Notes on Probability and Measure theory

Naina Praveen
May, 2022

Contents

Acknowledgements

These are some notes I took while preparing for my undergraduate thesis at Ashoka back in 2022, and are probably more geared towards the particular problem I was working on. These notes unfortunately do not have any citations because I didn't think too much about it back then (bad practice, I have learnt better since). The parts on probability theory are taken primarily from my course notes on lectures provided by Dr. Kumarjit Saha and Prof. Sandeep Juneja. The parts on measure theory are taken from my course notes on the lecture provided by Prof. Rajendra Bhatia by the same name. Any errors in these notes are mine alone.

1. Probability through a Measure Theory perspective

1.1 Probability Measures

Measure Theory in simple terms is the theory about the distribution of mass over a set Ω. Informally speaking, if the mass is uniformly distributed and the set Ω is a Euclidean space $\mathbb{R}^k$ (which corresponds to length in $\mathbb{R}$, area in $\mathbb{R}^2$, volume in $\mathbb{R}^3$ and so on and so forth), then the measure is called the Lebesgue measure. Probability theory is concerned with when Ω is the sample space of a random experiment and the total mass is one.

The reason the measure theoretic framework was introduced to formalise the language of probability was to overcome certain obstacles while dealing with infinite sample spaces. Consider the example of uniform distribution on the interval $[0, 1]$. What is the probability that a ball thrown on this interval will land on a particular point? It is zero, which is to say it is impossible for the ball to land at any particular point. But if the probability of the ball landing at any particular point is zero, then how does one explain that the probability equals 1. Phenomena like discussed above seem rather counter-intuitive, demanding a more formal framework to understand probability theory, and measure theory is one such natural framework to do the same.

Suppose one has a field Ω of general shape filled with grass and one sprays 1 kg of pesticide K on it. If one wants to measure the amount of pesticide K in each possible subset $\mathcal{B}$ of Ω, one possible way to approach this is to measure the amount of pesticide K on a class of subsets of "nice shapes"– say triangles or squares (assuming one has the tools to do so) – and then use this to calculate the amount of pesticides for regions of general shape through some kind of limiting approximation. Let $\mathcal{B}$ denote the class of "standard shape" subsets for Ω for which one has the tools to obtain such a measure, and let $\mu(B)$ denote the amount of pesticide in $B \in \mathcal{B}$. Using measure $\mu(B)$ for $B \in \mathcal{B}$ it is natural to ask about of pesticides for more general sets. Let $\mathcal{B}'$ be a bigger set containing $\mathcal{B}$ such that for each $B \in \mathcal{B}'$, the notation $\mu(B)$ is defined and denotes the amount of pesticides in B. It is reasonable to assume that the following properties hold for $\mathcal{B}'$ and $\mu(.)$ hold:

Properties for $\mathcal{B}'$

(i) $A \in \mathcal{B}'$ implies $A^c \in \mathcal{B}'$ (if one can calculate the amount of pesticide in A, one can then calculate the amount of pesticide in A^c).

(ii) $A_1, A_2 \in \mathcal{B}'$ implies $A_1 \cup A_2 \in \mathcal{B}'$ (if one can measure the amount of pesticide on A_1 and A_2, one can do so on $A_1 \cup A_2$ as well).

(iii) If $\{A_n : n \geq 1\} \subseteq \mathcal{B}'$ and $A_n \subset A_{n+1} \in \mathcal{B}'$ for all $n \geq 1$, then $\lim_{n\to\infty} A_n \equiv \cup_{n=1}^{\infty} A_n \in \mathcal{B}'$ (if one can measure the amount of chemical on each A_n for each $n \geq 1$ for an *increasing sequence* of sets then one can do so on the limit of A_n).

Properties of $\mu(.)$

(i) $\mu(A) \geq 0$ for all $A \in \mathcal{B}'$ (the amount of pesticide on a set cannot be non-negative).

(ii) If $A_1, A_2 \in \mathcal{B}'$ such that $A_1 \cap A_2 = \varnothing$, then $\mu(A_1 \cup A_2) = \mu(A_1) + \mu(A_2)$ (the amount of pesticides on union of two disjoint subsets is equal to sum of the amount on each of the two sets).

(iii) If $\{A_n : n \geq 1\} \subset \mathcal{B}'$, are such that $A_n \subseteq A_{n+1}$ for all n, then

$$\mu(\lim_{n\to\infty} A_n) = \mu(\cup_{n=1}^{\infty} A_n) = \lim_{n\to\infty} \mu(A_n)$$

(if we can approximate a set A by an increasing sequence $\{A_n\}_{n\geq 1}$ from $\mathcal{B}'$ as $A = \cup_{n=1}^{\infty} A_n$, then $\mu(A) = \lim_{n\to\infty} \mu(A_n)$. This property is referred to *monotone continuity from below*.)

