Chapter 07: Introduction to Ergodicity and Dynamical Systems

Sources

- Thermodynamics of Chaotic Systems by C. Beck, and F. Schlögl (1993)
- Symbolic Dynamics and Coding by D. Lind, and B. Marcus (1995)
- Lecture Notes on Ergodic Theory by Prof. Omri Sarig (Penn State, Fall 2008)
- Ergodic problems of Classical Mechanics by V.I. Arnold and A. Avez (1968)
- A First Course in Dynamics: with a Panorama of recent developments by B. Hasselblatt and A. Katok (2003)
- Studies of Nonlinear Problems by E. Fermi, J. Pasta and S. Ulam (May 1955)

Topics

- Introduction to measure (probability) preserving transformations and Poincare’s recurrence theorem
- Definitions of Ergodicity and Mixing, Examples
- Ergodic theory of finite Markov Chains
- Symbolic Dynamics and Coding
- Fermi-Pasta-Ulam-Tsingou problem
- Discussion on Ergodicity and Chaos

** Due to lack of time and adequate mathematical background, some of the theorems will be stated without proof.
In this chapter, we will start our discussion with a classical Hamiltonian system with $3N$ degrees of freedom. Let a system with $N$ particles be confined to a volume $V$ by an idealized surface which prevents heat or particle flows. The position and momentum phase space variables are $q_1, \ldots, q_j, \ldots, q_{3N}$ and $p_1, \ldots, p_j, \ldots, p_{3N}$, respectively. They are also assumed to obey classical Hamiltonian dynamics: $\dot{q}_j = \frac{\partial H}{\partial p_j}$ and $\dot{p}_j = -\frac{\partial H}{\partial q_j}$ for $j = 1, 2, \ldots, 3N$. Thus, we have a $6N$-dimensional phase space and, at a particular instant, the state of the system can be represented by $6N$-dimensional vector $x$. If we specify the energy $E$ of the system, then the motion of the $N$-particle system will be confined to a $(6N - 1)$-dimensional hyper-surface $\Gamma(E)$ given by $H(x) = E$.

Now for such a system, a remarkable fact is given by the Liouville’s theorem. In the present context, consider a set of initial conditions in the $6N$-dimensional phase space with a certain hypervolume. Then, the hypervolume does not change as the system evolves under the action of the Hamiltonian dynamics described above.

**Theorem 0.1. (Classical version of Liouville’s Theorem)**

$$\frac{\partial}{\partial t} \rho + \sum_{i=1}^{3N} \left[ \frac{\partial \rho}{\partial q_i} \dot{q}_i + \frac{\partial \rho}{\partial p_i} \dot{p}_i \right] = 0$$

**Proof of Liouville’s theorem**

Let us consider $N$ particles with canonical coordinates $q_i$ and conjugate momenta $p_i$, with their dynamics governed by Hamilton’s equations:

$$\dot{q}_i = \frac{\partial H}{\partial p_i} \quad \text{and} \quad \dot{p}_i = -\frac{\partial H}{\partial q_i}$$

which yields the following identity:

$$\frac{\partial \dot{q}_i}{\partial q_i} + \frac{\partial \dot{p}_i}{\partial p_i} = 0$$

because the Hamiltonian $H$ is a potential function and therefore the mixed partial derivatives are independent of their order, i.e., $\frac{\partial^2 H}{\partial q_i \partial p_i} = \frac{\partial^2 H}{\partial p_i \partial q_i}$.

**Remark 0.1.** The state of the system is uniquely and completely defined by specifying the $6N$ variables $\{q_i\}, \{p_i\}$, which span a $6N$ dimensional space called the phase space. A system in a definite microstate is represented by a point in the phase space. As the system evolves in time according to Hamilton’s equations of motion, the point executes a trajectory in the phase space.
Next let us consider an “ensemble” of states, which can be thought of as a collection of “representative” micro-states of $N$ particles governed by the same Hamiltonian dynamics. This ensemble is represented as a collection of points in the phase space with density $\rho(\{q_i\}, \{p_i\}, t)$, which can in general have explicit time dependence. As usual, the density is defined so that the volume element

$$\prod_{i=1}^{3N} dq_i dp_i \equiv d^{3N}q \, d^{3N}p$$

at the point $(\{q_i\}, \{p_i\})$ contains

$$\rho(\{q_i\}, \{p_i\}, t) \, d^{3N}q \, d^{3N}p$$

points at time $t$. The Liouville theorem states that

$$\frac{d}{dt} \left( \rho(\{q_i\}, \{p_i\}, t) \right) = 0.$$  

Noting that $(\{q_i\}, \{p_i\})$ implicitly depends on time, the physical meaning of Liouville’s theorem is that if we follow one particular point in time, the density at this point is invariant. In other words, if we think of the ensemble of systems as a fluid of fictitious “particles” moving in the phase space, then this fluid is “incompressible,” i.e. maintains its density, and therefore volume, during the flow. This dynamics of systems in the phase space should not be confused with the motion of individual particles within each system; in particular, the “particles” in the phase space have no mutual interactions, as they correspond to different physical systems.

The Liouville theorem can be proven by noting that the number of points in the phase space is locally conserved. Since each point represents a system, it cannot suddenly disappear, and it must move continuously in the phase space. Therefore, the current density of points in the phase space, given by

$$\vec{J} = \rho \vec{v}$$

where $\vec{v} \triangleq (\{\dot{q}_i\}, \{\dot{p}_i\})$ is a $6N$ dimensional vector, which satisfies the continuity equation:

$$\frac{\partial}{\partial t} \rho + \nabla \cdot \vec{J} = 0$$
A combination of Eq. (6) and Eq. (7) yields
\[
\frac{\partial}{\partial t} \rho + \sum_{i=1}^{3N} \left[ \frac{\partial \rho}{\partial q_i} \dot{q}_i + \frac{\partial \rho}{\partial p_i} \dot{p}_i \right] + \sum_{i=1}^{3N} \left[ \frac{\partial \dot{q}_i}{\partial q_i} + \frac{\partial \dot{p}_i}{\partial p_i} \right] \rho = 0
\]

\[
\Rightarrow \frac{\partial}{\partial t} \rho + \sum_{i=1}^{3N} \left[ \frac{\partial \rho}{\partial q_i} \dot{q}_i + \frac{\partial \rho}{\partial p_i} \dot{p}_i \right] = 0
\]

by using the identity in Eq. (2). This proves the Liouville theorem that follows from Hamilton’s equations.

**Remark 0.2.** It is noted that Liouville’s theorem assumes nothing beyond Hamilton’s equations. In particular, Liouville’s theorem is valid independent of whether the system is in equilibrium or not, and it provides foundations for non-equilibrium statistical mechanics.

We also present the quantum mechanical version of Liouville’s Theorem for the sake of completeness of this section. Let \( \rho \) be the density operator defined by the density matrix \( \rho_{m,n} \) as:
\[
\rho_{m,n} \triangleq \langle \Phi_n, \rho \Phi_m \rangle = \delta_{mn} |b_n|^2
\]
where \( \{\Phi_n\} \) is a standard set of complete orthonormal wave functions and \( \{b_n\} \) is a set of complex weights whose phases are random numbers and the magnitude is either 1 or 0, depending on whether the energy of the \( n^{th} \) state \( E_n \in (E, E+\Delta) \) or \( E_n \notin (E, E+\Delta) \).

**Theorem 0.2.** (Quantum mechanical version of Liouville’s Theorem)
\[
i\hbar \frac{\partial \rho}{\partial t} = [\mathcal{H}, \rho]
\]

where \( \hbar = 1.054572 \times 10^{-34} \) Joule-sec is the the modified Planck constant; \( \mathcal{H} \triangleq -\frac{\hbar^2}{2m} \nabla^2 + V \) is the Hamiltonian operator; \( V \) is the potential energy function; and \([A, B] \triangleq AB - BA\) is the commutator of operators \( A \) and \( B \).

Next Liouville’s theorem is formally stated in a measure-theoretic sense. Let \( X \) be the set of all states in the phase space of a given dynamical system and \( T : X \to X \) be the map (i.e., the law of motion) that describes the system evolution as the time progresses. Then, the sequence \( \{T^n(x)\}_{n\in\mathbb{Z}} \) of maps is called an orbit of the system. With these terminologies and focusing our attention on conservation of volume measure, we proceed as follows.
Theorem 0.3. (Measure-theoretic version of Liouville’s Theorem) The Lebesgue measure \( \mu \) on \( X \) satisfies the condition: \( \mu(T^n(E)) = \mu(E) \) for all \( n \in \mathbb{Z} \) and all measurable \( E \subset X \).

Liouville’s theorem is a major motivation behind investigation of measure-preserving transformation, which leads to the Poincare’s recurrence theorem and eventually leads to the ergodic theory. The primary application of the ergodic theory is the fact that most often we are unable to work out the detailed dynamics of \( x(t) \), but we may have access to averages over the surface \( \Gamma(E) \). The simple assumption that makes this averaging meaningful is a form of the ergodic hypothesis: \textit{Over any significant time interval } \( \tau \), \textit{the phase point } \( x(t) \) \textit{will spend equal time intervals in all regions of } \( \Gamma(E) \).

An equivalent alternative statement will be presented later in this chapter. However, another hypothesis, stronger than the ergodic hypothesis is known as the mixing hypothesis. This refers to the notion that an initial compact distribution of points on the energy surface \( \Gamma(E) \) in the phase space very quickly permeates the entire surface while occupying the same volume (as required by the Liouville’s theorem). Both of the hypothesis play important roles in the fundamental arguments of statistical mechanics. In this chapter, we will study these notions with a special emphasis on the ergodic theory of Markov chains.

1 Measure-preserving Transformation and Poincare Recurrence

Definition 1.1. (Measure Space) A measure space is a triplet \((X, \mathcal{B}, \mu)\) where

1. \( X \) is a set (e.g., \( \mathbb{R} \)).

2. \( \mathcal{B} \) is a \( \sigma \)-algebra (i.e., a collection of subsets of \( X \) that contains \( \emptyset \) and which is closed under complements and countable unions). Elements of \( \mathcal{B} \) are called measurable sets.

3. \( \mu : \mathcal{B} \to [0, \infty] \) is called the measure and it satisfies the countable subadditivity property, i.e., if \( E_1, E_2, \cdots \in \mathcal{B} \), then \( \mu(\bigcup_i E_i) \leq \sum_i \mu(E_i) \) and the equality holds if, in addition, \( E_1, E_2, \cdots \) are (pairwise) disjoint subsets of \( X \).

4. If the measure is finite, i.e., \( \mu : \mathcal{B} \to [0, \infty), \) then we can normalize this measure by dividing by \( \mu(X) \) so that \( \mu : \mathcal{B} \to [0, 1] \), which is called the probability measure.
Definition 1.2. (Measure Preserving Transformation) A measure preserving transformation (MPT) is a quartet \((X, \mathcal{B}, \mu, T)\) where \((X, \mathcal{B}, \mu)\) is a measure space and the following two conditions hold:

1. \(T : X \rightarrow X\) is a \(\mathcal{B}\)-measurable function, i.e., \(E \in \mathcal{B}\) implies \(T^{-1}E \in \mathcal{B}\).
2. \(\mu\) is \(T\)-invariant, i.e., \(\mu(T^{-1}E) = \mu(E) \forall E \in \mathcal{B}\).

An Equivalent probabilistic Setting (Stochastic Process)

- \(X\) is the sample space, i.e., the collection of all possible states \(\omega\) of a random system.
- \(\mathcal{B}\) is the collection of all measurable events, i.e., all states \(E \subseteq X\), such that it is known whether \(\omega \in E\) or \(\omega \notin E\).
- \(\mu\) is a probability measure, i.e., \(Pr[\omega \in E] := \mu(E) \in [0, 1]\).
- Measurable functions \(f : X \rightarrow \mathbb{R}\) are random variables \(f(\omega)\)
- The random sequence \(\{X_n\}\), where \(X_n \triangleq f \circ T^n (n \geq 1)\), whose distribution is given by

\[
Pr[X_{i_1} \in E_{i_1}, \ldots, X_{i_k} \in E_{i_k}] \triangleq \mu\left(\bigcap_{j=1}^{k} \{\omega \in X : f(T^{i_j}\omega) \in E_{i_j}\}\right)
\]

In this setting, an MPT on a probability space is called a probability preserving transformation (PPT) which means

\[
Pr[X_{i_1+m} \in E_{i_1+m}, \ldots, X_{i_k+m} \in E_{i_k+m}] = Pr[X_{i_1} \in E_{i_1}, \ldots, X_{i_k} \in E_{i_k}] \ \forall m
\]

The above equation is known as the strict sense stationarity condition.

Ergodic theory is the study of measure preserving actions of measurable maps on a measure space (usually with a finite measure, e.g., a probability space). The above setting is a way of studying stochastic phenomena that evolve in a deterministic context. i.e., if the system state \(x\) is known at time \(t = 0\), then the system state will be \(T^n(x)\) at \(t = n\) with full certainty).

Theorem 1.1. (Poincaré’s Recurrence Theorem) Let \((X, \mathcal{B}, \mu, T)\) be a PPT. If \(E\) is a \(\mathcal{B}\)-measurable set, then for almost every \(x \in E\), there exists a sequence of PPTs such that \(T^k(x) \in E\) for some \(k \in \mathbb{N}\).

(Sketch of proof using Liouville’s theorem)

Let us consider a time-independent Hamiltonian \(H(p, q)\) for the case where all orbits are bounded. (Note that this occurs if the energy surface is bounded,
i.e., there are no solutions of \( E = H(p, q) \) with \( \| p \| \to \infty \) or \( \| q \| \to \infty \).

