Probability theory aims to understand statistical behaviour of complex events. Many (maybe all) of these events themselves, such as throwing a dice, propagation of a disease or economy are deterministic in nature yet too complex to model exactly. Therefore the aim here is to model the long term statistical behaviour of these events rather then to give an exact solution.

It is worth a while to digest this statement. Things happen in nature; somebody throws a dice, particles float around, planets orbit etc. But people are curious and they want to understand the nature of these events[1]. This leads us to study physical events from a theoretical point of view. Such a study of a physical system requires one to write a mathematical model that describes certain aspects of its behaviour. One might be interested in getting a model that tries to mimic the physical evolution of the system or one that produces some statistical information about the event (such as how average energy of a collection of particles changes in time or the asymptotic behaviour of the trajectory of an object). In many cases "statistical models" are much more simpler and easier to solve or simulate and the results they predict coincide to a good extend with what one would get from experiments. But then how does one rigorously justify whether if it is reasonable to model statistically a given physical phenomenon? What is the relation between "statistical" and "physical" models? Lets give these considerations some substance through dice throwing example.

Consider that you were able to model a k simultaneous dice throwing event physically (up to good agreement with experimental results of course!). This means that once you fix certain parameters $s_1, ..., s_\ell$ (such as say initial throwing speed and direction, weight and shape of the dice and million other things) then there is a mathematical mechanism which tells what you will get after the next throw. We can put this in a more mathematical form. Assume you have have the phase space $M = \{0, 1, 2, ..., 6\}^\infty$. Let $(x_1, x_2, ...) \in M$ stand for a sequence of dice throwing results where $x_i = 0$ if the i^{th} throw has not been performed yet. Then we assume we have a map $f_k : M \times S_1 \times .. \times S_\ell \to M$ so that if one fixes the parameters $s_i \in S_i$ we get an exact result for the k^{th} dice throw:

$$f_k((x_1, ..., x_k, ...,), s_1, ..., s_\ell) = (x_1, ..., x_{k-1}, y_k, x_{k+1}, ...)$$

Note that the k^{th} throw of dice only changes the dice result x_k. Then if we denote shortly $s(k) = (s_1(k), ...s_\ell(k))$ a typical dice throwing experiment that is composed of k throws is of the following form:

$$f_k(f_{k-1}(...f_1((0, 0, ...), s(1)), s(2), ...), s(k)) = (x_1, ..., x_k, 0, 0, ...)$$

$$(0, 0, 0, ...) \underset{f_1(\cdot, s(1))}{\longrightarrow} (x_1, 0, 0, ...) \underset{f_2(\cdot, s(2))}{\longrightarrow} (x_1, x_2, 0, 0, ...) \underset{f_3(\cdot, s(3))}{\longrightarrow} ...$$

Such a model is extremely difficult to formulate and if we are not interested in an exact solution, then probability theory kicks in. It says the following: Assume that in some "ideal way" we randomly select k numbers from the set $\{1, 2, ..., 6\}$ (with equal probability of getting each number) and put together the result into a string of the form $(z_1, ..., z_k)$. Then experience tells us that the string

[1] The golden chalice to obtain at the end of all this is the power to predict future

$(z_1, ..., z_k)$ and $(x_1, ..., x_k)$ have the same asymptotic "statistical behaviour" as $k \to \infty$, where here "statistical behaviour" means such things as "the frequency of the number 3 appearing in both strings converges to the same number as $k \to \infty$". So in a way when we employ probability theory to model the behaviour of a deterministic system, we already make the assumption that asymptotically statistical behaviours are the same. This gives rise to the following natural question:

Question 0.1. *Is it possible to give a precise mathematical condition which allows us to understand whether it is "reasonable" to model a given system statistically using probability?*

Starting with Poincaré and Boltzmann, people have started investigating this question for dynamical systems that arise in certain parts of physics. Both Poincaré and Boltzmann were advocating that physical systems composed of many parts (such as molecules or systems with many point particles) are generally to complex to model exactly yet it should be possible to study their collective or statistical behaviour in the long run in a probabilistic manner. To be able to justify this point of view one either needs to experimentally verify many times that this statement is true or find an answer to question 0.1 which supplies us with some conditions that can be used to understand while system at hand is suitable for probabilistic modelling.