The last assumption is what guarantees that different approximations lead to the same limit. These natural assumptions lead to developing a rich and widely practical theory called measure theory.

A triplet $(\Omega, \mathcal{B}, \mu)$ is called a **measure space**. If we further stipulate a condition that $\mu(\Omega) = 1$, we have a **probability space**. The assumptions on $\mathcal{B}$ and μ lead to the following definitions:

Definition 1.1. Let Ω be a any set and let $P(\Omega)$ denote its power set. Then $\mathcal{B} \subseteq P(\Omega)$ is said to be a **σ- algebra** (or **σ-field**) if the following properties hold:

(i) $\varnothing \in \mathcal{B}, \Omega \in \mathcal{B}$.

(ii) $\mathcal{B}$ is closed under complementation: If $A \in \mathcal{B}$, then so is its complement, $\Omega \setminus A$.

(iii) $\mathcal{B}$ is closed under countable unions: If $A_1, A_2, A_3 \ldots$ are in $\mathcal{B}$, then so is $A = A_1 \cup A_2 \cup A_3 \ldots$.

By De Morgan's laws, it follows that $\mathcal{B}$ is also closed under countable intersections. Elements of the σ- algebra are called **measurable sets**. The largest possible σ- algebra on Ω is $P(\Omega)$, while the smallest $\sigma-$algebra is $\{\Omega, \varnothing\}$.

Definition 1.2. Let Ω be a set, and $\mathcal{B}$ a $\sigma-$algebra on Ω. A function $\mu : \mathcal{B} \to [0,1]$ is called a **probability measure** if it satisfies the following properties:

(i) $\mu(\Omega) = 1$ and $\mu(\varnothing) = 0$.

(ii) if $A_i \in \mathcal{B}$, is a countable sequence of pairwise disjoint sets, i.e., $A_i \cap A_j = \emptyset$ for all $1 \leq i < j < \infty$, then

$$\mu\left(\bigcup_i A_i\right) = \sum_i \mu(A_i)$$

It is straightforward to show a probability measure μ on $\mathcal{B}$ satisfies $\mu(A) \geq \mu(B)$ for all $A, B \in \mathcal{B}$ with $B \subseteq A$.

Consider the following examples of σ-algebras on the same set $\Omega = \{1,2,3\}$

$$\mathcal{B}_1 := \{\{1\},\{2,3\},\Omega,\varnothing\} \quad ; \qquad \mathcal{B}_2 := \{\{1,2\},\{3\},\Omega,\varnothing\}.$$

Natural questions on may ask is if the intersection or the unions of σ-algebras is a σ-algebra? The answer is yes for the former and no for the latter, i.e., the intersections of $\sigma-$algebras is a $\sigma-$algebra, while the union need not be a $\sigma-$algebra as the above example illustrates. The concept of $\sigma-$algebras plays an important role in probability theory. In many instances, one is given an arbitrary collection of subsets of Ω and one would like to consider the *smallest* class of subsets that is a $\sigma-$algebra containing the given collection of sets. This yields the following definition:

Definition 1.3. If $\mathcal{A}$ is a class of subsets of Ω, then the **$\sigma-$algebra generated by** $\mathcal{A}$ and denoted as $\sigma\langle\mathcal{A}\rangle$ is defined as

$$\sigma\langle\mathcal{A}\rangle = \bigcap_{\mathcal{F}\in\mathcal{I}(\mathcal{A})} \mathcal{F}$$

where $\mathcal{I}(\mathcal{A}) := \{\mathcal{F} : \mathcal{A} \subset \mathcal{F} \text{ and } \mathcal{F} \text{ is a } \sigma\text{-algebra on } \Omega\}$ is the collection of all $\sigma-$algebras containing the class $\mathcal{A}$.

In other words, $\sigma\langle\mathcal{A}\rangle$ is the smallest σ-algebra containing $\mathcal{A}$. An important collection of $\sigma-$algebras are those generated by open sets of topological spaces.

Definition 1.4. The **Borel $\sigma-$algebra** on a topological space $\mathbb{S}$ is defined as the $\sigma-$algebra generated by the collection of open sets in $\mathbb{S}$. Such sets equipped with the Borel $\sigma-$algebra are called **Borel sets**.