Let us choose an arbitrary initial point in the phase space within an open region of arbitrarily small diameter \( \varepsilon > 0 \). Poincare’s Recurrence Theorem states that, if there are points which leave the initial region \( N^0 \), almost all of these points will return to the original region \( N^0 \) in finite time. Now let us consider the PPT \( T \) that evolves in forward time such that the initial region \( N^0 \) is mapped to \( N^1 \) outside the initial region. (Note that \( N^0 \cap N^1 \) could be empty as \( \varepsilon > 0 \) is sufficiently small). Now we recurrently perform the mapping so as to obtain the regions \( N^2, N^3, \ldots \). By Liouville’s Theorem, all these regions \( N^i \) have the same volume as that of \( N^0 \). Since the orbits are bounded, they are contained within a finite volume of the phase space. Thus, recurrent application of the the map \( T \) eventually overlaps a previously produced region. If this is not so, then we would eventually encounter the impossible situation of exceeding the volume of the bounded region. Now, by applying the inverse transformation \( T^{-1} \) to \( N^k \) for some \( k \in \mathbb{N} \) to obtain \( N^{k-1} \) repeatedly, we arrive at the original region \( N^0 \). Similar to non-Hamiltonian systems, Poincare sections provide useful information on the system evolution in the analysis of Hamiltonian systems.

**Remark 1.1.** In general, the above argument is not true for unbounded systems (e.g., \( T(x) = x + 1 \) on \( \mathbb{Z} \)) which are also measure preserving, but such a PPT does not recur. Furthermore, Poincare’s Recurrence Theorem may not be applicable to MPTs where the measure is not finite.

**Physical interpretation:** Given that a system satisfies the measure preserving condition, for almost every initial condition \( x \) of the system, it will return arbitrarily close to its initial state infinitely many times in arbitrarily far future. Note that it may not depend on the specifications of the system states or its laws of motion! This idea is also the heart of Statistical Mechanics that deals with general many-body systems without an explicit knowledge (e.g., dynamical equations) of the system model. In fact, the basic assumption made by Boltzman in his work on Statistical Mechanics is nothing but a quantitative version of the Poincare’s recurrence theorem, which also marks the birth of Ergodic theory. The assumption is known as the ”Ergodic Hypothesis”.

**Ergodic Hypothesis:** For certain invariant measures \( \mu \), many functions \( f : X \to \mathbb{R} \) and many states \( x \), the time average of \( f \), i.e., \( \lim_{T \to \infty} \frac{1}{T} \int_0^T f(T_t(x))dt \) exists in a specified sense, and this limit is equal to the ensemble average of \( f \), i.e., \( \frac{1}{\mu(X)} \int f d\mu \). Note that the division by \( \mu(X) \) in the last expression
normalizes the measure for compatibility with the probability measure. Let \( f \) be the indicator of the \( \epsilon \)-ball around a state \( x \) (e.g., a neighborhood of \( x \) with radius \( \epsilon > 0 \)). Then, the time average of \( f \) is the frequency of \( T_t(x) \)'s visits to the \( \epsilon \)-ball. This hypothesis leads to an idea of its value that is dependent on the size of the \( \epsilon \)-ball and the time taken by the dynamical system to enter the ball at least once.

**Ensemble Average and Time Average**

In general, one may use two ways to handle the analysis of a dynamical system. One way is to look at the time evolution of the map from a single initial condition \( x_0 \). The other way is to consider an ensemble of different initial conditions. Let \( \rho_0(x) \) be the probability density of initial conditions over the entire phase space. Therefore, the probability \( \mu_0(A) \) of having an initial condition in a subset \( A \) of phase space \( X \) can be expressed as

\[
\mu_0(A) = \int_A dx \rho_0(x)
\]

Given a map \( f \), let us analyze the time evolution of an ensemble of initial conditions. After \( n \) iterations \( \mu_n(A) \), the probability of finding an iterate \( x_n \) in \( A \) is

\[
\mu_n(A) = \int_A dx \rho_n(x)
\]

where \( \rho_n(x) \) is the corresponding probability density after \( n \) iterations.

Now, \( \mu_{n+1}(A) = \mu_n(f^{-1}(A)) \) implies that the probability of finding the \((n + 1)\)th iterate in \( A \) is the same as the probability of finding the \(n\)th iterate in the preimage of \( A \), i.e., \( f^{-1}(A) \). But as discussed earlier, we are especially interested in invariant probability distributions, i.e., \( \mu_{n+1}(A) = \mu_n(A) \). Such a measure is called an *invariant measure* and the corresponding density function is called an *invariant density*. For an invariant measure \( \mu \) and an invariant density \( \rho \), it follows that

\[
\mu(A) = \mu(f^{-1}(A)) \quad \text{and} \quad \int_A dx \rho(x) = \int_{f^{-1}(A)} dx \rho(x)
\]

Now let \( Q(x) \) be an arbitrary integrable function with respect to the invariant density \( \rho \). Then the *ensemble average* (over the entire phase space \( X \)) of \( Q \), denoted by \( \langle Q \rangle \) is expressed as

\[
\langle Q \rangle = \int_X dx \rho(x) Q(x)
\]
In an alternative notation, $\rho(x)dx$ is replaced by $d\mu(x)$. In this notation,

$$\langle Q \rangle = \int_X d\mu(x)\ Q(x) = \int_X d\mu(f^{-1}(x))\ Q(x) = \int_X d\mu(x)\ Q(f(x))$$

Note that in the last step, $x$ is replaced by $f(x)$. The significance of this observation is that the expectation of an observable remains invariant in an invariant ensemble under the action of the map $f$.

On the other hand, the time average of $Q$ is defined as

$$\overline{Q} = \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} Q(x_n)$$

provided that the limit converges.

The notions of ensemble average and time average are different, in general. When these two averages become equal in a given sense (e.g., mean square), then the map together with the invariant measure is called *ergodic* in that specific sense. The next section makes a formal introduction to the notions of ergodicity and mixing.

## 2 Ergodicity and Mixing

There are different notions of mixing but they are usually stronger than the notion of ergodicity. Let us start with basic concepts.

**Definition 2.1.** *(Invariant Set)* Let $(X, \mathcal{B}, \mu, T)$ be an MPT. A measurable set $E \in \mathcal{B}$ is called invariant (more precisely, $T$-invariant), if $T^{-1}(E) = E$ $\mu$-a.e., i.e., $\mu[T^{-1}(E)] = \mu[E]$. Therefore, $T$ can be split into two non-interacting MPTs as follows:

$$T|_E : E \to E \ \mu - a.e. \ \text{and} \ T|_{E^c} : E^c \to E^c \ \mu - a.e.$$  

**Definition 2.2.** *(Ergodicity)* An MPT $(X, \mathcal{B}, \mu, T)$ is called ergodic, if every invariant set $E$ satisfies the condition: either $\mu(E) = 0$ or $\mu(X \setminus E) = 0$ (obviously not both). In that case, $\mu$ is called an ergodic measure.

Let us explain the physical consequence of the above definition.

**Natural Invariant Measure:** A direct consequence of a map being ergodic is that an invariant measure of the map can be obtained by iterating (i.e., taking time evolution of) a single initial condition and producing a histogram; this is achieved by tracking the system evolution through different parts of
the phase space. For example, let $A$ be a measurable subset of the phase space $X$ and, given the ergodic map $T : X \to X$, the objective is to find an invariant measure of $A$ by using the ergodic property of the map. In this setting, the ensemble average is defined as
\[
\langle Q \rangle \triangleq \int_X dx \, \rho(x) \, Q(x)
\]
where $Q$ is an arbitrary integrable function with respect to the invariant density $\rho$. On the other hand, the time average is defined as
\[
\overline{Q} \triangleq \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} Q(x_n)
\]
Note that, in general, $\overline{Q}$ could be a random variable and $\langle Q \rangle$ could be time-dependent. However, if $\overline{Q} = \langle Q \rangle$ in a given sense (e.g., mean-square, or almost sure) and if $Q(x)$ is chosen as the characteristic function $\chi_A(x)$ of the set $A$, then it follows that
\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} \chi_A(x_n) = \int_X dx \, \rho(x) \, \chi_A(x) = \int_A dx \, \rho(x) = \mu(A)
\]
The fact of the matter is that there may exist several invariant measures for an ergodic map. However, the invariant measure $\mu$ identified from the above perspective is special and very useful in most cases due to its physical relevance. This measure is known as the natural invariant measure and the corresponding density function $\rho(x)$ is known as the natural invariant density. However, there are only a very few examples for which the natural invariant density can be determined analytically. Here is an example.

**Example 2.1.** *(Tent Map)* Let the unit interval $X = [0, 1]$ be the phase space. Consider the following one-dimensional map $f : X \to X$ defined as
\[
f(x) = \begin{cases} 
2x & \text{if } x \in [0, \frac{1}{2}) \\
2(1 - x) & \text{if } x \in [\frac{1}{2}, 1]
\end{cases}
\]
Referring to Fig. 1, total length of arbitrary intervals in the phase space remains invariant under the action of the map $f$; for example, the length segments satisfy the condition: $l_0 + l_1 = l_2$. Thus, in this case, $\rho(x) = 1$ is an invariant probability density function (in fact, the natural invariant density). However, in general, invariant density functions tend to be far more complicated than the uniform distribution.
Proposition 2.1. Let \((X, \mathcal{B}, \mu, T)\) be a measure-preserving transformation. Then, the following statements are equivalent.

1. The measure \(\mu\) is ergodic (see Definition 2.2).

2. If \(E \in \mathcal{B}\) and \(\mu(T^{-1}E \Delta E) = 0\), then \(\mu(E) = 0\) or \(\mu(X \setminus E) = 0\) (Note: \(\Delta\) is the symmetric difference operator, i.e., \(A \Delta B \triangleq (A \setminus B) \cup (B \setminus A)\));

3. If \(f : X \to \mathbb{R}\) is measurable and \(f \circ T = f\) \(\mu\)-a.e., then there exists \(c \in \mathbb{R}\) such that \(f = c\) \(\mu\)-a.e.

Proof. [1 \(\Rightarrow\) 2] Let \(\mu\) be ergodic and \(E\) be a measurable set such that \(\mu(T^{-1}E \Delta E) = 0\). Let us construct a measurable set \(E_0\) such that \(T^{-1}E_0 = E_0\) and \(\mu(E_0 \Delta E) = 0\). By ergodicity (see Definition 2.2), either \(\mu(E_0) = 0\) or \(\mu(X \setminus E_0) = 0\). Since \(\mu(E_0 \Delta E) = 0\) implies that \(\mu(E) = \mu(E_0)\) and \(\mu(X \setminus E) = \mu(X \setminus E_0)\), it follows that either \(\mu(E) = 0\) or \(\mu(X \setminus E) = 0\)

[2 \(\Rightarrow\) 3] Let us assume that if \(E \in \mathcal{B}\) and \(\mu(T^{-1}E \Delta E) = 0\), then \(\mu(E) = 0\) or \(\mu(X \setminus E) = 0\). Let \(f : X \to \mathbb{R}\) be measurable and \(f \circ T = f\) \(\mu\)-a.e. Then, for every \(t \in \mathbb{R}\),

\[
\{x \in X : f(x) > t\} \Delta T^{-1} \{x \in X : f(x) > t\} \subseteq \{x \in X : f(x) \neq f \circ T(x)\}
\]

Hence, \(\mu(\{x \in X : f(x) > t\} \Delta T^{-1} \{x \in X : f(x) > t\}) = 0\). In other words, either \(f > t\) \(\mu\)-a.e., or \(f \leq t\) \(\mu\)-a.e. defining \(c = \sup\{t \in X : f > t\ \text{a.e.}\}\), it follows that \(f = c\ \text{a.e.}\), which proves (3).

[3 \(\Rightarrow\) 1] It follows by choosing \(f\) to be the identity map. \(\square\)
Remark 2.1. Ergodicity implies measure-theoretic isomorphism, i.e., if two MPTs are isomorphic, then ergodicity of one MPT implies ergodicity of the other MPT.

Now we introduce the notion of strong mixing (also referred to as mixing). There are other notions of mixing (e.g., weak mixing) and, in general, the notions of mixing are stronger than that of ergodicity.

Definition 2.3. (Strong Mixing) A PPT \((X, B, \mu, T)\) is called strong mixing if for all \(E, F \in B\), \(\mu(E \cap T^{-k}F) \to \mu(E)\mu(F)\) as \(k \to \infty\).

Remark 2.2. Definition 2.3 implies that, given two measurable sets, if one set held fixed and the other set is allowed to evolve under \(T\) (which is a PPT), then they will eventually overlap. It is noted that there is no notion of strong mixing for general infinite measure spaces. Probabilistically speaking, \(T^{-k}(F)\) is “asymptotically independent” of \(E\) and physically it actually signifies uniform mixing of one substance into another (e.g., Rum and Coke!!).

Definition 2.4. (Weak Mixing) A PPT \((X, B, \mu, T)\) is called weak mixing (or mixing on the average) if, for all \(E, F \in B\),

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n} |\mu(E \cap T^{-k}F) - \mu(E)\mu(F)| = 0
\]

Remark 2.3. Strong mixing \(\Rightarrow\) weak mixing; and weak mixing \(\Rightarrow\) ergodicity. Therefore, strong mixing \(\Rightarrow\) ergodicity.

Proof. Let \((X, B, \mu, T)\) be a PPT and let \(E \in B\) be \(T\)-invariant, Then,

\[
\mu(E) = \mu(E \cap T^{-n}E) \to \mu(E)^2 \quad \text{as} \quad n \to \infty
\]

and this can only happen if either \(\mu(E) = 0\) or \(\mu(E) = 1 = \mu(X)\). This implies either \(\mu(E) = 0\) or \(\mu(X\setminus E) = 0\).

Definition 2.5. (Correlation and Covariance) The correlation coefficient of \(f, g \in L_2(\mu)\) is defined to be

\[
\rho(f, g) = \frac{\int fg d\mu - \int f d\mu \int g d\mu}{\| f - \int f d\mu \|_{L_2} \| g - \int g d\mu \|_{L_2}}
\]

where the numerator of the above equation is called the covariance

\[
\text{Cov}(f, g) = \int \left( (f - \int f d\mu) (g - \int g d\mu) \right) d\mu
\]
Remark 2.4. If $f$ and $g$ in Definition 2.5 are weakly correlated, then they are not likely to deviate consistently from their average values in the same direction, which may lead to lead to cancellations in the integral and thus yield a small value of $\rho(f, g)$. On the other hand, if $f$ and $g$ are strongly correlated, there will be fewer cancellations that would yield a larger value of $\rho(f, g)$. Note that the denominator in Definition 2.5 of $\rho$ is not important as it merely normalizes $\rho(f, g)$ to take values in the range $[-1, 1]$.

