An Introduction to Entropy and Subshifts of Finite Type

Abby Pekoske
Department of Mathematics
Oregon State University
pekoskea@math.oregonstate.edu

August 4, 2015
Abstract

This work gives an overview of subshifts of finite type, and three measures of entropy on these systems. We develop Markov chain theory in order to define a subshift of finite type. We briefly define both topological and measure-theoretic dynamical systems in more generality. The three types of entropy defined here are topological, Shannon’s, and Kolmogorov-Sinai’s entropy.
Contents

1 Introduction 2

2 Markov Chains 4

3 Symbolic Dynamics 14

4 General Dynamical System 22

5 Shannon and Metric Entropy 25

6 Example 30
Chapter 1

Introduction

Subshifts of finite type, or topological Markov shifts, are a subset of a symbolic dynamical system that are largely studied in Ergodic theory. A symbolic dynamical system is a sequence space on a finite set of symbols together with a shift map on this space. For example, if we consider the set of infinite sequences of 0’s and 1’s, $X = \{0, 1\}^\mathbb{N}$, then the shift map, $\sigma : X \to X$, shifts each element of the sequence one index to the left, $\sigma(x)_i = x_{i+1}$. Subshifts of finite type are a closed, shift-invariant subset, which are defined by an incidence matrix. An incidence matrix, $A$ is a $(0, 1)$ matrix that indicates which symbols are allowed to follow another. For $i, j$ in a finite symbol set, $j$ is allowed to follow $i$ if $(a)_{ij} = 1$.

There is natural a correspondence between subshifts of finite type and Markov chains. A Markov chain is a sequence of random variables defined on a common probability space taking values in a finite state space. Markov chains possess a memory-less property; the past states of the process have
no influence on the future states. Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a probability space, and let \(I\) be a countable set which denotes the state space. Let \(\lambda = (\lambda_i : i \in I)\) denote the counting measure on \(I\). Then, because \(I\) is countable, \(\lambda\) is a probability distribution if \(\sum_{i \in I} \lambda_i = 1\). A random variable \(X\), is a function on the probability space with values in the state space, \(X : \Omega \rightarrow I\). If \(\lambda\) is a distribution, then we denote \(\lambda_i = \mathbb{P}(X = i) = \mathbb{P}(\{\omega : X(\omega) = i\})\). A Markov chain is a sequence \((X_n)_{n \geq 0}\) with an initial distribution \(\lambda\) and transition matrix \(P\) such that \(\mathbb{P}(X_0 = i_0) = \lambda_{i_0}\) and \(\mathbb{P}(X_{n+1} = i_{n+1}|X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = \mathbb{P}(X_{n+1} = i_{n+1}|X_n = i_n) = p_{i_n, i_{n+1}}\).

We will further explore the similarities of Markov chains and subshifts of finite type starting with Markov chains as a foundation. We also define a general dynamical system in order to define three types of entropy on subshifts of finite type. From a dynamical systems perspective, entropy is a measure of the randomness in the system. From the perspective of information theory, entropy measures the average amount of information that the system can produce. The three types of entropy that we explore are topological entropy, Shannon’s entropy, and measure-theoretic entropy due to Kolmogorov and Sinai. We conclude this work by providing a comprehensive example; we begin with a Markov chain, which defines a subshift of finite type. We then show computations involving each of the three measures of entropy, and interpret their meaning.
Chapter 2

Markov Chains

At the very foundation of a subshift of finite type is a Markov chain. A Markov process in general is characterized by a “memory-less” property. That is, the past states of the process have no influence on the future states. This means the conditional distribution of the next state given the current state is equivalent to the conditional distribution of the next state given all previous states. Markov chains are then a subset of Markov processes which have only a countable set of states. The content from this section can be found in detail in [3].

As a setting to the following work, we let \((\Omega, \mathcal{F}, \mathbb{P})\) be a probability space, and let \(I\) be a countable set which denotes the state space. Let \(\lambda = (\lambda_i : i \in I)\) denote the counting measure on \(I\). Then, because \(I\) is countable, \(\lambda\) is a probability distribution if \(\sum_{i \in I} \lambda_i = 1\). A random variable \(X\) is a function on the probability space with values in the state space, \(X : \Omega \to I\). If \(\lambda\) is a distribution, then we denote \(\lambda_i = \mathbb{P}(X = i) = \mathbb{P}(\{\omega : X(\omega) = i\})\). We say a
matrix, \( P = (p_{i,j})_{i,j \in I} \), is stochastic if every row \((p_{i,j})_{j \in I}\) is a distribution.

Let \( P \) be a stochastic matrix. The process \((X_n)_{n \geq 0}\) is a Markov chain with initial distribution if

(i) \( X_0 \) has initial distribution \( \lambda \) i.e. \( \mathbb{P}(X_0 = i_0) = \lambda_i \)

(ii) \( \mathbb{P}(X_{n+1} = i_{n+1}|X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = \mathbb{P}(X_{n+1} = i_{n+1}|X_n = i_n) = p_{i_n, i_{n+1}}. \)

The matrix \( P \) is known as the transition matrix. Note that we only consider chains such that \( \mathbb{P}(X_n = i|X_{n-1} = j) = \mathbb{P}(X_m = i|X_{m-1} = j) \) for all \( m \) and \( n \) and for all states \( i \) and \( j \). Because the Markov chain has no dependence on \( n \), we say that it is time-homogeneous.

A Markov chain is a a finite sequence of random variables with initial distribution \( \lambda \) and transition matrix \( P \). So the probability \( \mathbb{P}(X_{n+1} = j|X_n = i) \) is given by the \((i, j)\) entry of the transition matrix, \( p_{i, j} \). We abbreviate this by saying that \((X_n)_{n \geq 0}\) is Markov \((\lambda, P)\). Equivalently, a system is Markov if and only if the following proposition holds.