In particular, consider the case when $\mathbb{S} = \mathbb{R}$. Then, the Borel $\sigma-$algebra is defined as $\mathcal{B}(\mathbb{R}) := \sigma\langle\{A : A \text{ is open in } \mathbb{R}\}\rangle$. The same $\sigma-$algebra $\mathcal{B}(\mathbb{R})$ is equivalently generated by each of the following classes too:

$$\mathcal{O}_1 := \{(a, b) : -\infty \leq a < b \leq \infty\}$$

$$\mathcal{O}_2 := \{(-\infty, a) : a \in \mathbb{R}\}$$

$$\mathcal{O}_3 := \{(a, b) : a, b \in \mathbb{Q}, a < b\}$$

$$\mathcal{O}_4 := \{(-\infty, a) : a \in \mathbb{Q}\}$$

It is important to observe that the Borel $\sigma-$algebra $\mathcal{B}(\mathbb{R})$ is fairly huge. Practically any 'natural' set $A \subset \mathbb{R}$ that one can think of is Borel. On the other hand it is *much smaller* than the power set $P(\mathbb{R})$. It is non-trivial to assume existence of a probability measure satisfying the properties mentioned in 1.2 defined on $\mathcal{B}(\mathbb{R})$. Theorem 1.1 provides a general recipe and suggests existence of plenty of such probability measures.

1.2 Random variables and Expectations

1.2.1 Measurable functions and random variables

An important concept in probability theory is that of random variables, which are essentially functions from the sample space into $\mathbb{R}$. One would like to look at this concept through the lens of measure theory, but, before talking about random variables one needs to understand functions on measure spaces. Suppose there are two sets X and Y and a function $f : X \to Y$. Definition 1.3 talks about $\sigma-$algebras generated by a class of subsets of Ω. Now we will talk about the minimal $\sigma-$algebra under which makes the function measurable.

Definition 1.5 (Measurable function)**.** Let $(X, \mathcal{F}_X)$ and $(Y, \mathcal{F}_Y)$ be measurable spaces, i.e., X and Y are sets equipped with respective σ-algebras $\mathcal{F}_X$ and $\mathcal{F}_X$. A function $f : X \to Y$ is said to be $\mathcal{F}_Y$ **measurable** if for every $E \in \mathcal{F}_Y$ the pre-image of E under f is in $\mathcal{F}_X$. In other words, for all $E \in \mathcal{F}_Y$

$$f^{-1}(E) := \{x \in X \mid f(x) \in E\} \in \mathcal{F}_X.$$

Clearly, for a function f from a set X to a set Y the notion of measurability depends on the σ-algebra $\mathcal{F}_Y$.

Definition 1.6. A function $f : (X, \mathcal{F}_X) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is said to be **Borel measurable function** if $f^{-1}(E) \in \mathcal{F}_X$ for all $E \in \mathcal{B}(\mathbb{R})$, where $\mathcal{B}(\mathbb{R})$ denotes the $\sigma-$algebra on $\mathbb{R}$.

Now, one has all the necessary background needed to define a random variable.

Definition 1.7 (Random variable)**.** A **random variable** X on a probability space $(\Omega, \mathcal{B}, \mathbb{P})$ is a Borel measurable function from Ω to $\mathbb{R}$.

It is important to observe that for any random variable X defined on a probability space $(\Omega, \mathcal{B}, \mathbb{P})$, the collection $\{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\} \subseteq \mathcal{B}$ gives a σ-field and it is the smallest σ-field w.r.t. which X is 'Borel' measurable. This σ-field is denoted as $\sigma(X)$. We say that $\sigma(X)$ is the smallest σ-field w.r.t. which X is measurable.

In practise, it is often more convenient to ignore the underlying probability space of a random variable in favour of its distribution function.

Definition 1.8. The **distribution function** $F_X : \mathbb{R} \to [0,1]$ of a random variable X is defined as

$$F(x) = \mathbb{P}\{\omega : X(\omega) \leq x\}.$$

It is a map that is increasing (i.e., $a < b$ implies $F(a) \leq F(b)$) and **right-continuous** (i.e., $\lim_{x \to x_0^+} F(x) = F(x_0)$).

The random variable X is said to be **absolutely continuous** iff there exists a non-negative real-valued Borel measurable function f on $\mathbb{R}$ such that

$$F(x) = \int_{-\infty}^{x} f(t)\, dt \quad \text{for all } x \in \mathbb{R}.$$

If such an f exists, it is called the **density function** of X.

As mentioned in Section 1.1, one can obtain a number of probability measures on $\mathcal{B}(\mathbb{R})$ with the help of the following theorem.

Theorem 1.1. *Let F be a distribution function on $\mathbb{R}$ and let $\mu(a,b] = F(b) - F(a)$, $a < b$. Then, there is a unique extension of μ to a probability measure on $\mathbb{R}$.*

While we do not provide a proof of the theorem, essentially what the theorem does is extend μ to a finitely additive set function on the field $\mathcal{F}_o(\mathbb{R})$ of finite disjoint unions of right-semiclosed intervals. Then, one can show that μ is countably additive on $\mathcal{F}_0(\mathbb{R})$ and an extension theorem (called the Caratheodory extension theorem) extends μ to $\mathcal{B}(\mathbb{R})$.