Proposition 2.2. A PPT $(X, \mathcal{B}, \mu, T)$ is strongly mixing if and only if

$$\int f \circ T^k d\mu \rightarrow \int f d\mu \int g d\mu \quad \forall f, g \in L_2(\mu) \text{ as } k \rightarrow \infty$$

Equivalently, $\text{Cov}(f, g \circ T^k) \rightarrow 0$ as $k \rightarrow \infty$.

Proof. Sketch of the proof; The condition, $\int f \circ T^k d\mu \rightarrow \int f d\mu \int g d\mu$ for all $f, g \in L_2(\mu)$, implies mixing. For example, let $f = 1_F$ and $g = 1_G$, where $F, G \in \mathcal{B}$.

To show the other way, we observe that

1. Given $\mu \circ T^{-1} = \mu$, we have $\| f \circ T \|_{L_2} = \| f \|_{L_2} \forall f \in L_2(\mu)$.
2. $\text{Cov}(f, g)$ is bilinear in $f, g$.
3. $|\text{Cov}(f, g)| \leq \| f \|_{L_2} \| g \|_{L_2}$ by using the Cauch-Schwarz inequality.

Example 2.2. (Circle Rotation) Let the following map $R_\alpha : [0, 1) \rightarrow [0, 1)$ be defined as

$$R_\alpha(x) = x + \alpha \mod 1$$

Let $\mathcal{B}$ be the Borel $\sigma$-algebra of $[0, 1)$ and let $\mu$ be the Lebesgue measure on $[0, 1)$. The map $R_\alpha$ is called circle rotation due to the following isomorphism $\Pi : [0, 1) \rightarrow S^1$ between $R_\alpha$ and rotation by an angle $2\pi \alpha$ as

$$\Pi(x) = e^{2\pi i x}$$

where $S^1 \triangleq \{(x_1, x_2) \in \mathbb{R}^2 : |x_1|^2 + |x_2|^2 = 1\}$ is the unit disk’s circumference.

The following facts are stated without a detailed proof.

1. $([0, 1), \mathcal{B}, \mu, R_\alpha)$ is a PPT.
2. $R_\alpha$ is ergodic if and only if $\alpha \notin \mathbb{Q}$, i.e., $\alpha$ is an irrational numbers.
3. $R_\alpha$ is never a mixing.

Sketch of the Proof:

1. For any interval $I \subseteq [0,1)$, the condition $\mu(R_\alpha^{-1}I) = \mu(I)$ holds. For a complete proof, it is necessary to show that $\mathcal{M} = \mathcal{B}$ (using the Monotone Class theorem), where

$$\mathcal{M} = \{E \in \mathcal{B} : \mu(R_\alpha^{-1}E) = \mu(E)\}$$

2. Fact: Let $(X, \mathcal{B}, \mu, T)$ be an MPT on a complex measure space, then “$\mu$ is ergodic” is equivalent to the following: If $f : X \to \mathbb{R}$ is measurable and $f \circ T = f$ a.e., then there is a constant $c \in \mathbb{R}$ s.t. $f = c$ a.e.

Let $f \circ R_\alpha = f$ a.e. in $L_2(\mu)$, i.e., the space of square ($\mu$-)integrable functions. Then, its inverse Fourier series expansion is

$$f = \sum_{n \in \mathbb{Z}} \hat{f}(n)e^{2\pi int} \quad \text{(convergence in } L_2)$$

Also,

$$f \circ R_\alpha = \sum_{n \in \mathbb{Z}} \hat{f}(n)e^{2\pi in\alpha}e^{2\pi int}$$

Equating coefficients between the two equations above, it follows that

$$\hat{f}(n) = \hat{f}(n)e^{2\pi in\alpha} \quad (n \in \mathbb{Z})$$

or, $\hat{f}(n)(1 - e^{2\pi in\alpha}) = 0$

Case 1: If $\alpha \notin \mathbb{Q}$, then $(1 - e^{2\pi in\alpha}) \neq 0 \forall n \neq 0$. This implies that $\hat{f}(n) = 0 \forall n \neq 0$. Therefore, $f = \hat{f}(0)$ is a constant a.e. Hence, $R_\alpha$ is ergodic.

Case 2: If $\alpha \in \mathbb{Q}$, then $\exists n_k \to \infty$ s.t. $e^{2\pi in_k\alpha} = 1$. Let us consider the following invariant continuous function, satisfying $f \circ R_\alpha = f$.

$$f(t) = \sum_{k=1}^{\infty} \frac{1}{2^n}e^{2\pi int}$$

Since $f$ is not a constant function, it is not ergodic.

3. This part requires the basic idea of Dirichlet theorem: $\forall \alpha$, $\exists$ a sequence $n_k \to \infty$ s.t. $((n_k \alpha) \mod 1) \to 0$ as $k \to \infty$. (It is clear for $\alpha \in \mathbb{Q}$, for $\alpha \notin \mathbb{Q}$, $(n_k \alpha)$ is even denser in $[0,1)$)

It follows that if $f(x)$ is 1-periodic on $\mathbb{R}$ and a non-constant, continuous function, then

$$f \circ R_\alpha^{n_k} \xrightarrow{k \to \infty} f \quad \text{pointwise}$$
as we have \( f(x + n_k \alpha \mod 1) \to f(x) \). Therefore,
\[
\int_0^1 f \cdot f \circ R_{\alpha}^{n_k} \, dx \xrightarrow{k \to \infty \text{ DCT}} \int_0^1 f^2 \, dx = (\int f)^2
\]

Hence, \( R_{\alpha} \) is never mixing. (the dominating function here is \( \sup |f|^2 \geq f \cdot f \circ R_{\alpha}^{n_k} \))

Next we present two important and rigorous results in ergodic theory. These results are stated below without a rigorous proof.

**Theorem 2.1.** *(Von Neumann’s Mean Ergodic theorem)* Suppose \((X, \mathcal{B}, \mu, T)\) is a PPT. If \( f \in L^2(\mu) \), then
\[
\frac{1}{n} \sum_{k=0}^{n-1} f \circ T^k \xrightarrow{n \to \infty} \overline{f}
\]
where \( \overline{f} \in L^2(\mu) \) is invariant. Moreover, if \( T \) is ergodic then \( \overline{f} = \int f \, d\mu \).

**Remark 2.5.** Invariance of \( \overline{f} \in L^2(\mu) \) means \( \| \overline{f} - f \circ T \|^2 = 0 \). However, if \( T \) is ergodic, then \( \overline{f} \) becomes a constant with value \( \int f \, d\mu \). As a matter of fact, \( \overline{f} \) is the projection of \( f \) on the space of invariant functions.

**Corollary 2.1.** A PPT \((X, \mathcal{B}, \mu, T)\) is ergodic iff \( \forall E, F \in \mathcal{B} \),
\[
\frac{1}{n} \sum_{k=0}^{n-1} \mu(E \cap T^{-k}F) \xrightarrow{n \to \infty} \mu(E)\mu(F)
\]
Therefore, “Ergodicity is Mixing on average”.

**Sketch of the proof:** If \( f_n \xrightarrow{L^2} f \), then
\[
\langle f_n, g \rangle \xrightarrow{n \to \infty} \langle f, g \rangle \ \forall g \in L^2
\]
Now let \( f_n = \frac{1}{n} \sum_{k=0}^{n-1} 1_B \circ T^k \) and \( g = 1_A \). Then, the proof of the corollary follows from the Mean Ergodic theorem.

**Theorem 2.2.** *(Birkhoff’s Pointwise Ergodic Theorem)* Let \((X, \mathcal{B}, \mu, T)\) be a PPT. If \( f \in L^1(\mu) \), then \( \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} f \circ T^k \) exists \( \mu \) a.e. The limit is a \( T \)-invariant function and if \( T \) is ergodic, it is equal to \( \int f \, d\mu \).

**Remark 2.6.** Birkhoff’s Pointwise Ergodic Theorem is also known as the strong ergodic theorem due to the strong convergence property. Direct consequences of the ergodic theorems are observed as follows.
1. The time spent in a measurable subset $A$ (i.e., sojourn time) of the phase space $X$ of an ergodic system is proportional to its measure, i.e.,

$$
\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} \chi_A(T^k x) = \frac{\mu(A)}{\mu(X)} \text{ for almost all } x \in X
$$

where $\chi_A$ is the indicator function of $A$.

2. (Quantification of Poincare’s recurrence theorem) Let us consider the recurrence time of a measurable subset $A$. Let $x \in A \subseteq X$ be an initial condition and let $k_0 < k_1 < k_2 < \cdots < k_i \cdots$ (with $k_0 = 0$) be a strictly increasing sequence such that $T^{k_i}(x) \in A$. The recurrence time, i.e., the time required by the system to return to $A$, is defined as $R_i = k_i - k_{i-1}$. With this definition, it follows from the ergodic theorems that

$$
\frac{R_1 + R_2 + \cdots + R_n}{n} \xrightarrow{a.s.} \frac{\mu(X)}{\mu(A)}
$$

That is, the smaller the set $A$ is, the longer it takes for the system to revisit $A$.

3 **Symbolic Dynamics and Coding**

The discussions on Ergodic theory are primarily suitable for discrete-state systems. Therefore, we need an analytical method to convert real physical systems with continuous states to corresponding discrete-state representations. Such an analytical method is known as **symbolic dynamics** that involves coarse-graining of the phase space as it is often referred to in the physics literature. This coarse graining process is crucial for correct representation of the dynamical system and is system-specific in order to retain the salient characteristics of the dynamical system. A brief introduction to Symbolic Dynamics is presented below.

**Definition 3.1.** (Partitioning) The phase space $X$ of a dynamical system is partitioned into a set of finitely many cells $A_j$, $j \in \{1, 2, \cdots, R\}$ such that these cells are mutually exclusive and exhaustive, i.e.,

$$
A_j \cap A_k = \emptyset \quad \forall j \neq k \quad \text{and} \quad \bigcup_{j=1}^{R} A_j = X
$$

where each of these cells $A_j$ is labeled as a distinct symbol and the set of all these symbols is called the alphabet $\Sigma$ such that the cardinality of the alphabet is $|\Sigma| = R$. 


Let the dynamical system with map \( f \) have an initial condition \( x_0 \) and, as a consequence of the coarse graining, \( x_0 \) belongs to a specific cell that is labeled by the symbol \( i_0 \in \Sigma \). With the evolution of the dynamical system, the next point \( x_1 \) on the trajectory belongs to a cell that is labeled as \( i_1 \in \Sigma \). Similarly, the cell containing the next point \( x_2 \) is labeled as \( i_2 \in \Sigma \) and so on. Thus, for each initial condition \( x_0 \), there exists a sequence of symbols, in which each symbol belongs to the alphabet \( \Sigma \),

\[
x_0 \rightarrow i_0 \ i_1 \ i_2 \ \cdots \ i_n \ \cdots
\]

that is called a symbol sequence as it describes the trajectory of the corresponding dynamical system in a coarse-grained fashion. This mapping from the continuous phase space to the discrete symbol space is called symbolic dynamics. Although the size of the cell indexed by \( i_0 \) is finite, it can accommodate (uncountably) infinitely many initial conditions. Similarly, a given symbol sequence of finite length may represent many different (continuous) phase evolutions of the system. However, given a dynamical system, there could be symbol sequences that can never occur. Such sequences are called forbidden.

It is understood that even if a symbol sequence is infinitely long, the existence and uniqueness of an initial condition \( x_0 \) is not guaranteed. That is, the symbol sequence may be forbidden (i.e., an empty cylinder) or it may correspond to multiple system trajectories. A partitioning that provides one-to-one correspondence between the set of all (infinitely long) symbol sequences to the set of all system trajectories is of interest.

**Definition 3.2.** (Generating Partition) A fixed and finite partitioning for a given map is called a generating partition, if an infinite symbol sequence \( i_0, i_1, i_2, \cdots \) uniquely determines the initial condition \( x_0 \). In other words, the symbolic dynamics mapping is bijective, i.e., not only the initial condition \( x_0 \) determines the symbol sequence, but the reverse is true as well.

In general, whether there exists a generating partition for a given map or not is a very tough question to answer. Apparently, it is possible to construct a tractable generating partition for a few simple maps. At this point, we make use of the concept of cylinders to be able to impose measures on symbol sequences for analysis of dynamical systems.

**Definition 3.3.** (Cylinders) For a finite symbol sequence \( i_0, i_1, i_2, \cdots, i_{N-1} \) of length \( N \), the set of all initial conditions that generate the symbol sequence
is called an $N$-cylinder and is denoted as $J[i_0, i_1, i_2, \cdots, i_{N-1}]$. If the sequence $i_0, i_1, i_2, \cdots, i_{N-1}$ is forbidden, then the $N$-cylinder $J[i_0, i_1, i_2, \cdots, i_{N-1}] = \emptyset$.

Note that the cylinders induce another partition on the phase space $X$ because of the following condition.

$$\bigcup J[i_0, i_1, i_2, \cdots, i_{N-1}] = X$$

where the union is taken over all possible symbol sequences of length $N$; and the above fact is true as each point in the phase space $X$ can be an initial condition of a specific trajectory. Furthermore,

$$J[i_0, i_1, i_2, \cdots, i_{N-1}] \cap J[j_0, j_1, j_2, \cdots, j_{N-1}] = \emptyset$$

if $i_n \neq j_n$ for at least one $n \in \{0, 1, 2, \cdots, N-1\}$, because different symbol sequences must correspond to different initial conditions.

The rationale for imposing a measure on a symbol sequence is to quantify the frequency of occurrence of that sequence. The objectives here are to understand the nature of forbidden sequences and to quantify the frequency of occurrence of allowable sequences from the map characteristics of the dynamical system and the partitioning. Let us denote the probability of the sequence $i_0, i_1, i_2, \cdots, i_{N-1}$ as $p(i_0, i_1, i_2, \cdots, i_{N-1})$ and consider an ensemble of initial conditions with probability density $\rho_0$. Then, it follows that

$$p(i_0, i_1, i_2, \cdots, i_N) = \mu(J[i_0, i_1, i_2, \cdots, i_N])$$

where $\mu$ is the natural invariant measure. Furthermore,

$$p(i_0, i_1, i_2, \cdots, i_N) = p(i_N|i_0, i_1, i_2, \cdots, i_{N-1})p(i_0, i_1, i_2, \cdots, i_{N-1})$$

**Definition 3.4 (Markov Chain).** The stochastic process represented by the symbol sequence $i_0, i_1, i_2, \cdots, i_{(N-1)}$ is called a Markov chain if

$$p(i_N|i_0, i_1, i_2, \cdots, i_{N-1}) = p(i_N|i_{N-1})$$

(13)

Definition 3.4 implies that the conditional probability depends only on the last event and not on the entire chain of events.