The latter attempt lead to what is known as Birkhoff ergodic theorem which precisely gives a a partial answer to this question by finding some conditions, which if satisfied results in a "deterministic dynamical system that behaves in a probabilistic manner". In these notes we will informally try to make this notion precise. Our aim will be to therefore first review some of the fundamental notions in measure theoretic probability theory using simple examples, so as to set up terminology. Then of course we will review needed definitions and concepts from dynamical systems, more precisely the notion of invariant measures, ergodicity and Birkhoff ergodic theorem. Finally we will demonstrate how to interpret these dynamical phenomenon from the point of view of probability.

1 Determinism and Probability

We will connect the world of deterministic dynamical systems to the world of probability by explaining the relation between the four theorems that we state in this section (see next sections for the explanations)

Theorem 1.1. *(*Strong Law of Large Numbers*) Let $(M, \mu, \mathcal{M})$ be a probability space and $\{X^k\}_{k=0}^{\infty}$ be a sequence of identically distributed and independent random observables. Then for μ-almost every point p one has that*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} X^i(p) = \int_M X^0 d\mu$$

Theorem 1.2. *(*Birkhoff Ergodic Theorem*) Let $(M, \mu, \mathcal{M})$ be a probability space and $f : M \to M$ a map such that μ is invariant and ergodic with respect to f. Given any random observable $X : M \to \mathbb{R}$, writing $X^i = X \circ f^i$, we have that for μ-almost every p,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} X^i(p) = \int_M X^0 d\mu$$

Remark 1.3. Here the integral on the right hand side can be with respect to any X^i, the particular choice of 0 does not matter. As we will see later this is due to the assumptions of identical distribution in the former and of μ being invariant in the latter theorem.

The first theorem is a fundamental theorem from probability theory which gives us an averaged information about the asymptotic behaviour of identically distributed independent random observables. The second theorem is a fundamental theorem from dynamical systems theory which tells us the asymptotic behaviour of a dynamical system by using random observables evaluated along the orbit of the system. We note that this also gives us information about where the orbit spends its time as if one takes a measurable set $A \subset M$ and as the random observable the characteristic function χ_A then ergodic theorem tells us that the frequency of visits of all most every point to the set A is equal to measure of A. This is precisely the kind of statistical information one hopes to get about general dynamical systems that posses an invariant measure.

The similarity between these two theorems is no coincidence and a deep connection between the two underlies here. In the next sections we hope to explain this connection at least scratch its surface. The key point is that invariance replaces identicality and ergodicity replaces (in the limit) independency.

On a more finer level we have the following two theorems:

Theorem 1.4. *(*Central Limit Theorem*) Let $(M, \mu, \mathcal{M})$ be a probability space and $\{X^k\}_{k=0}^{\infty}$ be a sequence of identically distributed and independent random observables with mean $\mathbb{E}(X)$ and variance σ. Then*

$$X^k \to_d X$$

where $\to_d$ denotes convergence in distribution and X is a normally distributed random observable with mean $\mathbb{E}(X)$ and variance σ.

Theorem 1.5. *Let $(M, \mu, \mathcal{M})$ be a probability space and $f : M \to M$ a map such that μ is invariant and ergodic with respect to f. Given any random observable $X : M \to \mathbb{R}$ such that $X \in L_\infty$, X has (at least) polynomial decay of correlations and $\int_M X d\mu = 0$. Writing $X^i = X \circ f^i$, there exists a normally distributed random variable X with mean 0 and some variance σ such that:*

$$X^k \to_d X$$

where

$$\sigma = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \int (X^i)^2 d\mu$$

Central limit theorem is a more fine characteristic of independent identically distributed random observable which not only gives statistical information about their average like the strong law of large numbers but it gives individual statistical information about them. Similarly decay of correlations is a strong requirement about a dynamical system which also gives finer information about the iterates of random observables that satisfy it.