In particular, let $F(x) = x$. The measure obtained as such is known as the **Lebesgue measure**, and coincides with the length of any interval. In higher dimensions of 2 and 3, the Lebesgue measure corresponds to area and volume of a set.

1.2.2 Lebesgue integrals and Expectation

In probability theory, expectation is a generalisation of the notion of a weighted average. Informally, it makes precise the notion of 'average value'. While its formulations in the discrete and continuous setting are well known, in the axiomatic foundation for probability provided by measure theory, the expectation is given by Lebesgue integration.

Unfortunately for the sake of brevity, we shall be omitting construction of the Lebesgue integral in this paper. However, this section shall ensue the necessary definitions one would need in order to have a working idea of how the Lebesgue integral is defined.

To begin with, the Riemann integral is defined in terms of approximation by step functions. In some sense, one could say that step functions acts as 'building blocks' of Riemann integrable functions. In order to define the Lebesgue integral, one way to proceed is to consider a generalization of step functions called 'simple functions' (see Definition 1.9). A function will be Lebesgue integrable if it can be approximated by these simple functions in some appropriate way.

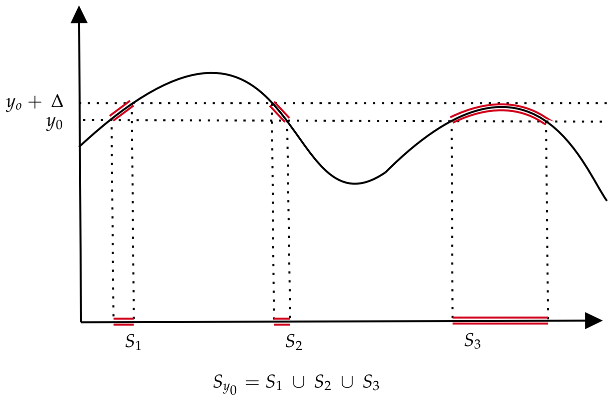
Definition 1.9. Let $(\mathbb{R}, \mathcal{B}, \mu)$ be a measure space. $\phi : \mathbb{R} \to [0, \infty)$ is a **simple function** if it is given by

$$\phi = \sum_{i=1}^{N} c_i I_{E_i}$$

where $c_i \geq 0$, $E_i \in \mathcal{B}$ and I_{E_i} is the indicator function for E_i. The integral of ϕ with respect to μ is

$$\int \phi d\mu = \sum_{i=1}^{N} c_i \mu(E_i).$$

Simple functions that lie directly underneath a given function f can be constructed by partitioning the range of f into a finite number of layers. The intersection of the graph of f with a layer identifies a set of intervals in the domain of f, which, taken together, is defined to be the pre-image of the lower bound of that layer, under the simple function. In this way, the partitioning of the range of f implies a partitioning of its domain. The integral of a simple function is found by summing, over these (not necessarily connected) subsets of the domain, the product of the measure of the subset and its image under the simple function (the lower bound of the corresponding layer). The following diagram helps illustrate the idea better.



Thus, for a non-negative function f that is Lebesgue measureable, its Lebesgue integral is defined as the (possibly infinite) quantity

$$\int_{\mathbb{R}} f d\mu := \sup\left\{\int_{\mathbb{R}} \phi d\mu : \phi \leq f, \phi \text{ simple}\right\}$$

In order to account for signed fucntions, the following provision is made. Given a Lebesgue measurable function $f : \mathbb{R} \to \mathbb{R}$, consider the disjoint subsets

$$S_+ := f^{-1}([0, \infty))$$
$$S_- := f^{-1}((-\infty, 0))$$

These functions are called the positive and negative parts of f, respectively. By definition, $f = f_+ - f_-$ and f_+, f_-. Also note that $|f| = f_+ + f_-$.

Definition 1.10. Let $f : \mathbb{R} \to \mathbb{R}$ be a measurable function. If one of $\int_{\mathbb{R}} f_+ d\mu$ or $\int_{\mathbb{R}} f_- d\mu$ is finite, the **Lebesgue integral** of f is defined to be the quantity:

$$\int_{\mathbb{R}} f d\mu = \int_{\mathbb{R}} f_+ d\mu - \int_{\mathbb{R}} f_- d\mu$$

f is said to be **Lebesgue integrable** if both $\int_{\mathbb{R}} f_+ d\mu$ and $\int_{\mathbb{R}} f_- d\mu$ are finite. Equivalently, since $|f| = f_+ + f_-$, f is Lebesgue integrable if $\int_{\mathbb{R}} |f| dm$ is finite, in which case $\left|\int_{\mathbb{R}} f d\mu\right| \leq \int_{\mathbb{R}} |f| d\mu$ by the triangle inequality. The Lebesgue integral satisfies the same properties as one expects with the Riemann integral. A quick remark to be made at this point is that the Lebesgue definition of integrals makes it possible to calculate the integral of a larger class of function as opposed to the Riemann integral, in fact the Riemann integral is subsumed by the Lebesgue integral. Lastly, while the description given above was defined over $\mathbb{R}$, the results can be extended in further generality over any measure space. One can now proceed to define expectation of a random variable in a more general sense.

