$. Suppose that the allowed patterns are as a and b in the Figure below.

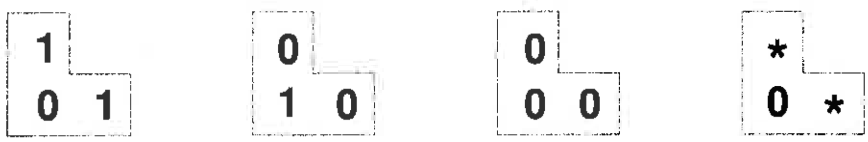

Figure 4.1a, b, c, d.

It is easy to see that the two legal patterns introduce a very rigid rule: there is a grand total of two elements in $X^{(W,P)}$ which are shifts of each other.

Of course we could use instead n symbols while still keeping the exclusion that the same symbol cannot neighbor horizontally or vertically. The case $n = 2$ is above. One might want to think of the symbols representing distinct colors so the problem is that of a map coloring and the formulations are sometimes refered to as

color models. Lieb has made significant progress on this problem when $n = 3$ (see [Ba]). In particular the topological entropy is then positive and one knows how to count periodic points. For $n > 3$ little seems to be known.

If S is as originally and P is augmented by the pattern in Figure 4.2. c (call the set P') it is easy to see that the configuration space becomes uncountable. However it is essentially one-dimensional subshift! All decreasing diagonals are constant and on increasing diagonals the rule is that of the one-dimensional full shift. Along rows and columns the rule is the golden mean of Example 1.1. i.e. no two 1's are allowed to be next to each other. Consequently the growth of the size of the space $X^{(W,P')}$ can be measured exactly as in the case the one dimensional golden mean. From this one-dimensional character it is easy to deduce that the topological entropy must be zero (exercise).

If instead the allowed patterns are as in Figures 4.1. b and d ($\star$ being the wild-card, here two independent copies of it) we have the two-dimensional golden mean which was introduced in Example 0.2. This is a rather subtle subshift – its analysis has eluded both the mathematicians and physicists for decades. The latter often refer to this model as "hard square gas": draw around each 1 a diamond with sidelength $\sqrt{2}$. The allowed configurations are exactly those where no two diamonds overlap. The same rule on the triangular lattice is the "hard hexagon model" which is exactly solvable! The model can be defined on various lattices, graphs and trees (on which there has been progress) and often refered to simply as the **Hard Core model.**

The topological entropy of the two-dimensional golden mean on $\mathbf{Z}^2$ is known at least to 14 digits (0.40749510126068...) but the exact value is unknown. As there is freedom in the choice of 1's in every diamond the entropy must be positive and loose bounds are fairly easy to show (exercise):

$$\frac{1}{2}\ln 2 < h_{top} < \ln\left(\frac{1+\sqrt{5}}{2}\right).$$

Example 4.7.: One-dimensional cellular automata as we have encountered in Chapter 3 can be viewed as two-dimensional subshifts. The change of viewpoint is easy: the window W is simply given by the geometry of the update. As an example Figure 4.2a. shows W for a radius r=2 rule. The set of acceptable patterns P in W is of course given by the CA rule.

Deterministic CA have zero topological entropy. This is an immediate consequence of the fact that the frame of width r determines the entire configuration in a $m \times m$ square. Indeed even less is needed. In the Figure we show the minimal frame determining the entire convex hull. However the area of the frame grows linearly in m (here it is 5m-4). So the number of possible configurations on the frame and hence inside the $m \times m$ square is bounded by 2^{cm}, some $c > 0$ independent of m. So h_{top} as defined by (4.2) must vanish.

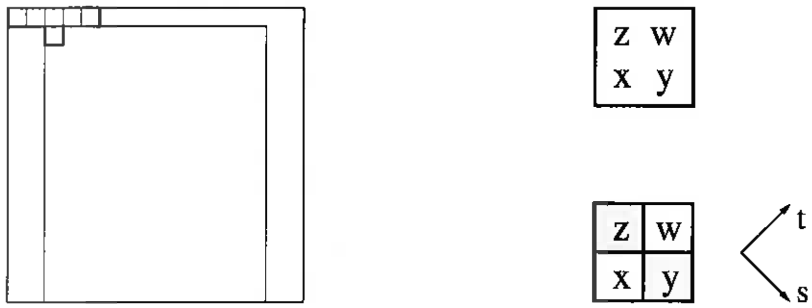

Figure 4.2a, b. Cellular automata as $\mathbf{Z}^d$-actions.