Another related concept in the analysis of dynamical systems is that of a topological Markov chain as defined below.

**Definition 3.5 (Topological Markov Chain).** The stochastic process represented by the symbol sequence $i_0, i_1, i_2, \cdots, i_{(N-1)}$ is called a topological Markov chain when $p(i_N|i_0, i_1, i_2, \cdots, i_{N-1}) = 0$ iff one of the following two conditions hold:

$$p(i_N|i_{N-1}) = 0 \text{ or } p(i_{N-1}|i_0, i_1, i_2, \cdots, i_{N-2}) = 0$$
Definition 3.5 implies that there are only two ways to have the zero probability of transition to a state – either the probability of transition to the current state from the previous state is zero, or the probability of occurrence of the previous state is zero.

If the stochastic process generated by a map \( f \) is neither a Markov chain nor a topological Markov chain, there is a way to enforce (expanding) maps to be converted to a topological Markov chain by introducing the concept of Markov partition.

**Definition 3.6.** *(Markov Partition)* Given a map, a partitioning is called Markov partition if it converts the (continuous) system evolution to a topological Markov chain.

The following example is presented to clarify the concepts introduced above.

**Example 3.1.** *(Angle Doubling)* Consider the map \( T(x) = 2x \mod 1 \) on the interval \([0, 1]\) equipped with the Lebesgue measure. This map is called the angle doubling map due to the following isomorphism \( \Pi \) between \( T \) and the map \( e^{i\theta} \rightarrow e^{2i\theta} \).

\[
\Pi(x) = e^{2\pi ix}
\]

Alternatively, the angle doubling map can be written as

\[
T(x) = \begin{cases} 
2x & \text{for } x \in [0, \frac{1}{2}) \\ 
2x - 1 & \text{for } x \in [\frac{1}{2}, 1] 
\end{cases}
\]

Recalling the formula for binary to decimal conversion in \([0, 1)\), let a fraction have the binary representation as \( 0.i_1i_2i_3\cdots i_n\cdots \), where \( i_n \in \{0, 1\} \). Then, the corresponding decimal representation would be \( i_1 \times 2^{-1} + i_2 \times 2^{-2} + i_3 \times 2^{-3} + \cdots + i_n \times 2^{-n} + \cdots \). For example, 0.001 is the binary representation of the fraction \( \frac{1}{8} \) in decimal notation. With the binary notation in the current phase space, the angle doubling map has the following property:

\[
T(0.i_1i_2i_3\cdots i_n\cdots) = 0.2i_1i_2i_3\cdots i_n\cdots
\]

i.e., the mapping \( T \) imposes a left shift operation on the sequence \( i_1, i_2, i_3, \ldots, i_n, \ldots \). In general such an operation is called a Bernoulli Shift. This also allows the initial condition \( x_0 \) to be expressed as:

\[
x_0 = \sum_{k=1}^{\infty} i_k 2^{-k}
\]
Furthermore, the $n^{th}$ iterate $x_n$ is expressed as:

$$x_n = \sum_{k=1}^{\infty} i_{k+n} 2^{-k}$$  \hspace{1cm} (17)

In this way, the first digit is removed from the sequence and the rest is shifted to the left. Thus, $T$ is also known as the binary shift map.

The binary formulation leads to a useful partition of the phase space consisting of two intervals, $A_0 = [0, \frac{1}{2})$ (representing the symbol 0) and $A_2 = [\frac{1}{2}, 1)$ (representing the symbol 1). In this framework, the symbol sequence $i_1, i_2, i_3, \ldots, i_n, \ldots$ provides a coarse-grained description of the trajectory. $x_n$ is in $A_0$ for $i_{n+1} = 0$ and $x_n$ is in $A_1$ for $i_{n+1} = 1$. From the expression of $x_0$ in this partition, the initial condition can be determined more and more precisely as the length of the symbol sequence increases and it can be uniquely determined with an infinite length sequence. Thus, the partition $\{A_0, A_1\}$ qualifies as a generating partition.

The following properties of the angle doubling map $T$ are presented below.

1. $T$ is a probability preserving transformation (PPT).
2. $T$ is strongly mixing and hence ergodic.

**Sketch of the proof:** For every symbol sequence $(i_1, i_2, \ldots, i_n)$ of finite length $n$, the cylinders are defined as:

$$[\hat{i}] \equiv [i_1, i_2, \ldots, i_n] \triangleq \{x \in [0, 1) : x = 0.i_1i_2\ldots i_n\varepsilon_1\varepsilon_2\ldots \text{ for some } \varepsilon_k \in \{0, 1\}\}$$

This is known as a dyadic interval with dyadic length $\mu([\hat{i}]) \triangleq \frac{1}{2^n}$ (probabilistic measure). Furthermore,

$$T^{-1}[i_1, i_2, \ldots, i_n] = [* , i_1, i_2, \ldots, i_n]$$

where ‘*’ stands for either ‘0’ or ‘1’.

1. It follows from the above analysis that

$$\mu(T^{-1}[\hat{i}]) = \mu(0, \hat{i}) + \mu(1, \hat{i}) = 2 \cdot \frac{1}{2^{n+1}} = \frac{1}{2^n} = \mu([\hat{i}])$$

Let the measurable space $\mathcal{M} \triangleq \{E \in \mathcal{B} : \mu(T^{-1}E) = \mu(E)\}$. By the Monotone Class theorem, it can be shown that $\mathcal{M} = \mathcal{B}$. 

20
2. Considering two symbol sequences, \([a] = [a_1, \cdots, a_\alpha]\) and \([b] = [b_1, \cdots, b_\beta]\), it follows that, for \(n > \alpha\)
\[ [a] \cap T^{-n}[b] = [a_1, \cdots, a_\alpha, *, \cdots, *, b_1, \cdots, b_\beta] \]
where the length of the sequence \(*, \cdots, *\) is \((n - \alpha)\). Hence,
\[ \mu([a] \cap T^{-n}[b]) = 2^{n-\alpha} \times \frac{1}{2^{\alpha+(n-\alpha)+\beta}} = \frac{1}{2^{\alpha+\beta}} = \mu([a])\mu([b]) \]

Through usage of indicator functions, it can be shown that the above expression is true for \(\forall \ E, F \in \mathcal{B}\).

In the above example, there are no forbidden sequences at all. Therefore, the condition for being a topological Markov chain is trivially fulfilled and hence, the current partition is also a Markov partition. Moreover, if we choose the natural invariant measure for the initial ensemble, we obtain even a conventional Markov chain. As the natural invariant density \(\rho(x)\) is uniform (= 1) and as each cylinder has equal size, we have \(\mu([i]) = \frac{1}{2^n} \forall n\). Hence,
\[ \mu(i_n|i_1, i_2, \cdots, i_{n-1}) = \frac{1}{2} = \mu(i_n|i_{n-1}) \]
(18)

The treatment above using measures on cylinders leads to the ergodic theory of finite Markov chains which is the next topic.

4 Ergodic Theory of Finite Markov Chains

This section addresses the ergodic properties of dynamical systems when they are modeled as Markov chains. Preliminary concepts are introduced and pertinent definitions are stated before embarking on the core concept of ergodicity and Markov properties.

**Definition 4.1 (Shift of Finite Type).** Let \(Q\) be a (nonempty) finite set of states, whose cardinality is \(|Q| \in \mathbb{N}\). Let \(A = (a_{ij})_{|S| \times |S|}\) be a matrix of zeros and ones without any zero row and zero column. The (one-sided) shift of finite type (SFT) associated with the set \(Q\) and the transition matrix \(A\) is defined as the set
\[ \Sigma_A^+ = \{(q_0, q_1, \cdots) \in S^\mathbb{N} : a_{q_iq_{i+1}} = 1 \ \forall i\} \]
(19)
that is equipped with the topology generated by the cylinders \([q_0, \cdots, q_{n-1}] \triangleq \{\underline{x} \in \Sigma_A^+ : x_i = q_i \text{ for } i = 0, \cdots, n-1\}\) and equipped with the left shift operation: \(\sigma(q_0, q_1, q_2, \cdots) = (q_1, q_2, \cdots)\).
Definition 4.1 implies that $\Sigma^+_A$ is the space of all infinite paths on a directed graph with vertices in $Q$ and edges $q \to \tilde{q}$ connecting vertices $q, \tilde{q} \in Q$ s.t. $a_{q\tilde{q}} = 1$. The following facts are stated in this context,

1. A stochastic matrix $P$ is a (square) array of non-negative real numbers such that $P = (P_{q\tilde{q}})_{q,\tilde{q}\in Q}$, where $P_{q\tilde{q}}$ is the transition probability from state $a$ to state $b$, which must satisfy the condition: $\sum_{q\in Q} P_{q\tilde{q}} = 1 \forall q \in Q$. That is, $P \mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ is the column vector with each entry equal to 1.

2. A probability vector is a (row) vector $\underline{p} = \{p_a \in [0,1] : a \in S\}$ and $\sum_a p_a = 1$. 

3. A stationary probability vector is a probability vector $\underline{p} = \{p_a : a \in S\}$ s.t. $\underline{p} P = \underline{p}$, i.e., $\sum_a p_a P_{ab} = p_b$

**Definition 4.2.** (Markov chain) A Markov chain is defined as the pair $(P, p)$, where the stochastic matrix $P$ is called the state transition matrix and the row vector is called the state probability vector. The Markov chain $(P, p)$ is said to be compatible with the SFT having the transition matrix $A = (A_{ab})_{|S|\times|S|}$ if $A_{ab} = 0$ implies $p_{ab} = 0$. In that case, a Markov measure ($\mu$) of the cylinder $[a_0, \cdots, a_{n-1}]$ is defined on $\Sigma^+_A$ as  

$$\mu[a_0, \cdots, a_{n-1}] \triangleq p_{a_0} p_{a_0 a_1} \cdots p_{a_{n-2} a_{n-1}}$$

Definition 4.2 is interpreted for operation of a Markov chain as presented below.

1. The probability of starting at a state $a_0 \in S$ is $p_{a_0}$.

2. The probability of transition from a state $a \in S$ to a state $b \in S$ is $p_{ab}$.

   (Note: It is possible that the states $a$ and $b$ are the same.)

3. Markov property ensures the independence of each transition.

**Proposition 4.1.** A measure $\mu$ is shift invariant if and only if the probability vector $\underline{p}$ is stationary with respect to $P$. Any stochastic matrix has a stationary probability vector.

**Proof.** $\mu$ is shift-invariant (stationary) iff $\mu[*]\underline{b} = \mu\underline{b} \forall \underline{b}$ i.e., 

$$\sum_a p_a p_{a_0 b_0} p_{b_0 b_1} \cdots p_{b_{n-2} b_{n-1}} = p_{b_0} p_{b_0 b_1} \cdots p_{b_{n-2} b_{n-1}}$$
Canceling the identical terms on both sides, it follows that \( \sum_a p_a p_{ab_0} = p_{b_0} \) for the stationary probability vector \( p \). The second part of the proposition is known as the Perron-Frobenius theorem that is stated without proof at the end of this section.

**Terminologies:** Let \( \Sigma^+_A \) be an SFT with \( A = \{t_{ij}\} \). Then, a single transition \( q \to \tilde{q} \) takes place if \( t_{\tilde{q}q} = 1 \) and \( n \) transitions \( a \xrightarrow{n} b \) take place if there exists states \( q_1, \ldots, q_{n-1} \) such that \( a \to q_1, q_1 \to q_2, \ldots, q_{n-1} \to b \). With these notations we introduce the following important concepts.

**Definition 4.3.** (Irreducibility) An SFT \( \Sigma^+_A \) is irreducible if \( \forall a, b, \exists n \in \mathbb{N} \) such that \( a \xrightarrow{n} b \).

**Remark 4.1.** Let the Markov chain \( (P, p) \) be compatible with an SFT \( \Sigma^+_A \). The Markov chain is called irreducible if there exists an \( n \in \mathbb{N} \) such a given entry of the \( n^{th} \) power of \( P \) is strictly positive. This positive integer \( n \) possibly depends on the indices of the given matrix entry. Furthermore, for an irreducible Markov chain, each element of the state probability vector \( p \) is strictly positive. This ensures that each state is recurrent and this property is equivalent to ergodicity. That is why a Markov chain that has an irreducible stochastic matrix is called ergodic.

**Example 4.1.** The Markov chain in Fig. 2 is not irreducible because \( n \in \mathbb{N} \) such that \( 2 \xrightarrow{n} 1 \), i.e., there are no finite number of transitions that can take the state 2 to the state 1. Now, let \( E = \{2, 3\} \), then \( E \) is an invariant set, but \( \mu(E) \neq 0, 1 \). Hence the Markov chain cannot be ergodic.

**Definition 4.4.** (Aperiodicity) An SFT \( \Sigma^+_A \) is aperiodic if \( \forall a \in S, \gcd\{n|a \xrightarrow{n} a\} = 1 \), and \( \gcd\{n|a \xrightarrow{n} a\} \) is known as the period of the SFT.

**Remark 4.2.** If \( \Sigma^+_A \) is irreducible, then \( \gcd\{n|a \xrightarrow{n} a\} \) is independent of \( a \) and is called the period of \( \Sigma^+_A \).
Example 4.2. The Markov chain in Fig. 3 is irreducible and it is observed that a non-trivial invariant set contains all the states. Hence, it is ergodic. But it is periodic of period 2 because a typical sequence is $0 \ast 30 \ast 30 \ast 30 \ast \cdots$. Hence, it follows that

$$\mu([0] \cap T^{-n}[0]) = 0 \text{ if } 3 \nmid n$$

But $\mu([0]) \neq 0$ which implies that on a subsequence

$$\mu([0] \cap T^{-n}[0]) \not\to \mu[0]^2$$

Hence, it cannot be mixing.

As a matter of fact, irreducibility and aperiodicity are the only two crucial aspects when it comes to ergodicity and mixing properties of a Markov chain.