Definition 1.11. The **expectation** of the random variable X on the probability space $(\Omega, \mathcal{B}, \mathbb{P})$ is the quantity

$$\mathbb{E}(X) := \int_{\Omega} X \, d\mathbb{P} = \int_{\Omega} X(\omega)\mathbb{P}(d\omega).$$

Th usual properties of expectation apply in this abstract definition as well. In general, it is not the case that $\mathbb{E}[X_n] \to \mathbb{E}[X]$ even if $X_n \to X$ pointwise. Thus, one cannot interchange limits and expectation, without additional conditions on the random variables. Some convergence theorems of expectation without proof are stated below:

- **Monotone convergence theorem:** Let $\{X_n : n \geq 0\}$ be a sequence of random variables, with $0 \leq X_n \leq X_{n+1}$ (a.s) for each $n \geq 0$. Furthermore, let $X_n \to X$ pointwise. Then, the monotone convergence theorem states that

$$\lim_n \mathbb{E}[X_n] = \mathbb{E}[X].$$

- **Fatou's lemma:** Let $\{X_n \geq 0 : n \geq 0\}$ be a sequence of non-negative random variables. Fatou's lemma states that $\mathrm{E}\left[\lim_n \inf X_n\right] \leq \liminf_n \mathrm{E}\left[X_n\right]$.

- **Dominated convergence theorem:** Let $\{X_n : n \geq 0\}$ be a sequence of random variables. If $X_n \to X$ pointwise (a.s.), $|X_n| \leq Y \leq +\infty$ (a.s.), and $\mathrm{E}[Y] < \infty$. Then, according to the dominated convergence theorem,

  (i) $\mathrm{E}|X| \leq \mathrm{E}[Y] < \infty$.

  (ii) $\lim_n \mathrm{E}\left[X_n\right] = \mathrm{E}[X]$.

  (iii) $\lim_n \mathrm{E}\left|X_n - X\right| = 0$.

1.3 Conditional Expectations

1.3.1 Introduction

Let X be a r.v. defined on a probability space $(\Omega, \mathcal{F}, P)$. Without loss of generality, assume that X has finite expectation. Recall the definition of $\sigma(X) \subseteq \mathcal{F}$.

Consider $\mathcal{G} \subseteq \sigma(X)$. One can think of $\mathcal{G}$ as the *partial* information one has at their disposal – as, for each $A \in \mathcal{G}$, one knows whether or not A has occurred. Given the information about the σ-field $\mathcal{G} \subseteq \mathcal{F}$, the conditional expectation of X given $\mathcal{G}$, denoted by $\mathbb{E}(X \mid \mathcal{G})$ is then the "best guess" of X in the following sense:

Definition 1.12. The **conditional expectation of** X **given** $\mathcal{G}$ denoted as $\mathbb{E}(X \mid \mathcal{G})$, is a $\mathcal{G}$ measurable random variable Y such that for all $A \in \mathcal{G}$, $\int_A X dP = \int_A Y dP$.

Any $\mathcal{G}$ measurable random variable Y satisfying Definition 1.12 is said to be a **version of** $\mathbb{E}(X \mid \mathcal{G})$ a.s. The first thing to be settled is that the conditional expectation exists and is unique.

Existence: The general proof of existence of such a random variable relies on the *Radon-Nikodym Theorem*, the proof for which will not be given in this section.

However, if we work with random variables that are square integrable, i.e., $\mathbb{E}(X^{)} < \infty$, then one can provide a "geometric intuition" of the same. Let $\mathcal{L}^2(\mathcal{F}_0) := \{Y \in \mathcal{F}_0 : \mathbb{E}(Y^{)}] < \infty\}$. Then one can show that $\mathcal{L}^2(\mathcal{F}_0)$ forms a Hilbert space and $\mathcal{L}^2(\mathcal{F})$ forms a closed subspace. Then $\mathbb{E}(X \mid \mathcal{F})$ is the random variable $Y \in \mathcal{F}$ that minimizes the **mean square error** i.e., $\mathbb{E}((X - Y)^2)$. In this case, one can view $\mathbb{E}(X \mid \mathcal{F})$ as the projection of X onto $\mathcal{L}^2(\mathcal{F})$, i.e, the point closest to X in the subspace. The proof that such a minimiser exists is non-trivial and relies on orthogonal projections.