**Proposition 2.0.1.** A discrete-time random process \( X_1, X_2, \ldots \) is Markov with initial distribution \( \lambda \) and transition matrix \( P \) if and only if for all \( i_0, \ldots, i_N \in I \),

\[
\mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_N = i_N) = \lambda_{i_0} p_{i_0, i_1} p_{i_1, i_2} \cdots p_{i_{N-1}, i_N}.
\]

**Proof.** The forward direction follows directly from the definition of a Markov chain. So, assume that \( \mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_N = i_N) = \lambda_{i_0} p_{i_0, i_1} p_{i_1, i_2} \cdots p_{i_{N-1}, i_N} \) holds for \( N \), then an induction argument shows it holds for all \( n \leq N \). □

Let \( \delta_i = (d_{ij} : j \in I) \) denote the Dirac delta function.
Proposition 2.0.2 (Markov Property). Let \((X_n)_{n \geq 0}\) be Markov \((\lambda, P)\). Then conditional on \(X_m = i\), \((X_{m+n})_{n \geq 0}\) is Markov \((\delta_i, P)\) and is independent of the random variables \(X_0, \ldots, X_m\).

Proof. The proof follows from a straight forward application of the previous theorem. We first show that this holds for cylinder sets of the form \(A = \{X_0 = i_0, \ldots X_m = i_m\}\). Then because any event \(E = \bigcup_{k=1}^{\infty} A_k\) is a countable, disjoint union of these events, the result holds in general.

There is a natural correspondence between Markov chains and directed graphs. Let \(G_P\) denote the graph corresponding to a Markov process \((\lambda, P)\). The graph \(G_P\) has assigned vertices from the state space \(I\) and directed edges from \(i\) to \(j\) if \((p_{ij}) > 0\). We can think of the initial distribution of the process assigning probabilities or weights to the vertices, and the transition matrix assigning transition probabilities to the edges between vertices.

Example 2.0.1. Consider the process \((X_n)_{n \geq 0}\) defined on the state space \(I = \{1, 2, 3\}\), and let \(\lambda\) be some initial distribution. The graph below is represented by transition matrix \(P\) given by:

\[
P = \begin{pmatrix}
0 & 1 & 0 \\
0 & 1/2 & 1/2 \\
1/2 & 0 & 1/2
\end{pmatrix}.
\]

Then, the entry \((p_{ij})_{i,j \in I}\) represents the probability of transitioning from state \(i\) to state \(j\) in one time step. The initial distribution \(\lambda\) represents a weighting of the nodes on the graph. An initial distribution of \(\lambda = (1/2, 1/8, 3/8)\) represents the probabilities \(\mathbb{P}(X_0 = 1) = 1/2, \mathbb{P}(X_0 = 2) = 1/8, \mathbb{P}(X_0 = 3) = 3/8\). 

6
Theorem 2.0.1. Let \((X_n)_{n \geq 0}\) be Markov \((\lambda, P)\). Then for all \(n, m \geq 0\),

\begin{enumerate}
\item \(\mathbb{P}(X_n = j) = (\lambda P^n)_j\)
\item \(\mathbb{P}(X_n = j | X_0 = i) = \mathbb{P}(X_{n+m} = j | X_m = i) = p_{i,j}^{(n)}\).
\end{enumerate}

Proof. By Proposition 2.0.1,

(i)

\[
\mathbb{P}(X_n = j) = \sum_{i_0 \in I} \cdots \sum_{i_{n-1} \in I} \mathbb{P}(X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = j) \\
= \sum_{i_0 \in I} \cdots \sum_{i_{n-1} \in I} \lambda i_0 p_{i_0} p_{i_1} \cdots p_{i_{n-1} j} \\
= (\lambda P^n)_j.
\]

(ii) By the Markov property, conditional on \(X_m = i\), \((X_{m+n})_{n \geq 0}\) is Markov \((\delta_i, P)\). Therefore, if we let \(\lambda = \delta_i\) above, the result follows.

\[
\square
\]

Intuitively, this theorem makes sense when considering a walk on a graph. This gives us the probability of landing on vertex \(j\) in \(n\) time steps and the
probability of reaching vertex \( j \) from vertex \( i \) in \( n \) steps. Note here, \( p_{i,j}^{(n)} \) represents the \( n \)-step transition probability from \( i \) to \( j \).

We say that \( i \) leads to \( j \) if \( \mathbb{P}(X_n = j | X_0 = i) > 0 \) for some \( n \geq 0 \). This corresponds to a directed edge from vertex \( i \) to vertex \( j \). Then \( i \) communicates with \( j \) if both \( i \) leads to \( j \), and \( j \) leads to \( i \). The relation communicates forms an equivalence relation on \( I \) and partitions \( I \) into communicating classes. Given a fixed vertex, the communicating class is the set of vertices which can be reached by a walk on the graph. A class \( C \) is closed if \( i \in C \) and \( i \) communicates with \( j \) implies that \( j \in C \). Finally, we say that a transition matrix \( P \) is irreducible if \( I \) is a single communicating class.

Let \( \mathbb{P}_i(A) \) denote the probability that \( A \) occurs given \( X_0 = i \). We define a stopping time as a random variable \( T : \Omega \to [0, \infty] \) if the event that \( [T = n] \) is a function of only \( X_0, X_1, \ldots X_n \) for \( n < \infty \). For example, consider the process \( (X_n)_{n \geq 0} \) on a discrete state space, \( I \). The first time that the process reaches state \( i \) is a stopping time. This is known as the hitting time, \( \tau \), and is formally defined as

\[
\tau = \min\{n : X_n = i\}.
\]

We see that the hitting time is a stopping time, because \( \mathbb{P}(X_n = i) = \mathbb{P}(X_0 \neq i, \ldots, X_{n-1} \neq i, X_n = i) \).

**Theorem 2.0.2** (Strong Markov Property). Let \( (X_n)_{n \geq 0} \) be Markov \((\lambda, P)\); let \( T \) be a stopping time of the process \( (X_n)_{n \geq 0} \). Then conditional on \( T < \infty \) and \( X_T = i \), \( (X_{T+n})_{n \geq 0} \) is Markov \((\delta_i, P)\) and independent of \( X_0, X_1, \ldots, X_T \).
Proof. Let $A$ be an event depending only on $X_0, X_1, \ldots, X_T$. Then the event $A \cap \{T = m\}$ is determined by $X_0, X_1, \ldots, X_m$. By the Markov property at time $m$,

$$
\mathbb{P}(\{X_T = j_0, X_{T+1} = j_1, \ldots, X_n = j_n\} \cap A \cap \{T = m\} \cap \{X_T = i\})
$$

$$
= \mathbb{P}_i(X_0 = j_0, X_1 = j_1, \ldots, X_n = j_n)\mathbb{P}(A \cap \{T = m\} \cap \{X_T = i\}).
$$