2 Strong Law of Large Numbers and Ergodicity

2.1 Strong Law of Large Numbers in Probability

Theorem 2.1. *(*Strong Law of Large Numbers*) Let $(M, \mu, \mathcal{M})$ be a probability space and $\{X^k\}_{k=0}^{\infty}$ be a sequence of identically distributed and independent random observables. Then for μ almost every point p one has that*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} X^i(p) = \int_M X^0 d\mu$$

Lets try to digest this theorem by a simple example and some definitions. First of all a probability space is some space M, with some collection of subsets $\mathcal{M} \subseteq 2^M$ (where 2^M represents set of all subsets of M) which includes M and $\emptyset$ and a measure $\mu : \mathcal{M} \to \mathbb{R}$ such that $\mu(M) = 1$. A random observable X on this space is simply a function $X : M \to \mathbb{R}$ which is Lebesgue integrable with respect to μ. For people who are not familiar with Lebesgue integrability with respect to arbitrary measures, for the purposes of this talk it suffices to think of it intuitively as an integral where the measure μ assigns a certain volume to each region in the space, not necessarily homogenous or translation invariant as in the case of usual volume. In practice $\mathcal{M}$ represents possible outcomes of a system that is modelled probabilistically and μ assigns to each subset the probability that the next event will take

Two random observables X, Y are said to be identically distributed if in some sense that the likely hood of observing a certain range of values for X and Y are the same. This can be formalized as saying that for all intervals $I \subset \mathbb{R}$ one has that

$$\mu(\{p \mid X(p) \in I\}) = \mu(\{p \mid Y(p) \in I\})$$

Note that given any set $A \in M$, its measure is computed as $\mu(A) = \int_{p\in A} d\mu(p)$ therefore one can also write the above equality as

$$\int_{p\in X^{-1}(I)} d\mu(p) = \int_{p\in Y^{-1}(I)} d\mu(p)$$

for all I. If one also makes a change of variables then actually we can write these as integrals over $\mathbb{R}$ than over M:

$$\int_{y\in I} dX^*\mu(y) = \int_{y\in I} dY^*\mu(y)$$

Here $(X^*\mu)$ is a measure on $\mathbb{R}$ defined by $X^*\mu(A) = \mu(X^{-1}(A))$. $dX^*\mu$ simply stands for the new measure after a change of variables. $dX^*\mu$ is sometimes called the probability distribution. In the case where $d\ell$ is the Lebesgue measure on real line and there exists a function h_X such that $dX^*\mu = h_X d\ell$ then h_X is called probability distribution function. We note in the passing that if the range of a random observable X is finite (such as $\mathbb{Z}$) it is better to use the counting measure on this set. Note that since

$$\mathbb{E}(X) = \int_{p\in M} X(p) d\mu(p) = \int_{y\in\mathbb{R}} y dX^*\mu(y)$$

identically distributed random variables have same expected values.

Two random observables are said to be independent if for all I, J subsets of $\mathbb{R}$

$$\mu(\ \{p \mid X(p) \in I\}\ \cap\ \{p \mid Y(p) \in J\}\ ) = \mu(\{p \mid X(p) \in I\})\mu(\{p \mid Y(p) \in J\})$$

which has the intuition that getting a certain value for the random observable X does not effect the "chances" of getting some value for the random observable Y and so the chance of getting those values are simply the product of chances of getting each separately.