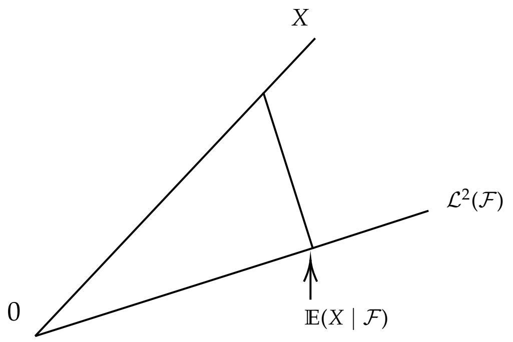

X
$\mathcal{L}^2(\mathcal{F})$
0
$\mathbb{E}(X \mid \mathcal{F})$

Figure 1.1: Conditional Expectation as a projection onto $\mathcal{L}^2$.

Uniqueness: If Y' also satisfies Definition 1.12 then

$$\int_A Y d\mathbb{P} = \int_A Y' d\mathbb{P} \quad \text{for all } A \in \mathcal{F}$$

Taking $A = \{Y - Y' \geq \varepsilon > 0\}$, notice that

$$0 = \int_A (X - X) d\mathbb{P} = \int_A (Y - Y') d\mathbb{P} \geq \varepsilon \mathbb{P}(A)$$

so $\mathbb{P}(A) = 0$. Since this holds for all ε, $Y \leq Y'$ a.s., and interchanging the roles of Y and Y', $Y = Y'$ a.s. One can use a similar argument to show that if $X_1 = X_2$ on $B \in \mathcal{F}$ a.s., then, $\mathbb{E}(X_1 \mid \mathcal{F}) = \mathbb{E}(X_2 \mid \mathcal{F})$ a.s. on B.

Following are some examples of conditional expectation:

Example 1. If $X \in \mathcal{F}$, then $\mathbb{E}(X \mid \mathcal{F}) = X$; that is, if we know X, then our "best guess" is X itself. Since X is $\mathcal{F}$ measureable it trivially satisfies the property mentioned in Defintion 1.12. A special case of this example is $X = c$, where c is a constant.

Example 2. At the other extreme from perfect information is no information. Suppose X is independent of $\mathcal{F}$. In this case, $\mathbb{E}(X \mid \mathcal{F}) = \mathbb{E}(X)$; that is, if one doesn't know anything about X, then the best guess is the mean $\mathbb{E}(X)$. To check the definition, note that $\mathbb{E}(X) \in \mathcal{F}$ so it is $\mathcal{F}$ measureable. To verify the property given in Definition 1.12, observe that if $A \in \mathcal{F}$, then since X and $\mathbf{1}_A \in \mathcal{F}$ are independent,

$$\int_A X d\mathbb{P} = \mathbb{E}\left(X \mathbf{1}_A\right) = \mathbb{E}(X)\mathbb{E}(\mathbf{1}_A) = \int_A \mathbb{E}(X) d\mathbb{P}.$$

The important thing to note here is that this is not a constructive definition; one is merely given the required property that a conditional expectation must satisfy.

1.3.2 Properties of Conditional Expectations

The following are certain properties of conditional expectations that are proved below: *Good to replace $\mathcal{F}$ by $\mathcal{G}$.*

Proposition 1.2. *(a) Conditional expectation is linear:*

$$\mathbb{E}(aX + Y \mid \mathcal{F}) = a\mathbb{E}(X \mid \mathcal{F}) + \mathbb{E}(Y \mid \mathcal{F}) \tag{1.1}$$

(b) Monotonicity: If $X \leq Y$, then

$$\mathbb{E}(X \mid \mathcal{F}) \leq \mathbb{E}(Y \mid \mathcal{F}) \tag{1.2}$$

Proof. To prove (a), one needs to check that the right-hand side is a version of the left. It clearly is $\mathcal{F}$-measurable as the individual conditional expectation random variables are $\mathcal{F}$ measurable. To check property **??**, observe that if $A \in \mathcal{F}$, then by linearity of the integral and the defining properties of $\mathbb{E}(X \mid \mathcal{F})$ and $\mathbb{E}(Y \mid \mathcal{F})$,

$$\begin{aligned}
\int_A \{a\mathbb{E}(X \mid \mathcal{F}) + \mathbb{E}(Y \mid \mathcal{F})\} d\mathbb{P} &= a \int_A \mathbb{E}(X \mid \mathcal{F}) d\mathbb{P} + \int_A \mathbb{E}(Y \mid \mathcal{F}) d\mathbb{P} \\
&= a \int_A X d\mathbb{P} + \int_A Y d\mathbb{P} \\
&= \int_A (aX + Y) d\mathbb{P}
\end{aligned}$$

which proves Eq 1.1.