By definition of conditional probability, if we sum over $m = 0, 1, 2, \ldots$ and divide by the probability that $\mathbb{P}(T < \infty, X_T = i)$, we find that

$$
\mathbb{P}(\{X_T = j_0, X_{T+1} = j_1, \ldots, X_n = j_n\} \cap A \mid T < \infty, X_T = i)
$$

$$
= \mathbb{P}_i(X_0 = j_0, X_1 = j_1, \ldots, X_n = j_n)\mathbb{P}(B \mid T < \infty, X_T = i).
$$

For a state $i \in I$, the probability that $X_n = i$ infinitely often is denoted by $\mathbb{P}(X_n = i \ i.o.)$. We say that state $i$ of a Markov process is recurrent if $\mathbb{P}_i(X_n = i \ i.o.) = 1$ and transient if $\mathbb{P}_i(X_n = i \ i.o.) = 0$. The first passage time $T_i$ is the first time that the process returns to state $i$: $T_i(\omega) = \inf\{n \geq 1 : X_n(\omega) = 1\}$.

The notion of passage time allows us to say that every state is either recurrent and transient.

**Proposition 2.0.3.** The following holds:

(i) If $\mathbb{P}_i(T_i < \infty) = 1$, then $i$ is recurrent, and $\sum_{n=0}^{\infty} p_{ii}^{(n)} = \infty$.

(ii) If $\mathbb{P}_i(T_i < \infty) < 1$ then $i$ is transient, and $\sum_{n=0}^{\infty} p_{ii}^{(n)} < \infty$.

Proof. A proof of this depends on a concept of the number of visits to a state, which is not covered in this work. As a result, the proof can be found in [3].

Note that this allows us to observe that every state is either transient or recurrent. All states of a closed class are either recurrent or all transient. Also,
every recurrent class is closed, and every finite closed class is recurrent. These conditions allow us easily observe recurrent and transient states. The only recurrent states are those in the closed classes, and all others are transient.

**Example 2.0.2.** Consider a Markov chain \((X_n)\) on the state space \(I = \{0, 1, 2, 3\}\), with some initial distribution \(\lambda\) with non-zero entries. If the transition matrix is

\[
P_1 = \begin{pmatrix}
\frac{1}{4} & 0 & \frac{3}{4} & 0 \\
1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 \\
\frac{2}{3} & 0 & 0 & \frac{1}{3}
\end{pmatrix}
\]

then starting in states 0, 2, or 3, there is no way to transition to state 1 with positive probability. This is easily observable from the graph:

Thus, state 1 is transient and also not contained in a closed class. If we let
the transition matrix be

\[ P_2 = \begin{pmatrix}
1/4 & 0 & 3/4 & 0 \\
1/2 & 0 & 1/2 & 0 \\
0 & 0 & 1 & 0 \\
2/3 & 0 & 0 & 1/3 \\
\end{pmatrix} \]

then every state is recurrent, and \( I \) is a closed class. We can verify this with the graph below:

Many of the asymptotic properties of Markov chains involve the idea of an invariant distribution. Recall that a measure \( \lambda \) is a row vector \((\lambda_i)\) with non-negative entries. The measure \( \lambda \) is invariant if \( \lambda \) is a left eigenvector of the transition matrix \( P \);

\[ \lambda P = \lambda. \]

We also say that \( \lambda \) is an equilibrium or stationary measure. The following proposition proves existence of invariant measures on a finite state space.

**Proposition 2.0.4.** Let \( I \) be finite. Suppose for some \( i \in I \) that \( p_{ij}^{(n)} \to \pi_j \) as \( n \to \infty \) for all \( j \in I \). Then \( \pi = (\pi_j : j \in I) \) is an invariant distribution.

**Proof.** In the following computation because \( I \), is finite we can exchange summation and limits. Using the hypothesis and definition of distribution, we
compute
\[ \sum_{j \in I} \pi_j = \sum_{j \in I} \lim_{n \to \infty} p_{ij}^{(n)} = \lim_{n \to \infty} \sum_{j \in I} p_{ij}^{(n)} = 1, \]
so \( \pi \) is a distribution. To show it is invariant, we find
\[ \pi_j = \lim_{n \to \infty} p_{ij}^{(n)} = \lim_{n \to \infty} \sum_{k \in I} p_{ik}^{(n)} p_{kj} = \sum_{k \in I} \lim_{n \to \infty} p_{ik}^{(n)} p_{kj} = \sum_{k \in I} \pi_k p_{kj}. \]

Recall that irreducibility of a single class, \( I \), implies that starting in a given state, with positive probability, the process can reach any other state within \( I \).

**Proposition 2.0.5.** Let \( P \) be irreducible, and suppose that \( P \) has an invariant distribution \( \pi \). Let \( \lambda \) be any distribution. Suppose \((X_n)_{n \geq 0}\) is Markov \((\lambda, P)\). Then
\[ P(X_n = j) \to \pi_j \text{ as } n \to \infty \text{ for all } j \]
and
\[ p_{ij}^{(n)} \to \pi_j \text{ as } n \to \infty \text{ for all } i,j. \]

**Proof.** A detailed proof of this proposition can be found in [3]. □

If the transition matrix \( P \) is irreducible, then an equilibrium distribution exists if and only if the expected time it takes for the process to return to a particular state \( i \) is finite. If an equilibrium distribution \( \mu \) exists, then it is also unique.
A proof of this can be found in [3]. Note that a Markov chain which is a single closed communicating class has an equilibrium distribution.
Chapter 3

Symbolic Dynamics

We now have the tools to begin defining a symbolic dynamical system. The content from this section can be found in detail in Kitchens Symbolic Dynamics [2]. The topic of main concern is a subshift of finite type which can be constructed by modifying a Markov chain.