Lets see these concepts through a simple example which will be the space of infinite dice throws. Let $S = \{1, 2, 3, 4, 5, 6\}$

- $M = S \times S \times ...$ (infinite product)
- $\mathcal{M} = 2^M$
- $\mu(A_1 \times A_2 \times ...) = m(A_1)m(A_2)...$ where $m : 2^S \to \mathbb{R}$ is given by $m(\{a_1, ..., a_k\}) = \frac{k}{6}$

Note that in the definition above m is simply the function that assigns to each subset of M, the probability of a dice throw resulting in a number in that set, for instance : $m(\{2, 3\}) = \frac{2}{6}$. μ is simply an extension of this to the ideal case of infinite dice throws. Indeed if one sets $A = \{x_1\} \times \{x_2\} \times ... \times \{x_n\} \times S \times S...$, then this set represents the event where first n throws are exactly $x_1, ..., x_n$ and the following throws can be anything. Then one has that $\mu(A) = \frac{1}{6^n}$ which is precisely the probability of getting a fixed sequence of numbers in the first n throw. It is obvious that as you increase n this probability goes to 0 since you are asking to exactly throw a very long sequence of prescribed numbers, for which you must be very lucky! Indeed this corresponds to computing the measure of a set of the form $A = \{x_1\} \times \{x_2\} \times ...$, which is single point in M, and $\mu(A) = 0$. In this setup a sequence of simple random observables could be

$$f^k((x_1, x_2, ...)) = x_k$$

which can be interpreted as the random observable that observes the result k^{th} throw. Then these are indeed identically distributed since the probability of getting a certain number is $\frac{1}{6}$ regardless which throw you are making. It is also easy to see that for instance if $m \neq k$

$$\mu(\ \{p \mid f^k(p) = a\} \cap \{p \mid f^m(p) = b\}\ ) = \frac{1}{36} = \mu(\{p \mid X(p) \in I\})\mu(\{p \mid Y(p) \in J\})$$

and again intuitively the result of the k^{th} dice throw does not effect the probability of getting a certain number in the m^{th} throw.

2.2 Birkhoff Ergodic Theorem

Theorem 2.2. *(*Birkhoff Ergodic Theorem*) Let $(M, \mu, \mathcal{M})$ be a probability space and $f : M \to M$ a map such that μ is invariant and ergodic with respect to f. Given any random observable $X : M \to \mathbb{R}$, writing $X^i = X \circ f^i$, we have that for μ-almost every p,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} X^i(p) = \int_M X^0 d\mu$$

The aim of this section is to somehow model an "ergodic dynamical system" as a sequence of identically distributed but not independent random observables which in some sense asymptotically becomes independent. The fact that it will be not independent will precisely have to do with the fact that we are working in a deterministic setting. The end result that they asymptotically become independent will have to do with ergodicity.

Let $(M, \mu, \mathcal{M})$ be a probability space and $f : M \to M$ be a map of sufficient regularity (such as measurability which is as low as one can really imagine). The measure μ is said to be *f-invariant* if for all $A \in \mathcal{M}$,

$$\mu(f^{-1}(A)) = \mu(A)$$

and ergodic if

$$f^{-1}(A) = A \Rightarrow \mu(A) = 1 \text{ or } \mu(A) = 0$$

The invariance simply means that the dynamics leaves the measure invariant no more no less. Note that in the case f is just a measurable function, then for any measurable set A, it only makes sense to speak about $f^{-1}(A)$ and not $f(A)$ and moreover transformations of the base space induced a transformation on the measures with the inverse mapping. These are the reasons why inverse appears in the definition rather than the forward map f. In the case f has measurable inverse, then μ is f invariant iff for every measurable set A, $\mu(f(A)) = \mu(A)$.

Ergodicity means that the dynamics can not be decomposed into positive measure invariant sets. Here $f^{-1}(A) = \{x \in M \mid f(x) \in A\}$ is a set valued inverse map. Indeed if A satisfies $A = f^{-1}(A)$ then $f(A) = A$ and so $f(A^c) = A^c$ and therefore we can define $M = A \cup A^c$ such that, $f : A \to A$, $f : A^c \to A^c$ which decomposes the dynamics into two non-interacting sets.