Theorem 4.1. (Ergodic theorem of Markov Chains) Suppose $P$ is a stochastic matrix and we write $P^n = (p_{ab}^{(n)})$, then $P$ has a stationary probability vector $\underline{p}$, and (for $a, b \in S$)

- If $P$ is irreducible, then $\frac{1}{n} \sum_{k=1}^{n} p_{ab}^{(k)} \to p_b$ as $n \to \infty$
- If $P$ is irreducible and aperiodic, then $p_{ab}^{(k)} \to p_b$ as $k \to \infty$

Corollary 4.1. The measure of a shift-invariant Markov chain on an SFT $\Sigma_A^+$ is ergodic if and only if $A$ is irreducible and it is mixing if and only $A$ is irreducible and aperiodic.

Remark 4.3. Let the Markov chain $(P, p)$ be compatible with an SFT $\Sigma_A^+$. The Markov chain is called regular (i.e., irreducible and aperiodic) if there exists an $n \in \mathbb{N}$ such that all entries of the $n^{th}$ power of $P$ are strictly positive. Note that this property is stronger than irreducibility alone where the $n$ could be different for matrix entries. This ensures that after a sufficiently large number of iterations, each state is accessible from all states (including itself).
in a single transition. Such stochastic matrices are often called primitive and this property is equivalent to mixing. Also note that an irreducible stochastic matrix becomes primitive if its trace is strictly positive, i.e., if any one of the diagonal elements is strictly positive.

**Theorem 4.2** (Perron-Frobenius theorem restricted to ergodic Markov chains). Let an ergodic Markov chain with \( n \) states be represented as \((P, p)\), i.e., \( P \) is an \((n \times n)\) irreducible stochastic matrix. Then, the following conditions hold.

- **Condition (i):** \( p P = p \), i.e., one of the eigenvalues of \( P \) is \( 1 \), and the corresponding left eigenvector \( p > 0 \), i.e., each entry of the \((1 \times n)\) vector \( p \) is strictly positive.

- **Condition (ii):** The eigenvalue \( 1 \) has algebraic multiplicity \( = 1 \) and consequently its geometric multiplicity \( = 1 \). All other eigenvalues \( \lambda_i \) of \( P \) are bounded above as: \( |\lambda_i| \leq 1 \), i.e., they must be contained in the complex plane on or within on the unit circle with its center at the origin.

- **Condition (iii):** The eigenvalues \( \lambda_i, i = 1, 2, \cdots, k - 1 \) of \( P \) such that \( |\lambda_i| = 1 \) are solutions of the equation \( \lambda^k - 1 = 0 \), i.e., the eigenvalues that are not equal to \( 1 \) must appear as \(-1\) and/or as complex conjugate pairs on the unit circle with its center at the origin.

**Corollary 4.2** (Perron-Frobenius theorem restricted to regular Markov chains). Let a regular Markov chain with \( n \) states be represented as \((P, p)\), i.e., \( P \) is an \((n \times n)\) primitive (i.e., irreducible and aperiodic) stochastic matrix. Then, the following conditions hold.

- **Condition (i):** \( p P = p \), i.e., one of the eigenvalues of \( P \) is \( 1 \), and the corresponding left eigenvector \( p > 0 \), i.e., each entry of the \((1 \times n)\) vector \( p \) is strictly positive.

- **Condition (ii):** The eigenvalue \( 1 \) has algebraic multiplicity \( = 1 \) and consequently its geometric multiplicity \( = 1 \). All other eigenvalues \( \lambda_i \) of \( P \) are bounded above as: \( |\lambda_i| < 1 \), i.e., these eigenvalues must be contained strictly within the unit circle with its center at the origin.

5 **Fermi-Pasta-Ulam (FPU) Problem**

Consider an isolated Hamiltonian system, if the map/flow is fully ergodic and all states are equally probable, then we have an equipartition of energy, i.e.,
the energy of the system is divided equally among the $N$ degrees of freedom.

More formally,

**Equipartition theorem**: In general (both in microcanonical ensemble with total energy of the system being constant, or in canonical ensemble with constant system temperature) for a physical system with Hamiltonian $H$ and degrees of freedom $x_n$, the following equipartition formula holds in *thermal equilibrium* $\forall \ m, n \in \mathbb{N}$.

$$\langle x_m \frac{\partial H}{\partial x_n} \rangle = \delta_{mn} k_B T$$

where

$$\delta_{mn} = \begin{cases} 
0 & \text{if } m \neq n \\
1 & \text{if } m = n 
\end{cases}$$

The expectation is an average over the phase space (i.e., ensemble average) or under the *ergodicity assumption*, the time average of a single entity. The dynamic process of reaching such equilibrium condition is known as *thermalization*.

The Fermi-Pasta-Ulam (FPU) problem was to investigate this thermalization process from a temporal perspective (hence, examining the validity of the ergodicity assumption) using a simple spring-mass system (and also to test Fermi’s idea of solving analytically intractable dynamical problems by a computer). The work was done during early 1950s in the Los Alamos National Laboratory using the MANIAC (Mathematical Analyzer, Numerical Integrator and Computer) system that was built in 1952 for the Manhattan project and was used to develop the first hydrogen bomb “Mike”. The system is illustrated in Fig. 4.

(For more information, please read: “The Fermi-Pasta-Ulam Numerical Experiment”: History and pedagogical perspectives by Dauxois et. al., Eur. J. Phys., 26 (2005), S3-S11)
Let $H$ be the Hamiltonian defined as

$$H \triangleq \sum_{i=0}^{N} \frac{p_i^2}{2} + \sum_{i=0}^{N} \frac{1}{2} (u_{i+1} - u_i)^2 + \frac{\alpha}{3} \sum_{i=0}^{N} (u_{i+1} - u_i)^3$$

where $u$ denotes displacement, $p$ denotes momentum, $\alpha$ denotes the degree of nonlinearity, $N$ denotes the number of mass entities and $i$ stands for mass number, and mass/stiffness is 1 without any loss of generality.

If the parameter $\alpha = 0$, then the system is linear and all the normal modes become independent, i.e., if the total energy is introduced into say mode 1, none of the other modes will be excited. Therefore, No Thermalization!! But what happens if $\alpha \neq 0$, that was the question. Will the energy introduced into a single mode become distributed among other modes leading to an equipartition of energy in an asymptotic sense?

From modal analysis perspective: Amplitude $A_k$ and frequency $\omega_k$ of mode $k$ are calculated as

$$A_k = \sqrt{\frac{2}{N+1}} \sum_{i=1}^{N} u_i \sin \left( \frac{ik\pi}{N+1} \right)$$

$$\omega_k^2 = 4 \sin^2 \left( \frac{k\pi}{2(N+1)} \right)$$

Using these definitions, $H$ is rewritten as

$$H = \frac{1}{2} \sum_{k=1}^{N} (\dot{A}_k^2 + \omega_k^2 A_k^2) + \frac{\alpha}{3} \sum_{k,l,m=1}^{N} C_{k,l,m} A_k A_l A_m \omega_k \omega_l \omega_m$$

(See Scholl, D.S., Phys. Lett. A, 149, 1990 for details on the above formulation and coefficients $C_{k,l,m}$).

The second term induces coupling between modes. But does it lead to thermalization? In fact, in the beginning of the experiment, it seemed like energy is being distributed among modes. However, when the simulation was run longer, accidentally it was found that the system actually comes back arbitrarily close to its initial condition with a finite period! The approximate recurrence time, $T_R \simeq 0.38 \frac{N^{5/2}}{\sqrt{aa}}$; where, $a$ is the amplitude of the sinusoidal initial condition. (Note, this is not Poincare’s recurrence as the recurrence time is much quicker and the scaling law with system size is different).

Next we present a resolution of the paradox.

1. **Generation of Solitons**: A sinusoidal initial excitation, with a high value
of the parameter $\alpha$, generates solitons (i.e., a series of pulses) and conservation of shapes and velocities of those pulses, which leads to the periodicity.

2. **Overlap of nonlinear resonances (Chirikov):** The Kolmogorov-Arnold-Moser (KAM) theorem states that, for most orbits with a slight perturbation in the parameter $\alpha$, the Hamiltonian remains quasi-periodic provided that they are integrable. In the same line, when the nonlinearity is small, chaotic orbits have smaller Lyapunov exponents that do not diffuse in phase space (minimal overlap of nonlinear resonance). Beyond the **strong stochastic threshold**, if a weak chaos turns into a strong chaos, thermalization may occur.

Please see *The Ergodic Hypothesis: A Complicated Problem in Mathematics and Physics* by Adrian Patrascioiu, Los Alamos Science Special Issue 1987 for the list of rigorous results obtained so far about the circumstances under which a system is fully ergodic.

## 6 Ergodicity and Chaos

The notion of chaos and specifically deterministic chaos in dynamical systems is closely related to mixing and ergodicity. In fact, the theory of ergodicity is used extensively to understand chaotic systems. Primarily, the main features of chaotic dynamical systems are: (i) sensitivity to initial conditions, (ii) topological mixing, and (iii) existence of dense periodic orbits.

**Definition 6.1.** (Topological Transitivity): Let $X$ be a topological space. A continuous map $f : X \to X$ is called topologically transitive if, for every pair of open sets $U, V \subseteq X$, $\exists n \in \mathbb{N}$ s.t. $f^n(U) \cap V \neq \emptyset$

**Definition 6.2.** (Topological Mixing): Let $X$ be a topological space. A continuous map $f : X \to X$ is called topological mixing if, for every pair of open sets $U, V \subseteq X$, $\exists N \in \mathbb{N}$ s.t. $f^n(U) \cap V \neq \emptyset \forall n > N$.

**Examples:**

- **Ergodic and chaotic:** The angle doubling map has already been proven to be an ergodic one. Now, considering the binary formulation, it is intuitively understood that, with a slight perturbation in the initial condition, the corresponding symbol sequences diverge after a finite length. This is a simple example of an ergodic and chaotic map.
• **Ergodic but not chaotic:** The irrational rotation map has also been shown to be ergodic. However, with slight perturbation in the initial condition, the perturbation does not increase as the systems evolve. However, it has been established that this map cannot have dense periodic orbits. Thus, the map is ergodic but not chaotic.

• **Not ergodic but chaotic:** This notion is not yet understood properly, see Broken ergodicity in classically chaotic spin systems by Borgonovi et. al., J. Stat. Phys., 2005 for a constructive example.

For ergodic theoretic approach to chaos, please read Ergodic theory of chaos and strange attractors by Eckmann & Ruelle, 1985.

**APPENDIX: Markov Processes**

Markov processes are naturally associated with the state-space modeling of dynamical systems, where the state of the system at any given time $t$ contains all information about the system up to and including the time $t$. The concept of a state is captured in stochastic processes by using the notions of conditional independence and Markov property.

Let us further elucidate the notion of conditional independence. Recall that, for an event $F$ in a probability space $(\Omega, \mathcal{E}, P)$, $P[F] = E[\chi_F]$, where $\chi_F$ is the indicator function of $F$, i.e., $\forall \zeta \in \Omega$,

$$
\chi_F(\zeta) = \begin{cases} 
1 & \text{if } \zeta \in F \\
0 & \text{if } \zeta \notin F 
\end{cases}
$$

Furthermore, let $Y$ be a random vector such that

- $P[F|Y] \overset{as}{=} E[\chi_F|Y]$ is a measurable function of $Y$, which has a finite second moment.

- $E[g(Y)P[F|Y]] = E[g(Y)\chi_F]$ for a measurable function $g(\bullet)$ with a finite second moment.

**Definition 6.3.** (Conditional Independence) Let $A$, $B$, and $C$ be Borel sets. Then, the events $X \in A$ and $Z \in C$ are said to be conditionally independent of an event $Y \in B$, which is denoted as $X \bowtie Y \bowtie Z$, if

$$
P[X, Z|Y] = P[X|Y] \cdot P[Z|Y]
$$

**Remark 6.1.** The statement $X \bowtie Y \bowtie Z$ implies truth of the following two statements:


**Definition 6.4. (Markov Process)** A stochastic Process \( \{X_t : t \in T \} \) is defined to be Markov if any one of the two following (equivalent) conditions hold for any \( t_1, \cdots, t_{n+m} \in T \) with \( t_1 < \cdots < t_{n+m} \) and \( n, m \in \mathbb{N} \):

1. The conditional independence: \( (X_{t_1}, \cdots, X_{t_n}) \indep X_{t_n} \indep X_{t_{n+1}} \)

2. The conditional independence: \( (X_{t_1}, \cdots, X_{t_n}) \indep X_{t_n} \indep (X_{t_n}, \cdots, X_{t_{n+m}}) \)

**Remark 6.2.** In Definition 6.4, the first condition is easier to check than the second condition but the second condition is more appealing, because it is symmetric in time. For example, letting \( t_n \) to be the current time, the Markov property implies that the past and future are conditionally independent given the present state. Equivalent to Definition 6.4, a Markov process satisfies the following definition:

A stochastic process \( X_t \) is called Markov if, for every \( t_{k-1} < t_k \),

\[
P[X_{t_k} | X_t, \forall t \leq t_{k-1}] = P[X_{t_k} | X_{t_{k-1}}]
\]

In other words, if \( t_1 < t_2 < \cdots < t_{k-1} < t_k \), then

\[
P[X_{t_k} | X_{t_{k-1}}, X_{t_{k-2}}, \cdots, X_{t_1}] = P[X_{t_k} | X_{t_{k-1}}]
\]

In essence, the past has no bearing on the future if the present is known.