Example 4.8.: Let the alphabet and window be as in Example 4.6. and let the three entries in view be x, y, z. Requiring $x + y + z \equiv 0 \pmod 2$ for the legal patterns defines a two-dimensional SOFT. Once we identify time running towards SW this model becomes exactly the cellular automaton FCA 6 on binary doubles of Section 3.3.

Suppose W is a 2×2 window and the allowed patterns on the binary alphabet are determined by the rule $x + y + z + w \equiv 0 \pmod 2$ (Figure 4.2b.). Since each element can be solved as a sum of the three others, this rule can be viewed as a block map on two consequtive diagonals. This is an example of a **totalisticCA** since the update depends only on the *sum* of the entries in the block map. It is clearly an example of a reversible CA.

The rule exhibits a lattice isotropy in its arguments i.e. there is no distinguished time- or space-direction. In this set-up one can readily introduce anisotropy while preserving the invertibility of the CA by defining e.g. $w = f(z, y) + x \pmod 2$. Here f implements the spatial dependency on the nearest neighbors in the current generation. If f is symmetric in its arguments the rule still possesses spatial isotropy.

Example 4.9.: Let $S = \{0, 1, 2, 3\}$, and W as in Example 4.5. Suppose that the legal sceneries are

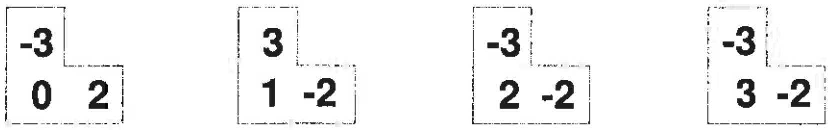

Figure 4.3.

Here the minus sign stands for exclusion: $-s$ means "any symbol but s". Checkerboard pattern of say 0's and 2's quickly indicates that $X^{(W,P)}$ is non-empty. Playing around with the patterns for moment shows that indeed there is a lot of legal configurations and that they have a certain character: the symbols always come in pairs. To see the geometry of the configurations let us call 0 "left", 2 "right", 1 "bottom" and 3 "top". The labels refer to the ends of a 1×2 or 2×1 tiles. The patterns just tell how to lay these in the plane in order to create a perfect tiling i.e. all the lattice sites covered and no overlapping tiles. For this reason the generation of the configurations of this subshift is also known as the "**domino/dimer problem**". In the theoretical computer science context the problem is formulated as finding the perfect matching on the graph $\mathbf{Z}^2$ and usually referd to as "independent sets" in their parlance.

The number of domino covers of a domain grows rather rapidly – for example the chess board has 12.988.816 different ones. It is easy to see that the system is in fact of positive entropy. The exact value of the topological entropy was established in the work of Kasteleyn, Temperley and Fischer (1961, see [Ba]). The analysis actually reveals a lot of detail and in particular the following. Define the set of configurations with horizontal period k and vertical period l simply by

$$P_{k,l} = \left\{ x \in X^{(W,P)} \mid \sigma^{(k,0)}x = \sigma^{(0,l)}x = x \right\}.$$

One can show that

$$\lim_{k,l\to\infty} \frac{1}{kl} \ln |P_{k,l}| = \frac{1}{4} \int_0^1 \int_0^1 \ln\left(4 - 2\left(\cos 2\pi t_1 + \cos 2\pi t_2\right)\right) dt_1 dt_2.$$

As the configurations of period (k, l) are a subset of all configurations this provides a lower bound for h_{top}. But it can be shown that it is a representative one: the number (≈ 0.2916) is exactly the topological entropy.

Example 4.10.: Let again $S = \{0, 1\}$, take a 3×3 square window, fix k, $0 \leq k \leq 9$ and let P_k be the set of patterns with exactly k 1's in the window. All the non-trivial subshifts $X_k = X^{(W,P_k)}$ are non-empty (convince yourself of this). In fact for $k \neq 0, 9$ they are uncountable (you only need to check $k = 1, 2, 3$ and

4, why?). Using a frame-argument analogous to that in Example 4.7. one can show that the topological entropy is zero for all of them. The systems are indeed rather rigid. Constructing their configurations one discovers that they have tiling-like qualities. This is due to a certain periodicity imposed by the rule. Moreover closer examination reveals that they have the following property indicating extreme rigidity or long range order: there are finite allowed configurations which can be extended in uncountably many different ways to global configurations but which can never be parts of the same configuration!