To prove (b), using the definition

$$\int_A \mathbb{E}(X \mid \mathcal{F}) d\mathbb{P} = \int_A X d\mathbb{P} \leq \int_A Y d\mathbb{P} = \int_A \mathbb{E}(Y \mid \mathcal{F}) dP$$

Letting $A = \{\mathbb{E}(X \mid \mathcal{F}) - \mathbb{E}(Y \mid \mathcal{F}) \geq \epsilon > 0\}$, it is easy to observe that the indicated set has probability 0 for all $\epsilon > 0$, and thus Eq 1.2 follows through. □

Proposition 1.3. *If $\mathcal{F} \subset \mathcal{G}$ and $\mathbb{E}(X \mid \mathcal{G}) \in \mathcal{F}$, then $\mathbb{E}(X \mid \mathcal{F}) = \mathbb{E}(X \mid \mathcal{G})$*

Proof. By assumption, $\mathbb{E}(X \mid \mathcal{G}) \in \mathcal{F}$ so it is $\mathcal{F}$ measureable. To check the other part of the definition, note that if $A \in \mathcal{F} \subset \mathcal{G}$, then

$$\int_A X d\mathbb{P} = \int_A \mathbb{E}(X \mid \mathcal{G}) d\mathbb{P}.$$

□

Proposition 1.4. *If $\mathcal{F}_1 \subset \mathcal{F}_2$, then*

(i) $\mathbb{E}\left(\mathbb{E}\left(X \mid \mathcal{F}_1\right) \mid \mathcal{F}_2\right) = \mathbb{E}\left(X \mid \mathcal{F}_1\right)$,

(ii) $\mathbb{E}\left(\mathbb{E}\left(X \mid \mathcal{F}_2\right) \mid \mathcal{F}_1\right) = \mathbb{E}\left(X \mid \mathcal{F}_1\right)$.

In words, the smaller σ-field always gives the better approximation.

Proof. Notice that $\mathbb{E}\left(X \mid \mathcal{F}_1\right) \in \mathcal{F}_2$, then (i) follows from Example 1. To prove (ii), notice that $\mathbb{E}\left(X \mid \mathcal{F}_1\right) \in \mathcal{F}_1$, and if $A \in \mathcal{F}_1 \subset \mathcal{F}_2$, then

$$\int_A \mathbb{E}\left(X \mid \mathcal{F}_1\right) d\mathbb{P} = \int_A X d\mathbb{P} = \int_A \mathbb{E}\left(X \mid \mathcal{F}_2\right) d\mathbb{P}.$$

□

Proposition 1.5 (Tower property). *Let X be an integrable random variable and let $\mathcal{G} \subset \mathcal{F}$. Then*

$$\mathbb{E}\left(\mathbb{E}\left(X \mid \mathcal{F}\right) \mid \mathcal{G}\right) = \mathbb{E}\left(X \mid \mathcal{G}\right).$$

Proof. Let A be an arbitrary event in $\mathcal{G}$. Then, since $A \in \mathcal{F}$, by the definition of conditional expectation applied first to $\mathcal{F}$ and then to $\mathcal{G}$, implies

$$\mathbb{E}\left(\mathbf{1}_A \mathbb{E}\left(X \mid \mathcal{F}\right)\right) = \mathbb{E}\left(\mathbf{1}_A X\right) = \mathbb{E}\left(\mathbf{1}_A \mathbb{E}\left(X \mid \mathcal{G}\right)\right).$$

Thus, $\mathbb{E}\left(X \mid \mathcal{G}\right)$ satisfies Definition 1.12 with X replaced by $\mathbb{E}\left(X \mid \mathcal{F}\right)$. Further, as $\mathbb{E}\left(X \mid \mathcal{G}\right)$ is $\mathcal{G}$-measurable, it must equal $\mathbb{E}\left(\mathbb{E}\left(X \mid \mathcal{F}\right) \mid \mathcal{G}\right)$. □

Following are some additional properties of conditional expectation that are provided without proof.

- If X is $\mathcal{F}$-measurable, then $\mathbb{E}\left(X \mid \mathcal{F}\right) = X$. If X is $\mathcal{F}$-measurable, then $\mathbb{E}\left(XY \mid \mathcal{F}\right) = X\mathbb{E}\left(Y \mid \mathcal{F}\right)$.
- If X is independent of $\sigma(Y, \mathcal{F})$, then $\mathbb{E}\left(XY \mid \mathcal{F}\right) = \mathbb{E}\left(X\right)\mathbb{E}\left(Y \mid \mathcal{F}\right)$. Note that this is not necessarily the case if X is only independent of $\mathcal{F}$ and of Y.
- If X, Y are independent, $\mathcal{G}, \mathcal{F}$ are independent, X is independent of $\mathcal{F}$ and Y is independent of $\mathcal{G}$, then $\mathbb{E}\left(\mathbb{E}\left(XY \mid \mathcal{G}\right) \mid \mathcal{F}\right) = \mathbb{E}\left(X\right)\mathbb{E}\left(Y\right) = \mathbb{E}\left(\mathbb{E}\left(XY \mid \mathcal{F}\right) \mid \mathcal{G}\right)$.
- For random variables X, Y, Z we have $\mathbb{E}\left(\mathbb{E}\left(X \mid Y, Z\right) \mid Y\right) = \mathbb{E}\left(X \mid Y\right)$.
- **Law of total expectation:** $\mathbb{E}\left(\mathbb{E}\left(X \mid \mathcal{F}\right)\right) = \mathbb{E}\left(X\right)$.
- **Monotone convergence:** If $0 \leq X_n \uparrow X$ then $\mathbb{E}\left(X_n \mid \mathcal{F}\right) \uparrow \mathbb{E}\left(X \mid \mathcal{F}\right)$.