**Remark 6.3.** A Markov process \( X_t \) has the following properties:

- Retention of Markov property under time reversal: If \( t_k < t_{k+1} < \cdots, t_{k+\ell}, \cdots \), then the Markov process \( X_t \) satisfies the following condition:

\[
P[X_{t_k} | X_{t_{k+1}}, \cdots, X_{t_{k+\ell}}] = P[X_{t_k} | X_{t_{k+1}}]
\]

- Separation of the joint conditional probability: If \( t_k < t_\ell < t_m \), then

\[
P[X_{t_m}, X_{t_k} | X_{t_\ell}] = \underbrace{P[X_{t_m} | X_{t_\ell}]}_{\text{Markov Property}} \underbrace{P[X_{t_k} | X_{t_\ell}]}_{\text{Conditional Probability}}
\]
**Proposition 6.1.** (Chapman-Kolmogorov Equation) Let $X_t$ be a continuous Markov process. Let $t_1 < t_2 < t_3$. Then,

$$f_X(x_3|x_1;t_3,t_1) = \int_{\mathbb{R}} dx_2 \ f_X(x_3|x_2;t_3,t_2) f_X(x_2|x_1;t_2,t_1)$$

**Proof.** Marginal density $f_X(x_3,x_1;t_3,t_1) = \int_{\mathbb{R}} dx_2 \ f_X(x_3,x_2,x_1;t_3,t_2,t_1)$ leads to

$$f_X(x_3,x_1;t_3,t_1) = \int_{\mathbb{R}} dx_2 \ f_X(x_3|x_2;t_3,t_2) f_X(x_2,x_1;t_2,t_1)$$

by using the Markov property of $X_t$. Then, dividing both sides by $f_X(x_1;t_1)$ yields

$$\frac{f_X(x_3,x_1;t_3,t_1)}{f_X(x_1;t_1)} = \int_{\mathbb{R}} dx_2 \ f_X(x_3|x_2;t_3,t_2) \frac{f_X(x_2,x_1;t_2,t_1)}{f_X(x_1;t_1)}$$

$$\Rightarrow f_X(x_3|x_1;t_3,t_1) = \int_{\mathbb{R}} dx_2 \ f_X(x_3|x_2;t_3,t_2) f_X(x_2|x_1;t_2,t_1)$$

$\square$

**Definition 6.5.** (Increment and Independent Increment) The increment of a stochastic process $\{X_t : t \in \mathcal{T} \subseteq [0, \infty)\}$ over an interval $[a,b] \subseteq \mathcal{T}$ is a random variable $X_b - X_a$. A stochastic process is said to have independent increments if, for any positive integer $n$ and any $0 < t_1 < \cdots < t_n \in \mathcal{T}$, the increments, $X_{t_1} - X_0$, $X_{t_2} - X_{t_1}$, $\cdots$, $X_{t_n} - X_{t_{n-1}}$ are mutually independent.

**Remark 6.4.** (Markov property of independent-increment processes) It follows from Definition 6.5 that a stochastic process $\{X_t : t \in \mathcal{T}\}$ is an independent-increment process if the following conditions hold:

- $X_0$ is a real constant.
- For any $t_1 < \cdots < t_{n+1}$, the vector $(X_{t_1}, \cdots, X_{t_n})$ is a function of the $n$ increments, $X_{t_1} - X_0$, $X_{t_2} - X_{t_1}$, $\cdots$, $X_{t_n} - X_{t_{n-1}}$, and is thus independent of the increment $X_{t_{n+1}} - X_{t_n}$. But $X_{t_{n+1}}$ is determined by $X_{t_{n+1}} - X_{t_n}$ and $X_{t_n}$ only. Thus, $\{X_t : t \in [0, \infty)\}$ is a Markov process. Later, we shall encounter random walk, Brownian motion, and Poisson process that are all independent-increment processes.

In other words, the vector $(X_{t_1}, \cdots, X_{t_n})$ can be generated from the $n$ increments, $X_{t_1} - X_0$, $X_{t_2} - X_{t_1}$, $\cdots$, $X_{t_n} - X_{t_{n-1}}$, and is thus independent of the increment $X_{t_{n+1}} - X_{t_n}$. But $X_{t_{n+1}}$ is determined by $X_{t_{n+1}} - X_{t_n}$ and $X_{t_n}$ only.
Examples of Markov Processes

This section presents examples of Markov processes that are commonly encountered in science and engineering, such as Wiener process and Poisson process.

Random Walk Sequence

This subsection deals with a discrete-time Markov chain that is the cumulative sum of infinite-length independent Bernoulli trials. Let us consider independent flips of a fair coin, where the sample space is $\Omega = \{H, T\}$; the event space is $2^\Omega$; and $P[\{H\}] = \frac{1}{2}$ and $P[\{T\}] = \frac{1}{2}$. If the outcome of a coin flip is $H$, we move from the current position to the right by one step; and if the outcome is $T$, we move from the current position to the left by one step. Let the step length be $s > 0$ after each flip.

Starting from the zero position, if we are at the position $m$ after $n$ flips, then $m = r - (n - r) = 2r - n$ if there are $r$ occurrences of $H$ out of $n$ flips. Hence, $r = \frac{n+m}{2}$. Let us denote the $k^{th}$ flip as a random variable $w_k$, where

$$w_k = \begin{cases} s & \text{if } H \text{ occurs (with probability } \frac{1}{2}) \\ -s & \text{if } T \text{ occurs (with probability } \frac{1}{2}) \end{cases}$$

Let the cumulative effect of $n$ consecutive flips be realized as a random variable $X_n \triangleq \sum_{k=1}^{n} w_k$ with $X_0 = 0$ with probability 1. Then,

$$P[X_n = ms] = \begin{cases} \frac{n!}{(\frac{n+m}{2})! \cdot (\frac{n-m}{2})!} & \text{if } (n+m) \text{ is an even integer} \\ 0 & \text{if } (n+m) \text{ is an odd integer} \end{cases}$$

implying $(\frac{n+m}{2})$ occurrences of $H$ and $(\frac{n-m}{2})$ occurrences of $T$ in a total of $n$ flips.

Remark 6.5. Individual flips of the coin are assumed to be independent events. Therefore, $F_{X_n} = F_{w_1 \cdots w_n} = F_{w_1} \cdots F_{w_n}$ and the sum of independent Bernoulli trials leads to a binomial process.

Since $E[w_k] = \frac{s}{2} - \frac{s}{2} = 0$ and $\text{var}[w_k] = \frac{s^2}{2} + \frac{(-s)^2}{2} = s^2$ and $n$ trials are pairwise independent events, it follows that

$$E[X_n] = 0 \text{ and } \text{var}[X_n] = ns^2$$

Let us normalize $X_n$ to have unity standard deviation for each $n \in \mathbb{N}$ to yield $W_n \triangleq \frac{X_n}{\sqrt{ns}}$, which yields the following results:

$$E[W_n] = 0 \text{ and } \text{var}[W_n] = 1 \ \forall n \in \mathbb{N}$$
By Central Limit Theorem, as $n \to \infty$, the random walk sequence converges in distribution to the zero-mean unity-variance Gaussian random variable, i.e., $W_n \to N(0,1)$. So, for a sufficiently large $n$,

$$P[\alpha < W_n \leq \beta] = P[\alpha \sqrt{ns} < X_n \leq \beta \sqrt{ns}] \approx (\text{erf}(\beta) - \text{erf}(\alpha))$$

where the error function is defined as: $\text{erf}(x) \triangleq \frac{2}{\sqrt{\pi}} \int_{0}^{x} dt \exp(-t^2)$.

**Brownian Motion** Brownian motion is the consequence of random events in the transport phenomena of particles suspended in a fluid (i.e., a liquid or a gas) resulting from their collision with the quick atoms or molecules in the fluid. The Wiener process refers to the mathematical model used to describe the Brownian motion.

In 1827, while looking through a microscope at particles trapped in cavities inside pollen grains in water, the botanist Robert Brown noted that the particles moved through the water but he was not able to determine the mechanisms that caused this motion. Although atoms and molecules had long been theorized as the constituents of matter, it was Albert Einstein who published a paper in 1905, which explained in precise details of how the Brownian motion is a result of the pollen being moved by individual water molecules. This explanation of Brownian motion served as definitive confirmation that atoms and molecules actually exist. In 1908, this observation was further verified experimentally by Jean Perrin who was awarded the Nobel Prize in Physics in 1926 for his work on *discontinuous structure of matter*. The underlying concept of Brownian motion is as follows:

*As the direction of the force of atomic bombardment is constantly changing, the particles are hit more on one side than another, leading to the seemingly random nature of the motion.*

**Einstein’s Theory of Brownian Motion**: The first part of Einstein’s theory formulates a diffusion equation for Brownian particles, in which the diffusion coefficient is related to the mean squared displacement of a Brownian particle, while the second part relates the diffusion coefficient to measurable physical quantities. This observation led Einstein to determine the size of atoms, and how many atoms there are in a mole, or the molecular weight in grams, of a gas. Let us digress this point a little further.

*Avogadro’s law states that, under the same conditions of temperature and pressure, equal volumes of different perfect (ideal) gases contain*
an equal number of molecules. This empirical relation can be derived from the kinetic theory of gases under the assumption of a perfect (ideal) gas.

In accordance to Avogadro’s law, the volume under consideration is the same for all ideal gases, which is $\sim 22.4$ liters at standard temperature and pressure; one gram mole (i.e., molecular weight in grams) of an ideal gas at standard temperature and pressure contains $6.022140857 \times 10^{23}$ molecules, known as the Avogadro’s number, and the determination of this number is equivalent to the knowledge of the mass of an atom as the latter is obtained by dividing the mass of a mole of the gas by Avogadro’s number. We present a simplified form of the first part of Einstein’s theory of Brownian motion in the next paragraph.

Einstein’s argument was to determine how far a Brownian particle travels in a given time interval. Classical mechanics is unable to determine this distance because of the enormous number of bombardments a Brownian particle undergoes, roughly of the order of $\sim 1,021$ collisions per second. Einstein considered the collective motion of Brownian particles, where the increment of particle positions in unrestricted one-dimensional domain was modeled as a random variable with a probability density function under coordinate transformation so that the origin lies at the initial position of the particle. Furthermore, assuming conservation of particle number, he expanded the density (i.e., the number of particles per unit volume) in a Taylor series. Then, he derived that the mass density function $\rho(x, t)$ of Brownian particles at point $x$ at time $t$ satisfies the diffusion equation:

$$\frac{\partial \rho}{\partial t} = \alpha \frac{\partial^2 \rho}{\partial x^2},$$

where $\alpha$ is the mass diffusivity.

With the initial condition of $\rho(x, 0) = \delta(x)$ and boundary conditions of $\rho(\infty, t) = \rho(-\infty, t) = 0$, the diffusion equation yields the solution:

$$\rho(x, t) = \frac{1}{\sqrt{4\pi \alpha t}} \exp \left( -\frac{x^2}{4\alpha t} \right)$$

This expression allowed Einstein to calculate the moments directly. The first moment is seen to vanish, meaning that the Brownian particle is equally likely to move to the left as it is to move to the right. The second moment is, however, non-vanishing, being given by:

$$\overline{x^2} = 2\alpha t$$
which expresses the mean squared displacement in terms of the time elapsed and the diffusivity. From this expression Einstein argued that the rms displacement of a Brownian particle is *not* proportional to the elapsed time, but rather to its square root. Einstein’s argument is based on a conceptual switch from the *ensemble* of Brownian particles to the *single* Brownian particle, where the relative number of particles at a single instant is the time that a Brownian particle takes to reach a given point.

**Derivation of Brownian Motion as a limit point of the Random Walk Sequence:** The Brownian motion process, also known as Wiener-Levy process, is a continuous-time Markov process that can be derived from the random walk sequence described in the previous subsection. In the one-dimensional form, the Brownian motion process is expressed as:

$$
\beta : [0, \infty) \times \Omega \to \mathbb{R}
$$

at a given time instant $t \in [0, \infty)$, the random variable $\beta_t$ is zero-mean Gaussian and its variance is proportional to the time parameter $t$. Let us consider a random walk sequence, where the (uniform) step size $s$ and the (uniform) time interval $\Delta$ between consecutive steps are both infinitesimally small. Let us denote this (random walk) process as $X_\Delta^t$ that is characterized by the time interval $\Delta > 0$, where $X_\Delta^t$ is piecewise constant between consecutive step $s$. Starting from the zero position at the time $t = 0$ to some finite final time $t = n\Delta$, it follows that

$$
X_\Delta^t = \sum_{k=0}^{n-1} w_k \Rightarrow X_{n\Delta}^\Delta = \sum_{k=0}^{n-1} w_k
$$

where, by following Subsection 6, the Bernoulli trial of the flip of a fair coin is:

$$
w_k = \begin{cases} 
  s & \text{if } H \text{ occurs (with probability } \frac{1}{2} \text{)} \\
  -s & \text{if } T \text{ occurs (with probability } \frac{1}{2} \text{)} 
\end{cases}
$$

Then, $E[X_\Delta^t] = 0$ and $\text{var}[X_\Delta^t] = ns^2 = \frac{s^2}{\Delta}$. By following Einstein’s derivation, we set $s^2 = 2\alpha\Delta$, where $\alpha$ is the diffusion constant, i.e., the step size is proportional to the square root of the time interval ($s = \sqrt{2\alpha\Delta}$).

Having $s^2 = 2\alpha\Delta$, as $\Delta \to 0^+$, $s \to 0^+$. Then, for any finite time $t > 0$ as $\Delta \to 0^+$, the limiting condition of the random sequence becomes the Brownian motion (or Wiener-Levy) process

$$
\beta_t = \lim_{\Delta \to 0^+} X_\Delta^t \quad \text{where } \beta_0 = 0 \text{ with probability } 1
$$
i.e., for all finite $t$ (i.e., $t \in [0, \infty)$),

$$\beta_t = \lim_{n \to \infty} \sum_{k=0}^{n-1} w_k$$

which converges in distribution to a Gaussian process by virtue of Central Limit Theorem because $w_k$’s are i.i.d. random variables. Then, for all $t \in [0, \infty)$, $E[\beta_t] = 0$ and \(\text{var} [\beta_t] = \lim_{n \to \infty} ns^2 = 2\alpha t\). In other words, the probability density function of $\beta_t$ for all finite $t$ is:

$$f_{\beta}(x; t) = \frac{1}{\sqrt{4\pi \alpha t}} \exp \left( - \frac{x^2}{4\alpha t} \right)$$

Birth and Death Process

Let $N_t : t \in [0, \infty)$ be a continuous-time Markov chain having exactly one of the non-negative integer values, $0, 1, 2, 3, \cdots$ at any given instant $t$. Examples of the birth and death process are:

- Population of a city where people enter and leave the city in continuous time.
- Queue length at the buffer of a computer.
- Jobs waiting at a transfer station within a manufacturing environment.