If the requirement "exactly k 1's in the window" is relaxed to "at most k 1's in the window" the resulting system has positive topological entropy and in many cases resembles the Hard Core model. It is possible to think this model as a high density packing limit of a Hard Core model. For further details on these models and their relations we refer to e.g. [E4].

4.3. Why is 2 so different from 1?

To really treat the Examples above and a few others properly would require another set of lecture notes. So in conclusion we'll instead be a bit philosophical and try to indicate what is the root of the difficulties encountered in two dimensions.

Given a SOFT, one- or two-dimensional, here is a list of some very basic questions one would like to settle.

1. Is the SOFT nonempty?
2a. Given a rational direction p_2/p_1 is there a point in $X^{(W,P)}$ which is periodic to that direction?
2b. Is there a rational direction such that the points periodic to that direction are actually dense?
3a. Are there periodic points?
3b. If the answer to 3a. is affirmative are they dense?
4. Does a given finite configuration extend to a global one?

First we note that the answers to the questions appropriate in one dimension (1, 3 and 4) all follow from the matrix formulation. Recall that the defining matrix of a SOFT gives us a directed graph representing the allowed transitions ($x_i \to x_{i+1}$). A global configuration is a bi-infinite path through this graph. But the graph is finite (since the matrix is) so there must be closed loops in it. Transversing such a loop gives a periodic point hence the SOFT is non-empty exactly when there are periodic points. Similarly the finiteness of the graph enables us to check (in a finite number of steps) whether a given finite block can be generated from the graph and if so we also know whether it can be extended to a global one. By the same vein we also know if this extension can be periodic and hence the question on the density of the periodic points is resolved.

Two dimensions (and higher) is different. But as in one dimension the questions formulated above are closely related. This is indicated also in the following result which was originally formulated and proved by Wang (1961) in tiling context.

Theorem 4.11.: *Let W be a $k \times k$ window and $X^{(W,P)}$ be a two-dimensional topological Markov shift.*
(i) It contains a point periodic to some direction p_2/p_1 iff it contains a periodic point.
(ii) If every non-empty $X^{(W,P)}$ contains a periodic point then the non-emptiness of $X^{(W,P)}$ is algorithmically decidable.

Proof: (i) The "if " is immediate. Without loss of generality we suppose that x is fixed under some power p of σ_h. Consider the infinite vertical strip in x between $(0,\cdot)$ and $(p+k-2,\cdot)$. By the horizontal periodicity the $k-1$-blocks $\left[x_{(0,n)},\ldots,x_{(k-2,n)}\right]$ and $\left[x_{(p,n)},\ldots,x_{(p+k-2,n)}\right]$ agree for all n. Then consider horizontal strips of height $k-1$. As there is an infinite number of them on a finite alphabet two of them must agree. So consider a finite window of width $p+k-1$ and height such that the top $k-1$ rows agree with the bottom $k-1$ rows. This is a periodic tile which we can use to tile the plane (with overlaps by strips of width $k-1$). Clearly the resulting configuration is periodic to any rational direction.

For (ii) we assume that every non-empty $X^{(W,P)}$ contains a periodic point. First check if any of the allowed patterns in P are both horizontally and vertically periodic i.e. whether the left and right $k-1$-columns agree and the top and bottom $k-1$ rows agree. If not, generate all allowed configurations of size $(k+1)\times(k+1)$. Check for the same periodicities. Continue to larger sizes. At some point we will either find a periodic square or the extension becomes impossible. The former implies the existence of a periodic point hence $X^{(W,P)}$ is non-empty. Continuing the extension indefinitely without encountering a periodic point contradicts the hypothesis as the limit is in $X^{(W,P)}$. ∎

From this we see that the questions 1. and 3a. as well as 2a. and 3a. are intimately connected. Similarly one can show that if every non-empty $X^{(W,P)}$ has a dense set of periodic points then the extension problem 4. is algorithmically decidable.

However it turned out that Theorem 4.11. (ii) was not quite as useful as Wang might have thought. Berger proved 1966 the following:

Theorem 4.12.: *There is a two-dimensional Markov shift without any periodic points. Moreover the question 1. is undecidable in two dimensions.*

Berger's original formulation is in terms of tilings. Then the statement is that given an arbitrary set of tiles the question whether they can be used to tile the plane is undecidable. This is a rather deep result – its shortest proof still seems to be a monograph... The idea has certain simplicity to it though. One shows that the set of subshifts under consideration is big enough to have extremely complicated elements. Specifically it has elements which if seen through a suitable coding are Turing machines i.e. capable of universal computation. From yet earlier results it is known that the halting problem for them is undecidable (i.e. given a universal computer and an input will the execution on it halt in a finite number of steps).