We first construct a sequence space and describe its topology. Consider the finite set $A = \{0, 1, 2, \ldots, n\}$. This set is known as the alphabet and is analogous to the state space for Markov chains. We endow this set with the discrete topology and consider the sets of infinite and biinfinite sequences. That is if we let $\mathbb{N}$ denote the set $\{0, 1, 2, \ldots\}$ and $\mathbb{Z}$ denote the integers, then $A^\mathbb{N}$ is the one-sided sequence space with elements of the form $(a_0a_1a_2a_n \ldots)$ for $a_i \in A$. The two-sided sequence space is $A^\mathbb{Z}$ with elements of the form $(\ldots a_{-2}a_{-1}a_0a_1a_2a_n \ldots)$ for $a_i \in A$. These sequence spaces are endowed with the product topology generated by the usual product metric. We will use the following metrics on these sequences spaces, which are equivalent to the prod-
uct metrics. For \( x, y \in \mathcal{A}^N \), \( d(x, y) = 1/2^k \), where \( k \) is the first occurrence that the sequences \( x \) and \( y \) differ: \( k = \max\{m : x_i = y_i \text{ for } i < m\} \). Analogously, for \( x, y \in \mathcal{A}^\mathbb{Z} \), \( d(x, y) = 1/2^k \), where \( k = \max\{m : x_i = y_i \text{ for } |i| < m\} \).

We denote the space of infinite and biinfinite sequences by \( X_n \) and \( \Sigma_n \) respectively. We are mainly concerned with the dynamics of these spaces. Thus, we define the left shift transformation. This transformation shifts the sequence elements by one index to the left. Formally, \( \sigma : X_n \to X_n \) and \( \sigma : \Sigma_n \to \Sigma_n \) by \( \sigma(x)_i = x_{i+1} \). For example, if we consider a \((0,1)\) alternating sequence in \( \mathcal{A}^\mathbb{Z} \), \((\ldots 0101.0101\ldots)\), then the shift transformation acts on this sequence in the following way:

\[
\sigma(\ldots 0101.0101\ldots) = (\ldots 1010.1010\ldots),
\]

and

\[
\sigma^2(\ldots 0101.0101\ldots) = (\ldots 0101.0101\ldots),
\]

where \( \sigma^k(x) \) denotes the \( k \)th iteration of the map \( \sigma \). These spaces together with the shift transformation \((X_n, \sigma)\) and \((\Sigma, \sigma)\) are known as the one-sided shift on \( n+1 \) symbols, or the full one-sided \((n+1)\)-shift and the two-sided shift on \( n + 1 \) symbols, or the full \((n+1)\)-shift.

We are mainly concerned with characterizing long-term behaviors of this dynamical system under iteration of the shift map. We will mostly be working with the full one-sided shift; however, there are analogous definitions for the full shift as well. A point \( x \in X_n \) is periodic, if \( \sigma^p(x) = x \) for some \( p \in \mathbb{N} \). The point \( x \) has period \( p \) if \( x \) is periodic and \( p \) is the minimal, nonzero power
satisfying \( \sigma^p(x) = x \). For example, in \( \{0, 1, 2\}^\mathbb{N} \), the point \( x = (.01201201\ldots) \)
is \(3k\) periodic for \( k \in \mathbb{N} \), and has period 3, because

\[
\sigma(.01201201\ldots) = (.12012012\ldots)
\]

\[
\sigma^2(.01201201\ldots) = (.20120120\ldots)
\]

\[
\sigma^3(.01201201\ldots) = (.01201201\ldots).
\]

A point is eventually periodic if after a finite number of iterates, it is periodic. That is \( x \in X_n \) is \emph{eventually periodic} if \( \sigma^k(x) = \sigma^{k+p}(x) \) for \( p \in \mathbb{N} \).

The types of sequences we will be working with are subshifts of finite type or SSFT. There is a natural correspondence between a SSFT and a Markov chain. Consider a square, \((0,1)\), irreducible matrix. The matrix \( A \) is the \emph{incidence matrix} of a graph \( G_A \) if there is an edge from vertex \( i \) to \( j \) if and only if \( A_{ij} = 1 \). Note that \( A \) cannot have any complete rows or columns of zeros. The incidence matrix \( A \) agrees with the transition matrix \( P \) of a Markov chain if \( (p_{ij}) > 0 \) if and only if \( (a)_{ij} = 1 \). Recall that in order for \( P \) to be a stochastic matrix, the row sums of \( P \) must equal 1. Any matrix \( P \) which agrees with \( A \) defines the incidence matrix of a Markov chain with some initial distribution \( \lambda \). Recall that a matrix \( A \) is irreducible if for a finite state space \( I \), \( I \) is a single class, meaning the graph \( G_A \) is strongly connected. Thus, the matrix \( A \) is irreducible if for every pair of indices \( i, j \) there exists a \( k > 0 \) with \( (A^k)_{ij} > 0 \).

The dynamical system with sequences defined by an incidence matrix are SSFT’s. Let \( A \) be an incidence matrix with rows and columns indexed by
\{0, 1, \ldots, n\}. We let the matrix \( A \) define the allowable sequences in \( \{X_n\} \) as the set \( \{x \in X_n : A_{x,x_{i+1}} = 1 \text{ for all } i \in \mathbb{N}\} \). These define a set of closed, shift-invariant sequences in \( X_n \). We define a one-sided subshift of finite type defined by \( A \), or one-sided topological Markov shift defined by \( A \) as the dynamical system consisting of these sequences, along with the restricted shift map. This new dynamical system is denoted by \( (X_A, \sigma) \). There is also an analogous definition of Subshifts of Finite Type for the full-shift, denoted by \( (\Sigma_A, \sigma) \).

**Example 3.0.3.** Consider the subshift of finite type \( (X_A, \sigma) \) given by the incidence matrix

\[
A = \begin{pmatrix}
0 & 1 \\
1 & 0
\end{pmatrix}.
\]

This consists only of the sequences (.010101\ldots) and (.101010\ldots), because the subwords 00 and 11 are not allowed. This system is compatible with the Markov chain, \( (X_n) \), with any initial vector, say \( \lambda = (1/5, 4, 5) \), and transition matrix

\[
P = \begin{pmatrix}
0 & 1 \\
1 & 0
\end{pmatrix}.
\]

This is the only allowable transition matrix, because the entries \( (p_{ij}) > 0 \) if and only if \( (a_{ij}) = 1 \), and \( P \) must be a stochastic matrix.

Notice that the transition matrix defines the allowable words or sequences. We can think of these as infinite walks on the graph associated with the transition matrix. Another way that a subshift of finite type is defined is by the finite list of unallowable transitions. This list then excludes every sequence that has these symbols as subwords or every sequence within the cylinder set.
generated by the excluded list of words. In the previous example, the set of unallowable words is \{00, 11\}.

Recall that a matrix \( A \) is irreducible and aperiodic if for some power \( k \), \((A^k)_{ij} > 0 \) for all \( i,j \) in the state space.