Let us denote the state of the system at time $t$ as $i$ if $N_t = i$. Let the conditional probability of transition from the state $i$ to its neighboring states $(i+1)$ and $(i-1)$ over the (closed) interval $[t, t+\delta]$ (where $\delta > 0$ and $\delta \to 0$) to be as follows:

$$P[N_{t+\delta} = (i+1)|N_t = i] = \lambda_i \delta + o(\delta) \quad \text{and} \quad P[N_{t+\delta} = (i-1)|N_t = i] = \mu_i \delta + o(\delta)$$

where the parameters $\lambda_i \geq 0$ and $\mu_i \geq 0$ are respectively called the birth rate and the death rate at the state $i$; and $o(\bullet)$ is a class of functions such that, if $\lim_{\delta \to 0} \frac{|g(\delta)|}{\delta} = 0$, then the function $g(\bullet)$ belongs to the class of $o(\bullet)$. Figure 5 elucidates the state transitions of the birth and death process.

Remark 6.6. If $\theta$ is any finite real number, then $(g(\bullet) \sim o(\bullet)) \Rightarrow (\theta g(\bullet) \sim o(\bullet))$.

Next two major assumptions are made to derive the governing equations for the birth and death process, as stated below.
• Assumption 1: Every change of state occurs instantaneously.
• Assumption 2: The probability of any change of a state by more than 1 in an infinitesimally small interval \([t, t + \delta]\) is \(o(\delta)\), i.e., as \(\delta \to 0\), it follows that
\[
P[|N_{t+\delta} - N_t| > 1] = o(\delta)\]
and hence
\[
\sum_{j: |j-i| \geq 2} P[N_{t+\delta} = j|N_t = i] = o(\delta) \quad \forall i
\]
Since \(\sum_{j=0}^{\infty} P[N_{t+\delta} = j|N_t = i] = 1 \forall i\), it follows that, as \(\delta \to 0\),
\[
P[N_{t+\delta} = i|N_t = i] = 1 - (\lambda_i + \mu_i)\delta + o(\delta) \quad \forall i
\]
Then, by using the theorem of total probability, which is stated below as:
\[
P[N_{t+\delta} = j] = \sum_{i=0}^{\infty} P[N_{t+\delta} = j|N_t = i] P[N_t = i] \quad \forall j
\]
and denoting \(P_j(t) \equiv P[N_t = j]\) for brevity and assigning \(\lambda_{-1} = 0; \mu_0 = 0\),
and \(P_{-1}(t) = 0 \quad t \in [0, \infty)\), it follows that, as \(\delta \to 0\),
\[
P_i(t + \delta) = \lambda_{i-1}P_{i-1}(t) + \mu_{i+1}P_{i+1}(t) + (1 - (\lambda_i + \mu_i)\delta)P_i(t) + o(\delta)
\]
which leads to:
\[
\lim_{\delta \to 0} \frac{P_i(t + \delta) - P_i(t)}{\delta} = \lambda_{i-1}P_{i-1}(t) + \mu_{i+1}P_{i+1}(t) - (\lambda_i + \mu_i)P_i(t) + \lim_{\delta \to 0} \frac{o(\delta)}{\delta}
\]
To solve the resulting differential-difference equation:
\[
\frac{dP_i(t)}{dt} = \lambda_{i-1}P_{i-1}(t) + \mu_{i+1}P_{i+1}(t) - (\lambda_i + \mu_i)P_i(t)
\]
with \(\lambda_{-1} = 0; \mu_0 = 0\); and \(P_{-1}(t) = 0 \quad t \in [0, \infty)\), we need the initial
conditions for individual state probabilities: \(P_i(0), i = 0, 1, 2, 3, \ldots\), at time \(t = 0\).

**Example 6.1. (The Poisson Process)** Let us consider a pure birth process
with a constant birth rate, i.e., \(\lambda_i = \lambda > 0\) and the death rate \(\mu_i = 0 \forall i\), and
initially occupying the zero state with probability 1, i.e.,
\[
P_i(0) = \begin{cases} 
1 & \text{for } i = 0 \\
0 & \text{for } i > 0 
\end{cases}
\]
The governing equation for the generic birth and death process is thus simplified as: \( \frac{dP_i(t)}{dt} = \lambda (P_{i-1}(t) - P_i(t)) \) with the above initial conditions. A real-life example is an operator recording incoming telephone calls that arrive at random instants of time.

With \( P_{-1}(t) = 0 \ \forall t \), the dynamics at state \( i = 0 \) becomes: \( \frac{dP_0(t)}{dt} = -\lambda P_0(t) \) at the initial condition \( P_0(0) = 1 \), which yields the solution: \( P_0(t) = \exp(-\lambda t) \ \forall t \geq 0 \). Next consider \( i = 1 \): \( \frac{dP_1(t)}{dt} = -\lambda P_1(t) + \lambda \exp(-\lambda t) \) with the initial condition \( P_1(0) = 0 \), which yields the solution: \( P_1(t) = (\lambda t) \exp(-\lambda t) \ \forall t \geq 0 \). Similarly for \( j = 2: \) \( \frac{dP_2(t)}{dt} = -\lambda P_2(t) + (\lambda^2 t) \exp(-\lambda t) \) with the initial condition \( P_2(0) = 0 \), which yields the solution: \( P_2(t) = (\lambda t)^2 \exp(-\lambda t) \ \forall t \geq 0 \). By applying the method of induction, the following result is generated:

\[
P_k(t) = \frac{(\lambda t)^k}{k!} \exp(-\lambda t) \ \forall t \geq 0 \ \forall k \geq 0
\]

which describes the Poisson process with a constant birth rate \( \lambda > 0 \). Poisson process plays a key role in the queueing theory.

**Example 6.2. (The Binomial Process)** Let us consider a pure death process with the death rate being proportional to the state, i.e., \( \mu_j = j \mu \), where \( \mu > 0 \) is a constant, and birth rate \( \lambda_j = 0 \ \forall j \), and initially occupying the zero initial state, i.e.,

\[
P_j(0) = \begin{cases} 
1 & \text{for } j = n \\
0 & \text{for } j < n 
\end{cases}
\]

The following result is generated by by solving the generic governing equation of the birth and death process:

\[
P_j(t) = \binom{n}{j} \exp(-j \mu t) (1 - \exp(-\mu t))^{(n-j)} \ \forall t \geq 0 \ \text{and } j = 0, 1, 2, \cdots, n
\]

which describes the binomial process with \( \mu > 0 \). Here, the departure events at random time instants are i.i.d. Bernoulli.

**Remark 6.7.** One may view the Poisson process to be a continuous time version of the Bernoulli trials process. To see this, let us suppose that each success in the Bernoulli trials process as a random point in discrete time. Then the Bernoulli trials process, like the Poisson process, has the strong renewal (memoryless) property, i.e., at the instant of each arrival, the process starts over independently of the past. With this analogy in mind, the Poisson process has connections with Bernoulli trials and binomial processes as seen below.
• The inter-arrival times have independent geometric distributions in the Bernoulli trials process; they have independent exponential distributions in the Poisson process.

• The arrival times have Pólya (also called negative binomial) distributions in the Bernoulli trials process; they have gamma distributions in the Poisson process. Note: A negative binomial distribution is the probability mass function, which is denoted as:

\[ NBin(k; r, p) \sim \binom{k + r - 1}{k} p^k (1 - p)^r \]

where the (non-random) non-negative integer \( r \) signifies the number of failures; \( k \) is the number of successes; and \( p \) is the probability of successes. Compare it with the binomial distribution having a total of \( n \) Bernoulli trials

\[ Bin(k; n, p) \sim \binom{n}{k} p^k (1 - p)^{n-k} \]

• The number of arrivals in an interval has a binomial distribution in the Bernoulli trials process; it has a Poisson distribution in the Poisson process.

Law of Large numbers

This section is devoted to two versions of the law of large numbers, namely,

• Weak law of large numbers;
• Strong law of large numbers.

Weak Law of Large Numbers

We present two versions although there are many other versions.

**Version 1 of Weak Law of Large Numbers (Uniform Variance)**

**Theorem 6.1.** Let \( \{X_k\} \) be a sequence of independent random variables on a probability space \((\Omega, \mathcal{E}, P)\) such that

\[ E[X_k] = \mu_X \quad \text{and} \quad \text{var}[X_k] = \sigma_X^2 \quad \forall k \in \mathbb{N} \]
Define another random sequence \( \{ \hat{\mu}_X(n) \} \) as
\[
\hat{\mu}_X(n) \triangleq \frac{1}{n} \sum_{k=1}^{n} X_k \quad \forall n \in \mathbb{N}
\]
Then, \( \hat{\mu}_X(n) \xrightarrow{p} \mu_X \), i.e., converges in probability.

Proof. By Chebyshev inequality, it follows that
\[
P[|\hat{\mu}_X(n) - \mu_X| \geq \delta] \leq \frac{\text{var}(\hat{\mu}_X(n))}{\delta^2} = \frac{\sigma_X^2}{n \delta^2}
\]
Therefore, the convergence is established in the \( L_2 \) sense, which suffices convergence in probability.

**Version 2 of Weak Law of Large Numbers (Non-uniform Variance)**

Let \( \{X_k\} \) be a sequence of independent random variables on a probability space \((\Omega, \mathcal{E}, P)\) such that
\[
E[X_k] = \mu_X \quad \text{and} \quad \text{var}[X_k] = \sigma_X^2(k) \quad \forall k \in \mathbb{N}
\]
Define another random sequence \( \{ \hat{\mu}_X(n) \} \) as
\[
\hat{\mu}_X(n) \triangleq \frac{1}{n} \sum_{k=1}^{n} X_k \quad \forall n \in \mathbb{N}
\]
If \( \sum_{k=1}^{n} \frac{\sigma_X^2(k)}{n^2} < \infty \), then \( \hat{\mu}_X(n) \xrightarrow{p} \mu_X \), i.e., converges in probability.

Proof. By Chebyshev inequality, it follows that
\[
P[|\hat{\mu}_X(n) - \mu_X| \geq \delta] \leq \frac{\sum_{k=1}^{n} \sigma_X^2(k)}{n^2 \delta^2}
\]
Therefore, the convergence is established in the \( L_2 \) sense, which suffices convergence in probability.

**Strong Law of Large Numbers**

The strong law of large numbers is stated below.

(Strong Law of Large Numbers) Let \( \{X_k\} \) be a sequence of wide-sense stationary independent random variables on a probability space \((\Omega, \mathcal{E}, P)\) such that
\[
E[X_k] = \mu \quad \text{and} \quad \text{var}[X_k] = \sigma^2 \quad \forall k \in \mathbb{N}
\]
Define another random sequence \( \{\hat{\mu}_n\} \) as
\[
\hat{\mu}_n \triangleq \frac{1}{n} \sum_{k=1}^{n} X_k \quad \forall n \in \mathbb{N}
\]

Then, \( \hat{\mu}(n) \xrightarrow{a.s.} \mu \), i.e., converges almost surely to \( \mu \).

As the proof of strong law of large numbers requires Martingale Convergence theorem, we first introduce the notion of Martingales.

**Definition 6.6.** *(Martingale)* A stochastic process \( \{X_t : t \in \mathcal{T}\} \) is called a martingale if the following conditions hold:

- \( X_t \in L_1(\Omega, \mathcal{E}, P) \), i.e., \( E[|X_t|] \) is finite for all \( t \in \mathcal{T} \),
- \( E[X_{t_{n+1}} | X_{t_n}, \cdots, X_{t_2}, X_{t_1}] = X_{t_n} \) for any \( n \in \mathbb{N} \) and any \( t_0 < t_1 < \cdots < t_n \in \mathcal{T} \).
  
  Equivalently, \( E[X_{t_{n+1}} - X_{t_n} | X_{t_n}, \cdots, X_{t_2}, X_{t_1}] = 0 \).

**Remark 6.8.** In Definition 6.6, if \( t_n \) is interpreted as the present time, then \( t_{n+1} \) is a future time and the vector \( (X_{t_1}, \cdots, X_{t_n}) \) represents the past and present. Then, the martingale property implies that the future increments of \( X \) have zero conditional means, given the past and present values of the process.

An alternative and more rigorous statement of martingale in Definition 6.6 is as follows.

Let \( \{X_t : t \in \mathcal{T}\} \) be a stochastic process on \( (\Omega, \mathcal{E}, P) \). We now bring in the notion of martingale in terms of filtration.

**Definition 6.7.** A filtration on the measurable space \( (\Omega, \mathcal{E}) \) is a family of monotonically increasing \( \sigma \)-algebras \( \{\mathcal{F}_t \subset \mathcal{E} : t \in \mathcal{T}\} \), i.e., \( (s < t) \Rightarrow (\mathcal{F}_s \subset \mathcal{F}_t) \). Then, a stochastic process \( X_t : t \in \mathcal{T} \) on \( (\Omega, \mathcal{E}, P) \) is called a martingale with respect to the filtration \( \{\mathcal{F}_t\} \) and the probability measure \( P \) if

(i) \( X_t \) is \( \mathcal{F}_t \)-measurable for all \( t \in \mathcal{T} \).

(ii) \( X_t \in L_1(\Omega, \mathcal{E}, P) \), i.e., \( E[|X_t|] < \infty \) for all \( t \in \mathcal{T} \).

(iii) \( E[X_t | \mathcal{F}_s] \xrightarrow{a.s.} X_s \) if \( s < t \) for all \( s, t \in \mathcal{T} \).

Definition 6.7 is summarized as follows: Let \( \{\mathcal{F}_t : t \in \mathcal{T}\} \) be a monotonically increasing set of \( \sigma \)-algebras, i.e., \( \mathcal{F}_s \subset \mathcal{F}_t \subset \mathcal{E} \). Then, a stochastic process
\{X_t : t \in T \}\) is called a martingale if, for all \(s < t\),
\[
E[X_t|F_s] \overset{a.s.}{=} X_s, \quad \text{where} \quad E[X_t|F_s] \triangleq E[X_t|X_\tau : \tau < s]
\]
Furthermore, a stochastic process \(\{X_t : t \in T \}\) is called a supermartingale if \(E[X_t|F_s] \overset{a.s.}{\leq} X_s\) and is called a submartingale if \(E[X_t|F_s] \overset{a.s.}{\geq} X_s\).