The undecidability results outlined here certainly indicate a very definite difference between one and two dimensions. 2 is not only twice as big as 1 – in some sense it is much bigger and for that reason even collections like the set of SOFTs that are quite innocent in one dimension are much less so in two dimensions.

But even if the questions 1-4 are resolved to a subclass of two-dimensional SOFTs their analysis may still be complicated by things not encountered in one dimension. The most basic ones stem from the fact that unlike in one dimension in $\mathbf{Z}^2$ there is a non-trivial neighborhood topology. This in turn is due to the fact that in one dimension it is possible to define total order whereas in two or higher it is not (for definition see [Rud]). Consequently things like counting the symbol variation allowed in a neighborhood (key to topological entropy) can be highly non-trivial.

References. There is no standard reference to higher dimensional subshifts. Several articles on them can be found e.g. in [Al]. In particular the one by Kitchens and Schmidt is relevant as it analyzes a class (essentially extensions of the cellular automaton Rule 6 on binary doubles) for which the decidability problems are no obstacle. Parts of [Sc] are useful in indicating the bigger picture in the algebraic context. The definite treatise on the work (up to 1982) on related lattice models in statistical mechanics is [Ba]. The undecidability results mentioned appeared first in the tiling context and can be found in a jewel of a book, [GS].

Appendix 1

The Perron-Frobenius Theory

For the purposes of 1-d symbolic dynamics we outline here the main results of the Perron-Frobenius theory. Some motivation and interpretation is given, but the actual proofs which are standard linear algebra can be found in matrix analysis treatises like [Ga] or in probability texts like [KT]. The presentation here follows the latter.

Let $A = (a_{ij})$ be a $n \times n$ matrix. If every element a_{ij} is non-negative, we write $A \geq 0$ and if additionally there is a positive element, we write $A > 0$ and call A **positive**. If furthermore every element is positive one marks $A \gg 0$. For vectors the notation is analogous: it is non-negative ($x \geq 0$) if all entries are non-negative, if at least one entry is positive, then $x > 0$ and if all are, then $x \gg 0$.

The First Perron-Frobenius Theorem: *If $A \gg 0$ then*
(i) there is $x^0 \gg 0$ such that $Ax^0 = \lambda_0 x^0$,
(ii) for an eigenvalue $\lambda \neq \lambda_0$ $|\lambda| < \lambda_0$,
(iii) λ_0 has geometric multiplicity 1 i.e. the corresponding eigenspace is one-dimensional

With some further work the result is recovered even under weaker condition $A > 0$ and $A^m \gg 0$ for some positive m. The latter part of this condition is equivalent to our formulation "irreducible and aperiodic".

To get some geometric intuition, it is useful to consider the matrix $P = xy$, where $Ax = \lambda_0 x$ and $yA = \lambda_0 y$ with the normalization $x^T y^T = 1$. So P is a matrix of rank 1 and furthermore it is easy to check that it also has the properties

(i) $Pz = \left(z^T y^T\right) x,\ wP = (wx)y$. In particular $Px = x$ and $yP = y$,
(ii) $P^2 = P$,
(iii) $AP = PA = \lambda_0 P$.

Theorem: *If $A > 0$ and $A^m \gg 0$ then $\frac{1}{\lambda_0^k} A^k \to P$.*

P is just the projection to the subspace spanned by the eigenvector associated to the dominant eigenvalue λ_0. So the Theorem says that high iterates of A look essentially like the action of this projection (together with some scaling and rotation if $\lambda_0 \neq 1$). The reason to this behavior and also the key part of the First Theorem is the existence of a spectral gap, (ii). Even without that some of the theory can be recovered in a weaker Cesaro form. This can be formulated as

The Second Perron-Frobenius Theorem: *Let $A > 0$ and λ_0 be as in the First Theorem. Then*
(i) the eigenvector x^0 corresponding to λ_0 is positive,
(ii) for any other eigenvalue λ it holds $|\lambda| \leq \lambda_0$,
(iii) $\frac{1}{m}\sum_{i=1}^{m} \frac{A^i}{\lambda_0^i}$ converges if $x^0 \gg 0$,
(iv) if λ is an eigenvalue of A and $|\lambda| = \lambda_0$, then $\eta = \lambda/\lambda_0$ is a root of unity and $\eta^m \lambda_0,\ m = 0, 1, 2, \ldots$ is an eigenvalue of A.