**Theorem 3.0.3** (Perron Frobenius Theorem). *Suppose that \( P \) is a nonnegative, square matrix. If \( P \) is irreducible and aperiodic, there exists a real eigenvalue \( \lambda > 0 \) such that:

(i) \( \lambda \) is a simple root of the characteristic polynomial;

(ii) \( \lambda \) has strictly positive left and right eigenvectors;

(iii) the eigenvectors of \( \lambda \) are unique up to constant multiple;

(iv) \( \lambda > |\mu| \), where \( \mu \) is any other eigenvalue;

(v) if \( 0 \leq B \leq P \) (entry by entry) and \( \beta \) is an eigenvalue for \( B \) then \( |\beta| \leq \lambda \) and equality can occur if and only if \( B = P \);

(vi) \( \lim_{n \to \infty} \frac{P^n}{\lambda^n} = rl \) where \( l \) and \( r \) are the left and right eigenvectors for \( P \) normalized so that \( l \cdot r = 1 \).

**Proof.** The heart of the proof is to show that there is a single eigenline passing through the first quadrant, and that every eigenvector approaches this line under iteration by \( A \). The Contraction Mapping theorem is then used to show that there is a unique fixed point under iteration, which turns out to be the single eigenline in the first quadrant. This ensures that the eigenvalue associated with this line is strictly larger than the other eigenvalues. \( \square \)
If the matrix $A$ is irreducible but not aperiodic, then there is another version of Perron-Frobenius theorem with slight changes. In part (iv), $\lambda > |\mu|$ becomes a non-strict inequality. Also, part (vi) of the theorem is dropped. The eigenvalue and eigenvector satisfying the above conditions are known as the Perron value and Perron vector.

Perron-Frobenius theorem allows us to characterize long-term behavior of the dynamical system just based on the transition matrix. An allowable block or word of length $k$ is the finite sequence of length $k$ that may appear in a sequence of the dynamical system. That is, $[i_0, \ldots, i_{k-1}]$ is an allowable word if $A_{i_ri_{r+1}} > 0$ for $r = 0, \ldots, k - 2$. Notice that we can find these allowable words by iterating the transition matrix; $(A^k)_{ij}$ gives the number of allowed words of length $k+1$ starting at $i$ and ending at $j$. We denote the set of allowable words of length $k$ by $W(A,k)$. For a given sequence $x \in X_n$, $|W_x(A,k)|$ counts the number of subwords of length $k$ in $x$ and is known as the complexity function. We obtain the total set of allowed words by taking the union over all $k$ of sets $W(A,k)$, and denote this by $W(A)$.

We now have the tools to be able to define the first notion of entropy. For subshifts of finite type, topological entropy is a measure of the exponential growth rate of the cardinality of the set $W(A)$, or the number of allowable words. Let $A$ be a transition matrix and $|W(A)|$ denote the cardinality of $W(A)$. Then the topological entropy of both the one-sided subshift of finite type defined by $A$ and the two-sided subshift of finite type defined by $A$ is defined as

$$\lim_{k \to \infty} \frac{\log |W(A,k)|}{k}.$$
Traditionally, the base of the logarithm is 2. Topological entropy is denoted by $h_{\text{top}}(X_A, \sigma)$ and $h_{\text{top}}(\Sigma_A, \sigma)$ for the shift acting on $X_A$ and $\Sigma_A$ respectively. Topological entropy can be thought of as a measure of randomness in the system; the larger the exponential growth rate of allowable words, the more randomness the dynamical system possesses. Alternatively, entropy may be thought of as the amount of information that is stored in the set of allowed sequences. We will see more of this notion when defining Shannon entropy and metric entropy.

**Proposition 3.0.6 (3.0.6).** Let $A$ be an irreducible transition matrix. Then $h_{\text{top}}(X_A, \sigma) = h_{\text{top}}(\Sigma_A, \sigma) = \log \lambda$, where $\lambda$ is the spectral radius of $A$. If $A$ is irreducible, then $\lambda$ is the Perron value of $A$.

**Proof.** Note that $(A^k)_{ij}$ represents the number of allowed words of length $k+1$ beginning with $i$ and ending with $j$. Thus, the cardinality $|W(A, k)|$ is given by $\sum_{ij}(A^{k-1})_{ij}$. By Perron-Frobenius theorem, there exist constants $c$ and $d$ such that

$$c\lambda^k \leq |W(A, k)| \leq d\lambda^k$$

for sufficiently large $k$. When we take the logarithm, dividing by $k$ and taking the limit as $k$ approaches infinity, we see that

$$\log \lambda \leq \lim_{k \to \infty} \frac{1}{k} \log |W(A, k)| \leq \log \lambda.$$  

Note that both $\log c$ and $\log d$ disappear in the limit. Therefore, by the definition of topological entropy, $h_{\text{top}}(X_A) = h_{\text{top}}(\sigma_A) = \log \lambda$. \qed
Note that the result also holds for reducible matrices.
Chapter 4

General Dynamical System

We now define a measure theoretic dynamical system in order to define the last two notions of entropy. The content in this section can be found in detail in Fogg’s Substitutions in Dynamics, Arithmetics and Combinatorics [1].

We will still be working with symbolic dynamical systems, just in higher generality. We must define a topological dynamical system in order to define a measure-theoretic dynamical system. A topological dynamical system \((X, T)\) is a compact metric space \(X\) together with an onto, continuous map \(T : X \rightarrow X\).

Notice that for a given transition matrix the spaces of subshifts of finite type with the shift map, \((X_A, \sigma)\) and \((\Sigma_A, \sigma)\), satisfy the definition of a topological dynamical system. The spaces \(X_A\) and \(\Sigma_A\) are both compact under the metric, \(d(x, y) = 1/2^k\), where \(k\) is the first place that \(x\) and \(y\) differ, and the associated map \(\sigma\) is onto and continuous.

Let \(f\) be a Borel measurable function defined on \(X\). Then \(f\) is \(T\)-invariant if for all \(x \in X\), \(f(Tx) = f(x)\). We want to endow the space \((X, T)\) with a
measure. Note that a topological dynamical system, \((X, T)\), always has an invariant probability measure. This follows from Fogg’s proof of the existence of an invariant measure on a compact metrizable space [1]. A measure-theoretic dynamical system is a system \((X, \mathcal{B}, \mu, T)\), where \((X, T)\) is a topological dynamical system, \(\mu\) is a probability measure defined on the Borel measurable subsets of \(X\), denoted by \(\mathcal{B}\), and \(T : X \to X\) is a measurable map which preserves measure; that is, for \(x \in X\), \(\mu(Tx) = \mu(x)\).