**Example 6.3.** Let \(\{W_k\}\) be a sequence of iid Bernoulli trials with \(E[W_k] = 0\) and let \(X_n = \sum_{k=0}^{n} W_k\). Then, \(\{X_k\}\) is a martingale sequence as seen below.
\[
E[X_n|X_{n-1} \cdots X_0] = E\left[\sum_{k=0}^{n} W_k|X_{n-1} \cdots X_0\right] = \sum_{k=0}^{n} E[W_k|X_{n-1} \cdots X_0]
\]
\[
= \sum_{k=0}^{n} E[W_k] + E[W_n] = X_{n-1}
\]
What happens if \(E[W_k] \neq 0\)?

**Example 6.4.** Let \(\{X_k\}\) be a sequence of zero-mean random variables with independent increments. Then, \(\{X_k\}\) is a martingale sequence as seen below.
\[
E[X_n|X_{n-1} \cdots X_0] = E[(X_n - X_{n-1})|X_{n-1} \cdots X_0] + E[X_{n-1}|X_{n-1} \cdots X_0]
\]
\[
= E[(X_n - X_{n-1})] + X_{n-1} = X_{n-1}
\]
What happens if \(\{X_k\}\) be a sequence of non-zero-mean random variables with independent increments?

The next lemma and the theorem are required for the martingale sequence to converge.

**Lemma 6.1.** Let \(\{X_n\}\) be a martingale sequence for non-negative integers \(n\). Then, for all \(\varepsilon > 0\) and any positive integer \(n\),
\[
P\left[\max_{0 \leq k \leq n} |X_k| \geq \varepsilon \right] \leq \frac{E|X_n|^2}{\varepsilon^2}
\]
**Proof.** For \(0 \leq j \leq n\), let us define mutually exclusive events,
\[A_j \triangleq \{|X_k| \geq \varepsilon \text{ the first time at } j\}\]
Then, the event \(\max_{0 \leq k \leq n} |X_k| \geq \varepsilon\) is just a union of the events \(A_j\)'s, where \(0 \leq j \leq n\) and the respective indicator functions are defined as:
\[
\chi_j \triangleq \begin{cases} 
1 & \text{if } A_j \text{ occurs} \\
0 & \text{if } A_j \text{ does not occur}
\end{cases}
\]
It is noted that at most one of these $\chi_j$’s is equal to 1 and the rest are 0’s. That is, $E[|X_n|^2] \geq \sum_{j=0}^{n} E[|X_n|^2 \chi_j]$. Then, by expanding $|X_n|^2$ as $(X_j + (X_n - X_j))^2$ and substituting the above inequality, it follows that

$$E[|X_n|^2] \geq \sum_{j=0}^{n} E[|X_j|^2 \chi_j] + 2 \sum_{j=0}^{n} E[X_j(X_n - X_j)\chi_j]$$

Denoting $X_j \chi_j$ as $Z_j$ (that depends only on $X_0, \cdots, X_j$), it follows that

$$E[Z_j(X_n-X_j)] = E[E[Z_j(X_n-X_j)|X_0, \cdots, X_j]] = E[Z_jE[(X_n-X_j)|X_0, \cdots, X_j]]$$

$$= E[Z_j(X_j - X_j)] = 0$$

In the above derivation, we have used the identity $E[Y] = E[E[Y|X]]$ in the first equality and the Martingale property in the third equality. Therefore,

$$E[|X_n|^2] \geq \sum_{j=0}^{n} E[|X_j|^2 \chi_j] \geq \varepsilon^2 \sum_{j=0}^{n} E[\chi_j] = \varepsilon^2 P\left[ \bigcup_{j=0}^{n} A_j \right] = \varepsilon^2 P\left[ \max_{0 \leq k \leq n} |X_k| \geq \varepsilon \right]$$

\[\Box\]

**Theorem 6.2. (Martingale Convergence Theorem)** Let $\{X_n\}$ be a martingale sequence for non-negative integers $n$, satisfying the condition: $E[|X_n|^2] \leq C < \infty$ for some $C \in \mathbb{R}$. Then, $X_n \xrightarrow{a.s.} X$ as $n \to \infty$, where $X$ is the limiting random variable.

**Proof.** Having the integers $m \geq 0$ and $n \geq 0$, let $Y_n \triangleq (X_{n+m} - X_m)$. Then, $\{Y_n\}$ is a Martingale sequence and it follows from Lemma 6.1 that

$$P\left[ \max_{0 \leq k \leq n} |X_{m+k} - X_m| \geq \varepsilon \right] \leq \frac{E[|X_n|^2]}{\varepsilon^2} \quad \forall \varepsilon > 0$$

Having $E[|Y_n|^2] = E[|X_{n+m}|^2] - 2E[X_mX_{n+m}] + E[|X_m|^2]$ and the middle term is rewritten as

$$E[X_mX_{n+m}] = E[X_mE[X_{n+m}|X_m, \cdots, X_0]] = E[|X_m|^2]$$

where the above expression follows from Definition 6.7 because $\{X_n\}$ is a martingale sequence and $(n + m) \geq m$. Then, it follows that $\{E[|X_n|^2]\}$ is monotonically non-decreasing sequence of real numbers, because

$$E[|Y_n|^2] = E[|X_{n+m}|^2] - E[|X_m|^2] \geq 0 \quad \forall m, n \geq 0$$
Since \( \{ E[|X_n|^2]\} \) is bounded from above by \( C < \infty \), it must converge to a limit. Therefore, \( E[|Y_n|^2] \to 0 \) as \( m, n \to 0 \). Therefore,

\[
\lim_{m \to \infty} P \left[ \max_{k \geq 0} |X_{m+k} - X_m| \geq \varepsilon \right] = 0
\]

which by continuity of the probability measure, implies that

\[
P \left[ \lim_{m \to \infty} \max_{k \geq 0} |X_{m+k} - X_m| \geq \varepsilon \right] = 0
\]

Finally, due to completeness of Borel-measurable spaces, it is concluded that there exists a random variable \( X \) such that \( X_n \overset{a.s.}{\to} X \) as \( n \to \infty \).

Now we present the proof of strong law of large numbers, which is restated below.

**Theorem 6.3. (Strong Law of Large Numbers)** Let \( \{X_k\} \) be a sequence of wide-sense stationary independent random variables on a probability space \((\Omega, \mathcal{E}, P)\) such that

\[
E[X_k] = \mu \quad \text{and} \quad \text{var}[X_k] = \sigma^2 < \infty \quad \forall k \in \mathbb{N}
\]

Define another random sequence \( \{\hat{\mu}_n\} \) as

\[
\forall n \in \mathbb{N} \quad \hat{\mu}_n \overset{a.s.}{=} \frac{1}{n} \sum_{k=1}^{n} X_k \quad \text{(i.e., with probability 1)}
\]

Then, \( \hat{\mu}_n \overset{a.s.}{\to} \mu \), i.e., almost sure convergence.

**Proof.** Let us define \( X^c_k \triangleq X_k - \mu \) and \( Y_n \triangleq \sum_{k=1}^{n} \frac{X^c_k}{k} \) for \( n \in \mathbb{N} \) with \( Y_0 \forall n \overset{a.s.}{=} 0 \). \( \{Y_n\} \) is a Martingale sequence in view of the fact that \( \{X_k\} \) is a sequence of independent random variables,

\[
E[Y_n|Y_{n-1}, \ldots, Y_1] = E \left[ \left( \frac{X^c_n}{n} + Y_{n-1}\right)|Y_{n-1}, \ldots, Y_1 \right]
\]

\[
= E \left[ \frac{X^c_n}{n}|Y_{n-1}, \ldots, Y_1 \right] + Y_{n-1} = Y_{n-1}
\]

Since \( E[|Y_n|^2] = E\left[ \left| \sum_{k=1}^{n} \frac{X^c_k}{k} \right|^2 \right] = \sum_{k=1}^{n} \frac{\sigma^2}{k^2} = \sigma^2 \sum_{k=1}^{n} \frac{1}{k^2} < \infty \), because \( X_k \)'s are pairwise independent. Now, by applying Theorem 6.2, it follows that the sequence \( \{Y_n\} \to Y \) (a.s.), where \( Y \) is some random variable. Since \( Y_0 \overset{a.s.}{=} 0 \),

\[
X^c_k = (Y_k - Y_{k-1})k \Rightarrow \frac{1}{n} \sum_{k=1}^{n} X^c_k = \frac{1}{n} \left[ \sum_{k=1}^{n} kY_k - \sum_{k=1}^{n} kY_{k-1} \right] = \frac{1}{n} \left[ \sum_{k=1}^{n} kY_k - \sum_{k=1}^{n-1} (k+1)Y_k \right]
\]

44
\[
\frac{1}{n} \left[ \sum_{k=1}^{n} kY_k - \sum_{k=1}^{n-1} kY_k - \sum_{k=1}^{n-1} Y_k \right] = \frac{1}{n} \left[ nY_n - \sum_{k=1}^{n-1} Y_k \right] = Y_n - \frac{1}{n} \sum_{k=1}^{n-1} Y_k
\]

Therefore, \( \lim_{n \to \infty} X^c_k = \lim_{n \to \infty} Y_n - \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n-1} Y_k = Y - Y = 0 \) (a.s.), which implies \( \hat{\mu}_n \xrightarrow{a.s.} \mu \).

**Remark 6.9.** The convergence \( \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} Y_k = Y \) (a.s.) in the proof of Theorem 6.3 is based on the following fact.

If \( \{a_k\} \) is a (bounded) real sequence converging to \( a \) in the usual metric, then \( \left( \frac{1}{n} \sum_{k=1}^{n} a_k \right) \to a \) as \( n \to \infty \).

**Proof:** Since \( a_n \to a \), it follows that \( \forall \varepsilon > 0 \exists M \in \mathbb{N} \), such that \( |a_n - a| < \frac{\varepsilon}{2} \), \( \forall n \geq M \). Let \( \sum_{k=1}^{M-1} (a_k - a) < C < \infty \) and let \( n = \max \{ \frac{2C}{\varepsilon}, M \} \). Then,

\[
\left| \frac{1}{n} \sum_{k=1}^{n} (a_k - a) \right| \leq \left| \frac{1}{n} \sum_{k=1}^{M-1} (a_k - a) \right| + \left| \frac{1}{n} \sum_{k=M}^{n} (a_k - a) \right| < \frac{C}{n} + \frac{\varepsilon}{2} \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon
\]

i.e., \( \forall \varepsilon > 0 \exists M \in \mathbb{N} \) such that \( \left| \frac{1}{n} \sum_{k=1}^{n} a_k - a \right| < \varepsilon \), \( \forall n \geq M \).

**Remark 6.10.** (Converse to Strong Law of Large Numbers) (Ref: Feller Vol. 2, p. 241) Let \( \{X_k\} \) be a sequence of pairwise independent random variables with a common distribution and let \( E[X_k] = \infty \). Then, with probability 1, \( \lim \sup \left| \frac{1}{n} \sum_{k=1}^{n} X_k - c_n \right| = \infty \) for any real sequence \( \{c_k\} \).

Next we present an application of the strong law of large numbers for identification of empirical distributions in nonparametric statistics, which is an important topic of research in Statistical Learning.

**Theorem 6.4.** (Glivenko-Cantelli Theorem) Let \( X_1, X_2, \ldots \) be an iid sequence with distribution \( F \). Let \( F_n(x) \) be the frequency of occurrence of values of \( X_k \) such that \( X_k \leq x \), which is denoted as:

\[
F_n(x) \overset{\Delta}{=} \frac{1}{n} \sum_{k=1}^{n} 1_{(X_k \leq x)} \text{ where } 1_{*} = 1 \text{ if } * \text{ is true; otherwise } 1_{*} = 0.
\]

Then, \( F_n \) converges uniformly to \( F \) as \( n \to \infty \). In other words, \( \sup_x |F_n(x) - F(x)| \xrightarrow{a.s.} 0 \) as \( n \to \infty \).

**Proof.** First we present the proof for \( \{F_n\} \) being a sequence of monotonically increasing distribution functions converging to a bounded and continuous
distribution $F$. Let us fix $x$ and set $Z_n = 1_{(X_n \leq x)}$. Since $Z_n$ are iid with $E[Z_n] = P[(X_n \leq x)] = F(x)$, the strong law of large numbers implies that $F_n(x) = \frac{1}{n} \sum_{k=1}^{n} Z_k \overset{a.s.}{\to} F(x)$.

Next we relax the above assumption (e.g., the distribution $F$ is allowed to have finitely many discontinuities). As before, let us fix $x$ and set $Z_n = 1_{(X_n < x)}$. Since $Z_n$ are iid with $E[Z_n] = P[(X_n < x)] = F^-(x)$, the strong law of large numbers implies that $F_n(x^-) = \frac{1}{n} \sum_{k=1}^{n} Z_k \overset{a.s.}{\to} F^-(x^-)$. For $1 \leq j \leq k - 1$, let $x_{j,k} \triangleq \inf\{y : F(y) \geq \frac{j}{k}\}$; then, pointwise convergence of $F_n(x)$ and $F_n(x^-)$ implies that there exists a positive integer $k \geq 2$ such that if $n \geq k$, then

$$|F_n(x) - F_n(x_{j,k})| < \frac{1}{k} \quad \text{and} \quad |F_n(x^-) - F_n(x_{j,k}^-)| < \frac{1}{k} \quad \text{for} \quad 1 \leq j \leq k - 1$$

Letting $x_{0,k}$ approach $-\infty$ and letting $x_{k,k}$ approach $\infty$, the above two inequalities hold for $j = 0$ or $j = k$. If $x \in (x_{j-1,k}, x_{j,k})$ with $1 \leq j \leq k$ and $n \geq k$, then

$$F_n(x) \leq F_n(x_{j,k}^-) \leq F_n(x_{j,k}^-) + \frac{1}{k} \leq F_n(x_{j-1,k}) + \frac{2}{k} \leq F(x) + \frac{2}{k}$$

$$F_n(x) \geq F_n(x_{j-1,k}) \geq F_n(x_{j-1,k}^-) - \frac{1}{k} \geq F_n(x_{j,k}^-) - \frac{2}{k} \geq F(x) - \frac{2}{k}$$

From the above results, it follows that $\sup_x |F_n(x) - F(x)| \leq \frac{2}{k}$, which leads to the result of the theorem as $k \to \infty$. \qed