Finally we note that in the context e.g. Markov transition matrices the theorems both simplify and are highly useful. These are stochastic matrices i.e. positive matrices with row sums one. One can easily show that then always $\lambda_0 = 1$. But the eigenvalue does not need to be simple and the matrix does not need to be irreducible and aperiodic.

Appendix 2

Endomorphisms on compact groups

Theorem A2. (Halmos): *Suppose that G is a compact abelian group and A a continuous map onto G. Then A is ergodic with respect to the Haar measure if and only if the trivial character is the only $h \in \hat{G}$ for which $h\left(A^n\right) = h$ for some finite $n > 0$.*

Recall that the Haar measure on a group is the uniform Bernoulli measure (the different name refers to the fact that the measure is invariant under the group action whereas the Bernoulli measure does not need any action to start with). The Theorem gives a useful criteria for ergodicity in a special set-up: we only have to verify that orbits from all $h \neq 0$ under the endomorphism are infinite. The proof can be found in [Wa]. It is actually a relatively easy generalization of the argument for the ergodicity of the irrational rotation of the circle.

Version 1.1 29.11.2011 Kari Eloranta

Some notation

$\mathcal{A}, \mathcal{B}, \mathcal{C}$	σ-algebra, partition, cover, pp.
$O(x)$	Bi-infinite orbit of x under the given transformation, p.
δ_x	Pointmass at x, p.
$A_n(x)$	Average of an indicator function of a given set and
$S_n(\psi)$	a given function ψ along an orbit of length n from x, p.
$\mathcal{P} \vee \mathcal{P}'$	Join of two partitions or covers, p.
$H(\mathcal{P})$	Entropy of a partition and
$H(\mathcal{P}\|\mathcal{P}')$	conditional entropy of a partition, p.
$h(T, \mathcal{P})$	Entropy of a transformation with respect to a partition and
$h(T)$	entropy of a transformation, p.
k-name	$\mathbf{k} = (m_0, \ldots, m_{k-1})$ such that $T^i x \in P_{m_i}$, $i = 0, \ldots, k-1$, p.
$B(p_1 \ldots, p_n)$	Bernoulli-shift on n symbols, p.
(p, P)	Markov-shift with transition probability matrix P
	and equilibrium state p, p
X_A	1-dimensional sequence space with adjacency matrix A, p.
N_p^{per}	Number of periodic orbits of length p, p.
$Z_m(\psi)$	Finite partition function for the given potential ψ, p.
$P(\psi)$	Topological pressure of the potential ψ, p.
$\underline{S_1 S_2}$	Configuration generated piecewise from subalphabets S_i without
$\underline{S_1 S_2}(j)$	and with boundary specification, p.
$NW(T)$, $\omega(x)$	Non-wandering set of a transformation, ω-limit set of x, p.
$X^{(W,P)}$	Multidimensional sequence space with the allowed patterns P
	in the window A, p.
$\sigma^{\mathbf{k}}$	Multidimensional joint shift action, p.
$\|A\|$	Cardinality of the set A, p.
$d(x, y)$	Metric on the sequence space, p.

Bibliography

[Al] Alexander, J. C. (ed.): *Dynamical systems, Proceedings of the Special Year held at the University of Maryland, College Park, 1986-7*, Springer-Verlag, 1988.

[Bal] Baladi, V.: *Positive transfer operators and decay of correlations*, World Scientific, 2000.

[Bax] Baxter, R. J.: *Exactly solved Models in Statistical Mechanics*, Academic Press, 1982.

[Bi] Billingsley, P.: *Ergodic theory and information*, Wiley, 1965.

[Bo] Bowen, R.: *Equilibrium states and the ergodic theory of Anosov diffeomorphisms*, Springer Lecture Notes in Mathematics **470**, 1975.

[BKS] Bedford, T., Keane, M., Series, C. (eds.): *Ergodic theory, symbolic dynamics and hyperbolic spaces*, Oxford University Press, 1991.

[Ch] Chung, K. L.: *A second course in probability theory*, Academic Press.

[CBG] Conway, J. H., Berlekamp, E. R., Guy, R. K.: *Winning ways for your mathematical plays*, Academic Press, 1982.