Notice that the system of subshifts of finite type defined by an incidence matrix \(A\) is a measure-theoretic dynamical system, when we associate \(A\) with a transition matrix, \(P\) of a Markov chain \((\lambda, P)\).

**Example 4.0.4.** Consider the full shift on \(\{0, 1\}^\mathbb{N}\). Here, the incidence matrix,

\[
A = \begin{pmatrix}
1 & 1 \\
1 & 1
\end{pmatrix},
\]

defines the topological dynamical system, \((X_A, \sigma)\), where \(\sigma\) is the shift map, \(\sigma : X_A \to X_A\) by \(\sigma(x)_i = x_{i+1}\). The Borel sigma-algebra is generated by the cylinder sets discussed in section 3, under a measure defined by a Markov chain. Consider the Markov chain \((X_n)\) with initial distribution \(\lambda = (3/5, 2/5)\) and transition matrix

\[
P = \begin{pmatrix}
1/2 & 1/2 \\
3/4 & 1/4
\end{pmatrix}.
\]

Notice that \(\lambda P = \lambda\), so \(\lambda\) defines an equilibrium distribution on \((X_n)\). This Markov process defines a measure, \(\mu\), on \((X_A, \sigma)\) by \(\mu([x_0 x_1 x_2 \ldots x_n]) = \lambda x_0 p_{x_0 x_1} p_{x_1 x_2} \ldots p_{x_{n-1} x_n}\) for \(x_i \in \{0, 1\}\) for all \(i\). For example, if we consider
the cylinder set [001011], the measure of this set is

\[ \mu([001011]) = \frac{3}{5} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{3}{4} \cdot \frac{1}{2} \cdot \frac{3}{4} = \frac{27}{640}. \]

Therefore, the system of the one-sided full shift with measure \( \mu \) and shift map \( \sigma \) defines a measure-theoretic dynamical system, \((X_A, B, \mu, \sigma)\).
Chapter 5

Shannon and Metric Entropy

We now have the tools to define Shannon and metric entropy. The content in this section can be found in detail in [5], [4], or [6]. Both Shannon entropy and metric entropy quantify the uncertainty and information content of a symbolic dynamical system. The following section illustrates this idea.

Let \( \alpha = \{A_1, A_2, \ldots, A_N\} \) be a measurable partition of the probability space \((X, \mathcal{B}, \mu)\), and let \( T : X \to X \) be a measurable map over \( X \). If we let \( \alpha(x) \) be the element of \( \alpha \) containing \( x \in X \), then we define the itinerary of \( x \) as the sequence \( (\alpha(x), \alpha(Tx), \alpha(T^2x), \ldots) \) which takes values in \( A_1, A_2, \ldots, A_N \). Suppose \( x \) is not known, but we know its itinerary through \( \alpha(T^{n-1}) \). We wish to quantify how much uncertainty we have about the value \( \alpha(T^n x) \) in order to define both types of entropy.

Example 5.0.5. Let \( S \subset \mathbb{C} \) denote the unit circle in the complex plane. That is, \( S = \{z \in \mathbb{C} : |z|=1\} \). We will consider the transformation \( T : S \to S \) defined by \( T(z) = z^3 \). This map triples the angle on the unit circle. Consider
the partition $\alpha = \{A_0, A_1\}$, where $A_0 = \{e^{i\theta} : 0 \leq \theta < \pi\}$ and $A_1 = \{e^{i\theta} : \pi \leq \theta < 2\pi\}$. If $z \in \mathbb{S}$ is unknown, then the itinerary of $(0,0,0,1,1\ldots)$ means that $z \in A_0$, $T(z) \in A_0$, $T^2(z) \in A_0$, $T^3(z) \in A_1$, $T^4(z) \in A_1$. If we know the first digits of the itinerary of $z$, what can we say about the $n$th digit of the itinerary? Knowing the first five digits of the itinerary tells us no new information about the sixth, because we do not know the initial angle, $\theta$. This means that there is an equal probability of the sixth element being either a 0 or a 1; thus the ‘information’ gained when learning $\alpha(T^5(z))$ is a constant $1/2$.

Let $(X, \mathcal{B}, \mu)$ be a probability space. Suppose that $x \in X$ is unknown. How can we quantify the information content of the event $x$ is in $A$? Intuitively, we would like the information content of an event $E$ to quantify the amount of uncertainty lost when we learn that $x \in A$. If the probability of event $A$ occurring is large, then the uncertainty lost when learning that $x$ belongs to $A$ is small. In other words, $I(A)$ is large when $\mu(A)$ is small. Let $I(A)$ denote information content. The information content function is defined by the following three axioms.

(i) $I(A)$ is a continuous function of $\mu(A)$,

(ii) $I(A)$ is a non-negative, decreasing function in $\mu(A)$, and if $\mu(A) = 1$, then $I(A) = 0$,

(iii) If $A, B$ are independent, meaning $\mu(A \cap B) = \mu(A)\mu(B)$, then $I(A \cap B) = I(A) + I(B)$.
The only functions $\phi : [0, 1] \to \mathbb{R}^+$ such that $I(A) = \phi[\mu(A)]$ that satisfy these three conditions for all probability spaces $(X, \mathcal{B}, \mu)$ are $c\ln t$ with $c < 0$.

We now are able to define the information content and Shannon’s notion of entropy. Let $(X, \mathcal{B}, \mu)$ be a probability space.

The \textit{Information Content} of a set $A \in \mathcal{B}$ is $I_\mu(A) = -\log \mu(A)$.

The \textit{Information Function} of a finite measurable partition, $\alpha$, is

$$I_\mu(\alpha)(x) = \sum_{A \in \alpha} I_\mu(A)1_A(x) = -\sum_{A \in \alpha} \log \mu(A).$$

The \textit{entropy} of a finite measurable partition, $\alpha$ is the average value of the information content of the elements of the partition. \textit{Shannon’s entropy} is defined as,

$$H_\mu(\alpha) = \int_X I_\mu(\alpha) \, d\mu = -\sum_{A \in \alpha} \mu(A) \log \mu(A).$$

Traditionally, log base 2 is used for defining information content, and $0 \log 0 = 0$.