2.3 Probabilistic Point of View of Dynamics

As we have said, the similarity of ergodic theorem to strong law of large numbers is no mistake. To understand this similarity, given a random observable $X : M \to \mathbb{R}$, define the sequence of random observables $X^k(p) = X \circ f^k(p)$. Then we get a sequence of random observables which by Birkhoff ergodic theorem satisfy the same conclusion of identically distributed independent random variables in the strong law of large numbers. These random variables that we have just defined, are they also independent and identically distributed? They are indeed identically distributed. Given any $I \subset \mathbb{R}$;

$$\{p \mid X \circ f^k(p) \in I\} = f^{-k}\{p \mid X \in I\}$$

and therefore by invariance

$$\mu(\{p \mid X \circ f^k(p) < r\}) = \mu(\{p \mid X < r\})$$

for all k, which means they are identically distributed.

But these variables are by no means independent. The intuition here is the following: if you are given that $X \circ f^k(p)$ has a certain value then you know where the initial point p can be and therefore you can fix the possible orbits that it can follow which fixes all the possible values that $X \circ f^m(p)$ assumes for all m. More precisely for instance let $I \subset \mathbb{R}$ and $A = X^{-1}(I)$, then

$$\mu(\ \{p \mid X^k(p) \in I\} \ \cap \ \{p \mid X^{k+m}(p) \in I\} \ ) = \mu(f^{-k}(A) \cap f^{-m-k}(A)))$$

If these random observables were independent we would require for all m

$$\mu(f^{-k}(A) \cap f^{-m-k}(A))) = \mu(A \cap f^{-m}(A)) = \mu(A)^2$$

It is obvious such a condition can not be satisfied generally for all A and m. For instance take a positive measure set B such that $f^{-1}(B) \neq B$, where the difference also has positive measure. Then there exists some positive measure set $A \subset B$ such that $f^{-1}(A)$ is disjoint from A. But then $\mu(A \cap f^{-m}(A)) = 0$ while $\mu(A)^2 > 0$. Therefore generally it is not expected that X^k defined above are independent. Thus determinism introduce dependences in the random observables we have defined, making strong law of large numbers inapplicable in this case.Yet we obtain a generalization of strong law of large numbers where ergodicity makes up for non-independence in the limit.

For another point of view, consider the characteristic functions χ_A for $A \subset M$. Then on a full measure set

$$\lim_{n\to\infty} \frac{1}{n} \sum_i \chi_A \circ f^i(p) \to \int \chi_A d\mu$$

The left hand side is the asymptotic frequency of visits of $f^i(p)$ to the set A while the right hand side is precisely the measure of A. On probabilistic terms, this says that "the chances that the orbit $\{f^i(p)\}$ falls in A is equal to its measure". Indeed also in probability theory, the measure of a subset $A \subset M$ tells how likely an experiment is likely to result in that set. Thus in some sense a typical orbit of an ergodic dynamical system in M behaves as if it is a sequence of random experiments with results in M. Or in other words as we have discussed in the dice example, if we randomly select points from M with probability μ and string them together to a sequence $(x_1, x_2, x_3, ...)$ then almost surely such a sequence and a a sequence of the form $(p, f^1(p), f^2(p), ...)$ asymptotically similar statistical behaviours.

2.4 Dynamical View of Probability

On the reverse direction, it is also possible to give an infinite sequence of probability experiments the structure of an ergodic dynamical system. Consider as before

- $M = S \times S \times ...$ (infinite product)
- $\mathcal{M} = 2^M$
- $\mu(A_1 \times A_2 \times ...) = m(A_1)m(A_2)...$ where $m : 2^M \to \mathbb{R}$ is given by $m(\{a_1, ..., a_k\}) = \frac{k}{6}$

Define the shift map $\sigma : M \to M$ by

$$\sigma((x_1, x_2, ...)) = (x_2, x_3, ...)$$

It is easy to check that since $m(S) = 1$, this leaves the measure μ invariant. One can also check that the measure μ is ergodic by observing that only invariant set is $M = S \times S \times S....$ and the empty set. Now again consider a subset of events $A \in \mathcal{S}$ and its characteristic function χ_A. Then by ergodic theorem applied to shift map, we get that the frequency of observing a result in the set of outcomes A is precisely equal to its measure $\mu(A)$.