[CFS] Cornfeld, I. P., Fomin, S. V., Sinai, Ya. G.: *Ergodic theory*, Springer-Verlag, 1982.

[CP] Coven, E., Paul, M.: *Math. Syst. Th.*, **8**, p. 167, 1974.

[DG] Dettmann, C.P., Georgiou, O.: Open mushroooms: stickiness revisited, *J. Phys. A: Math. Theor.*, **44**, 195102, 2011.

[E1] Eloranta, K.: Partially permutive cellular automata, *Nonlinearity*, **6**, No. 6, pp. 1009-1023, 1993.

[E2] Eloranta, K.: Random walks in cellular automata, *Nonlinearity*, **6**, No. 6, pp. 1024-1036, 1993.

[E3] Eloranta, K.: Cellular automata for contour dynamics, *Physica D* **89**, pp.184-203, 1995.

[E4] Eloranta, K.: A note on certain rigid subshifts, *Ergodic Theory of* $\mathbf{Z}^d$*-actions, M. Pollicott and K. Schmidt (eds.)*, London Math. Soc. Lecture Notes, **228**, pp. 307-317, Cambridge Univ. Press, 1996.

[EN] Eloranta, K., Nummelin, E.: The kink in the elementary cellular automaton Rule 18 performs a random walk, *J. of Stat. Phys.*, **69**, No.:5/6, 1992.

[Ga] Gantmacher, F. R.: *Applications of the theory of matrices*, Interscience, 1959.

[GH] Guckenheimer, J., Holmes, P.: Nonlinear oscillations, dynamical systems, and bifurcations of vector fields, Springer, 1983.

[GS] Grünbaum, B., Shephard, G.C.: Tilings and Patterns, Freeman, 1987.

[He] Hedlund G.A.: Endomorphisms and automorphisms of the shift dynamical system, *Math. Syst. Theor.*, **3**, 1969.

[Hi] Hilmy, H.: Sur les centresd'attractionminimaux des systemes dynamiques, *Compositio Math.*, **3**, pp. 227-238, 1936.

[HR] Hewitt, E., Ross, K.A.: *Abstract harmonic analysis*, Spriger, 1963.

[Hu] Hurley, M.: Varieties of periodic attractor in cellular automata, *Trans. AMS.*, **326**, pp. 701-726, 1991.

[Kar] Kari, J.: Reversible Cellular Automata, *DLT 2005, LNCS 3572*, pp. 57-68, 2005.

[Kat] Katznelson, Y.: *An introduction to harmonic analysis*, Dover, 1976.

[KH] Katok, A., Hasselblatt,B.: Introduction to the modern theory of dynamical systems, Cambridge univ. Press, 1995.

[Kr] Krengel, U.: *Ergodic Theorems*, de Grüyter, 1985.

[KT] Karlin, S., Taylor, H. M.: *A first course in stochastic processes*, Academic Press, 1975.

[KW] Katznelson, Y., Weiss, B.: A simple proof of some ergodic theorems, *Israel Journal of Mathematics*, **42**, No. 4, 1982.

[Li] Lind, D.: Applications of ergodic theory and sofic systems to cellular automata, *Physica D*, **10**, pp. 36-44, 1984.

[LM] Lind, D., Marcus, B.: *An introduction to Symbolic Dynamics and Coding*, Cambridge Univ. Press, 1995.

[M1] Milnor, J.: On the entropy geometry of cellular automata, *Complex Systems*, **2**, pp. 357-386, 1988.

[M2] Milnor, J.: On the concept of attractor, *Comm. Math. Phys.*, **99**, pp. 177-195, 1985; Correction and remarks **102**, pp. 517-519, 1985. item**[Pe]** Petersen, K.: *Ergodic Theory*, Cambridge University Press, 1983.

[Rud] Rudin, W.: *Real and complex analysis*, McGraw-Hill, 1978.

[Rue] Ruelle, D.: *Thermodynamic formalism*, Ensyklopedia of Mathematics, Addison Wesley, 1978.

[Sc] Schmidt, K.: *Algebraic ideas in ergodic theory*, CBMS Reg. Conf. Ser. in Math. **76**, American Mathematical Society, 1990.

[TM] Toffoli, T., Margolus, N.: *Cellular automata machines*, MIT Press, 1987.

[Wa] Walters, P.: *An introduction to ergodic theory*, Springer-Verlag, 1982.

[Wo] Wofram, S.: *Theory and applications of cellular automata*, World Scientific, 1986.