Shannon’s entropy has several useful properties that are analogous to many properties of a probability measure. For each of the three definitions above, we can define a conditional version based upon conditional probabilities.

Let $(X, \mathcal{B}, \mu)$ be a probability space. Let $\alpha$ be a finite measurable partition. Suppose that $\mathcal{F}$ is a sub-$\sigma$ field of $\mathcal{B}$. Let $\mu(A|\mathcal{F})(x)$ denote the conditional expectation of the indicator of $A$ given $\mathcal{F}$.

The \textit{information content} of $A$ given $\mathcal{F}$ is $I_\mu(A|\mathcal{F})(x) := -\log \mu(A|\mathcal{F})(x)$.

The \textit{information function} of $\alpha$ given $\mathcal{F}$ is $I_\mu(\alpha|\mathcal{F}) := \sum_{A \in \alpha} I_\mu(A|\mathcal{F})1_A$. 

27
The conditional entropy of $\alpha$ given $\mathcal{F}$ is $H_\mu(\alpha|\mathcal{F}) := \int_X I_\mu(\alpha|\mathcal{F}) \, d\mu$.

Let $\alpha, \beta$ be two countable partitions of a common probability space $X$. If every element of $\alpha$ is equal up to a $\mu$ set of measure 0 to a union of elements of $\beta$, then we say that $\alpha$ is courser than $\beta$, and $\beta$ is finer than $\alpha$. We define $\alpha \lor \beta = \{A \cap B : A \in \alpha \text{ and } B \in \beta\}$ as the smallest partition that is finer than both $\alpha$ and $\beta$. Note that if $\mathcal{F}_1$ and $\mathcal{F}_2$ are both $\sigma$-algebras, then $\mathcal{F}_1 \lor \mathcal{F}_2$ is the smallest $\sigma$-algebra which contains both $\mathcal{F}_1$ and $\mathcal{F}_2$.

Let $(X, \mathcal{B}, \mu, T)$ be a measure-theoretic dynamical system. The following definition is due to Kolmogorov and Sinai. Metric entropy is defined as

$$h_\mu = \sup \{h_\mu(T, \alpha) : \alpha \text{ is a countable measurable partition such that } H_\mu(\alpha) < \infty\}$$

where $h_\mu(T, \alpha) = \lim_{n \to \infty} H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i} \alpha\right)$.

Note that the limit that defines $h_\mu(T, \alpha)$ exists. A proof of this relies on the subadditivity of Shannon entropy, and can be found in [6].

Shannon entropy, $H_\mu(\alpha_n)$ gives the “average information content in the first $n$-digits of the $\alpha$-itinerary”. When we divide by $n$, this gives the average “information per unit time.” Therefore, we can think of metric entropy as measuring the maximal rate of “information production the system is capable of generating.”

**Theorem 5.0.4 (5.0.4).** Let $\mu$ be a translation invariant measure for a Markov chain with transition matrix $P = (p_{ij})$ and initial distribution $(p_i)$. Then

$$h_\mu(\sigma) = - \sum_{i,j} p_i p_{ij} \log p_{ij}.$$ 

**Proof.** Let $\alpha$ be the partition generated by the cylinder sets $\{[a] : a \in I\}$ and
let $\alpha_0^{n-1}$ denote the refinement $\sigma(\bigcup_{i=0}^{n-1} T^{-i} \alpha)$. Then a computation shows that
\[
H_\mu(\alpha_0^{n-1}) = n \left( \sum_{i,j} p_i p_{ij} \log p_{ij} \right) - \sum_i p_i \log p_i.
\]
Now dividing by $n$ and taking the limit as $n$ approaches infinity, we find,
\[
\lim_{n \to \infty} \frac{1}{n} H_\mu(\alpha_0^{n-1}) = \lim_{n \to \infty} \left( \sum_{i,j} p_i p_{ij} \log p_{ij} - \frac{1}{n} \sum_i p_i \log p_i \right).
\]
Therefore, $h_\mu(\sigma) = -\sum_{i,j} p_i p_{ij} \log p_{ij}$. Then because $\alpha$ generates the Borel sigma-algebra, by Sinai’s Generator Theorem in [5], $h_\mu(\sigma) = h_\mu(\sigma, \alpha)$. 

$\square$
Chapter 6

Example

Consider the Markov chain \((X_n)\) with initial distribution \(\lambda = (1/2, 1/4, 1/4)\) and permutation matrix \(P\) with associated graph representation below,

\[
P = \begin{pmatrix}
    1/4 & 3/4 & 0 \\
    0 & 0 & 1 \\
    1 & 0 & 0
\end{pmatrix},
\]

Notice that every state in \(I = \{0, 1, 2\}\) is recurrent, because the graph of the Markov chain \((\lambda, P)\) is a closed class, meaning we can transition from a given state to any other state with positive probability.
The corresponding incidence matrix for a subshift of finite type is given by

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix},$$

and dynamical system $(X_A, \sigma)$ under the shift operator. We see that $A$ is irreducible, with period 1, and is therefore aperiodic. This incidence matrix defines a shift with the set of unallowable subwords: $\{02, 10, 11, 21, 22\}$. We see that $(.00\ldots)$ is the only fixed point, because there is only a loop on state 0. There are no period 2 points, because the shortest complete walk on the graph, $G_A$, other than around the 0-loop has length 3. We can create periodic points of any period greater than or equal to 3 by the following. Note that $(.012012012\ldots)$ has a period of 3. Because 00 is an allowable subword, we can create a word of period $k \geq 3$ by $(.00\underbrace{\ldots012}_{k-2}\underbrace{012\ldots012}_{k-2}\ldots)$.
The Markov chain \((\lambda, P)\) defines a measure \(\mu\) on the shift space \((X_A, \sigma)\).

Consider the cylinder sets \([000]\), \([012]\), \([1200]\). Then

\[
\mu([000]) = 1/2 \cdot 1/4 \cdot 1/4,
\]
\[
\mu([012]) = 1/2 \cdot 3/4 \cdot 1,
\]
\[
\mu([1200]) = 1/4 \cdot 1 \cdot 1 \cdot 1/4,
\]

because \(\mu([a_0a_1a_2 \ldots a_n]) = \lambda_{a_0}p_{a_0a_1}p_{a_1a_2} \ldots p_{a_{n-1}a_n}\).