Therefore this completes a bidirectional link between dynamical systems and probability theory. On one hand ergodic dynamical systems can be realized as a sequence of identically distributed but not independent random observables which become independent in the limit (via Birkhoff ergodic theorem) and on the other hand an infinite sequence of probability experiments can be seen as an ergodic dynamical system equipped with the shift map.

3 Central Limit Theorem and Decay of Corellations

Strong law of large numbers is not the only statistical property satisfied by sequence of identically distributed independent random observables. Central limit theorem is a much stronger description of their character. To describe this theorem we need some preliminary definitions.

3.1 Central Limit Theorem

Theorem 3.1. *(*Central Limit Theorem*) Let $(M, \mu, \mathcal{M})$ be a probability space and $\{X^k\}_{k=0}^{\infty}$ be a sequence of identically distributed and independent random observables with mean $\mathbb{E}(X)$ and variance σ. Then*

$$X^k \to_d X$$

where $\to_d$ denotes convergence in distribution and X is a normally distributed random observable with mean $\mathbb{E}(X)$ and variance σ.

To understand central limit theorem we need to know some terminology related to convergence of random variables in probability theory. Let $(M, \mu, \mathcal{M})$ be a probability space. For a given $\sigma > 0$, let

$$f(x) = e^{\frac{x^2}{2\sigma}}$$

We say that a random variable X has normal distribution with zero mean and σ variance if for all $r \in \mathbb{R}$,

$$\mu(\{p \mid X(p) < r\}) = \int_{-\infty}^{r} f d\ell$$

A sequence of random observables X^k are said to converge to X in distribution (which we denote as $X^k \to_d X$) if for all $r \in \mathbb{R}$,

$$\lim_{k \to \infty} \mu(\{p \mid X^k(p) < r\}) = \mu(\{p \mid X(p) < r\})$$

3.2 Decay of Corellations in Dynamical Systems

Theorem 3.2. *Let $(M, \mu, \mathcal{M})$ be a probability space and $f : M \to M$ a map such that μ is invariant and ergodic with respect to f. Given any random observable $X : M \to \mathbb{R}$ such that $X \in L_\infty$, X has polynomial decay of correlations and $\int_M X d\mu = 0$. Writing $X^i = X \circ f^i$, there exists a normally distributed random variable X with mean 0 and some variance σ such that:*

$$X^k \to_d X$$

Given a random observables $\phi \in L_\infty$, such that $\int_M \phi d\mu = 0$, it is said have decay of correlations of if there exists $\beta > 1$ and $C > 0$ such that for all random observables $\psi \in L_1$, and for all k

$$|\int_M \phi.(\psi \circ f^k) d\mu| \leq C \frac{1}{k^\beta} |\psi|_1$$

If for instance one uses characteristic functions $\phi = \chi_A, \psi = \chi_B$ then one gets that

$$\frac{\mu(A \cap f^{-k} B)}{\mu(B)} \leq K \beta^{-k}$$

Since $\mu(A) = 0$ the random observables χ_A and $\chi_B \circ f^k$ are independent if

$$\frac{\mu(A \cap f^{-k} B)}{\mu(B)} = 0$$

Therefore decay of correlations is also a way of saying that the dynamics makes the random observables independent asymptotically.

Then in the end what we obtain is similar to what has happened in the case of strong law of large numbers and ergodic theorem. The sequence of observables $X^k = X \circ f^k$ are identically distributed but not independent. Yet decay of correlations property implies that asymptotically they display some characteristic behaviour of sequence of identically distributed and independent distributions, that is they converge distributionally to a normally distributed random observable.