By Proposition [3.0.6], to compute topological entropy, we need to find the Perron root of the incidence matrix, \(A\). We compute the characteristic polynomial of \(A\) by finding the determinant, \(\det(A - xI)\), where \(I\) is the identity matrix. Thus, \(A\) has characteristic polynomial

\[
c_T(x) = -x^3 + x^1 + 1.
\]

We see that the roots of \(c_T\) are approximately \(-.23279 + .79255i, -.23279 - .79255i, and 1.46557, all of which are eigenvalues. By the Perron-Frobenius theorem, because \(A\) is a nonnegative, square, irreducible and aperiodic matrix, the Perron root is 1.46557. We will denote this by \(\psi = 1.46557\). The Perron vector is the associated eigenvector: \(v = (1.46557, .682328, 1)\).

We can now compute the topological entropy of \((X_A, \sigma)\). This is defined
as the limit of the number of allowable subwords per unit time:

\[ h_{\text{top}}(X_A, \sigma) = \lim_{k \to \infty} \frac{1}{k} |W(A, k)| \]

\[ = \log \psi \]

\[ \approx 0.551462. \]

In order to compute Shannon’s entropy, we must consider a partition of \( X_A \). Consider the partition of length 2 cylinder sets, \( \alpha_1 = \{[00], [01], [12], [20]\} \). Let these partition elements be \( \{A_1, A_2, A_3, A_4\} \) respectively. Then by the measure \( \mu \) defined by \((\lambda, P)\), we find

\[ \mu(A_1) = \frac{1}{8} \]
\[ \mu(A_2) = \frac{3}{8} \]
\[ \mu(A_3) = \frac{1}{4} \]
\[ \mu(A_4) = \frac{1}{4}. \]

Recall that for a finite partition \( \alpha \), Shannon’s entropy is defined as

\[ H(\alpha) = -\sum_{A \in \alpha} \mu(A) \log \mu(A). \]

So we compute Shannon’s entropy of the partition \( \alpha_1 \):

\[ H(\alpha_1) = -\frac{1}{8} \log \frac{1}{8} - \frac{3}{8} \log \frac{3}{8} - \frac{1}{4} \log \frac{1}{4} - \frac{1}{4} \log \frac{1}{4} \]

\[ \approx 1.905639 \]
using log base 2.

Now consider the partition of length 3 cylinder sets, 
\( \alpha_2 = \{ [000], [001], [012], [120], [200], [201] \} \). We find the measure of each partition element using the measure defined by the Markov chain \((\lambda, P)\),

\[
\begin{align*}
\mu(A_1) &= 1/32 \\
\mu(A_2) &= 3/32 \\
\mu(A_3) &= 3/8 \\
\mu(A_4) &= 1/4 \\
\mu(A_5) &= 1/16 \\
\mu(A_6) &= 3/16.
\end{align*}
\]

Again, if we find Shannon’s entropy of this new partition,

\[
H(\alpha_2) = - \sum_{A \in \alpha_2} \mu(A) \log \mu(A)
\]

\[
= -1/32 \log 1/32 - 3/32 \log 3/32 - 3/8 \log 3/8 \\
- 1/4 \log 1/4 - 1/16 \log 1/16 - 3/16 \log 3/16
\]

\[
\approx 2.20986.
\]

Because Shannon’s entropy is a measure of the average value of the information content of the elements of a partition, it makes sense that \( \alpha_2 \) has higher entropy than \( \alpha_1 \); the partition elements of length three provide more information on average than the partition elements of length two. We would
be slightly better at guessing a word from the system knowing the first three symbols versus the first two. However, because we can consider partitions of any length, Shannon’s entropy is unbounded. Thus an entropy value of approximately two is a small value.

Last, we find the metric entropy of the system \((X_A, \sigma)\). We must find an equilibrium distribution on the system defined by \((\lambda, P)\). We need to find a measure \(\pi\) such that \(\pi P = \pi\). Using

\[
P = \begin{pmatrix}
0 & 3/4 & 0 \\
0 & 0 & 1 \\
1 & 0 & 0
\end{pmatrix},
\]

by solving the system of linear equations, we find that

\[
\pi = (2/5, 3/10, 3/10).
\]

Note that the norm of \(\pi\) is one, because \(\pi\) defines a probability measure. By Theorem [5.0.4] we find,
\[ h_\mu(\sigma) = - \sum_{i,j} \lambda_i p_{ij} \log p_{ij} \]

\[ = -\left( \frac{2}{5} \cdot \frac{1}{4} \log \frac{1}{4} + \frac{2}{5} \cdot \frac{3}{4} \log \frac{3}{4} + \frac{1}{4} \cdot 0 \log 0 \right) \]

\[ + \frac{3}{10} \cdot 0 \log 0 + \frac{3}{10} \cdot 0 \log 0 + \frac{3}{10} \cdot 1 \log 1 \]

\[ + \frac{3}{10} \cdot 1 \log 1 + \frac{3}{10} \cdot 0 \log 0 + \frac{3}{10} \cdot 0 \log 0 \]

\[ = \frac{1}{5} + \frac{3}{10} \log \frac{4}{3} \]

\[ \approx .075489. \]

using log base 2. Note that by convention, \(0 \log 0 = 0\). This value of entropy represents the maximal rate of information production the system is capable of generating. Low entropy is not a surprising result. The system is highly predictable, because the only variability within a word is placement of 12 between a string of 0’s of a variable length. This is because according to the incidence matrix \(A\), a 2 must follow a 1, and a 0 must follow a 2.

This example is an illustration of the ways in which a Markov chain and the very structure of the subshift of finite type affect the measures of entropy. We see that topological entropy measures the exponential growth rate of the number of allowable words in a system. This is an invariant defined on a topological system. Shannon’s entropy on the other hand is highly dependent on the measure associated with the measure-theoretic system. It also measures the average amount of information of elements of a partition. Measure-theoretic entropy is defined using an invariant measure, and gives a measure
of the average amount of information production that the system is capable of generating.

Given the foundation of subshifts of finite type built upon Markov chains, we would like to explore more general dynamical systems. According to the variational principle, topological entropy of a system is equal to the supremum over all invariant measures of measure-theoretic entropy. The theory which develops this result seems highly interesting, but is beyond the scope of this paper. Future work would include learning more of this theory, and exploring other types of dynamical systems including substitutions.
Bibliography