1.4 Filtrations

In much of probability, especially when conditional expectation is involved, one is concerned with sets that represent only part of all the possible information that can be observed. This partial information can be characterized with a smaller $\sigma-$algebra which is a subset of the principal $\sigma-$algebra; it consists of the collection of subsets relevant only to and determined only by the partial information. The following example helps illustrate this idea –

Imagine two people A and B are betting on a game that involves flipping a coin repeatedly and observing whether it comes up Heads (H) or Tails (T). Since A and B are each infinitely wealthy, there is no limit to how long the game can last. This means the sample space Ω must consist of all possible infinite sequences of H or T:

$$\Omega = \{H, T\}^\infty = \{(x_1, x_2, x_3, \dots) : x_i \in \{H, T\}, i \geq 1\}$$

However, after n flips of the coin, A wants to determine or revise their betting strategy in advance of the next flip. The observed information at that point can be described in terms of the $2n$ possibilities for the first n flips. Formally, since A used subsets of Ω, this is codified as the $\sigma-$algebra

$$\mathcal{F}_n = \{A \times \{H, T\}^\infty : A \subset \{H, T\}^n\}$$

Observe that then

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \cdots \subset \mathcal{F}_\infty$$

where $\mathcal{F}_\infty$ is the smallest $\sigma-$algebra containing all the others. Each of these smaller $\sigma-$algebras are called **sub-$\sigma-$algebras**.

In the following we will introduce the notion of the canonical/natural filtration corresponding to a (discrete time) stochastic process $\{X_n : n \in \mathbb{N}\}$. A discrete time stochastic process should be described as an infinite collection jointly distributed on a common probability space indexed by set of natural numbers.

This concept of partial information is also particularly useful in the theory of stochastic processes.

Definition 1.13. Let $(X_n)_{n\in\mathbb{N}}$ be a stochastic process on the probability space $(\Omega, \mathcal{B}, P)$. Then

$$\mathcal{F}_n := \sigma\left(X_k \mid k \leq n\right)$$

is a σ-algebra and $\mathbb{F} = (\mathcal{F}_n)_{n\in\mathbb{N}}$ is a filtration. Here $\sigma\left(X_k \mid k \leq n\right)$ denotes the σ-algebra generated by the random variables $X_1, X_2, \dots, X_n$ which is the smallest σ-algebra w.r.t. which the random variables $X_1, X_2, \dots, X_n$ are measurable. It is not difficult to observe that $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$. This is known as the **natural/canonical filtration** for the stochastic process $\{X_n : n \in \mathbb{N}\}$.

2. Markov Chains

2.1 Introduction

Markov chains are among the most important stochastic processes. They are stochastic processes for which the description of the present state entirely captures all the information that could influence the future evolution of the process.

Predicting traffic flows, communications networks, genetic issues, and queues are examples where Markov chains can be used to model performance. Devising a physical model for these chaotic systems would be impossibly complicated but doing so using Markov chains makes their understanding a lot more simpler. In this chapter, we will mostly deal with *finite* state space Markov Chains. Let us for this chapter denote our default discrete state space $\mathcal{S} := \{1, \ldots, d\}$

Definition 2.1 (Markov Chains)**.** A discrete time stochastic process $\{X_n : n \in \mathbb{N}\}$ is said to have **Markov property** or called a **Markov chain** if

$$X_{n+1} \Big| (X_n = x, X_{n-1}, \cdots) \overset{d}{=} X_{n+1} \Big| X_n = x$$

which says that the conditional distribution of X_{n+1} given the entire past is same as the conditional distribution of X_{n+1} given the present X_n.

Further, a Markov chain is said to be **time homogeneous** if

$$X_{n+1} \Big| X_n = x \overset{d}{=} X_1 \Big| X_0 = x.$$

In what follows, unless specified otherwise, we will deal with time homogeneous Markov chains only. The following are some examples of Markov chains:

Example 3. A game of snakes and ladders or any other game whose random moves depend only on the present state of the game and an independent r.v. (e.g., outcome of a die here) is a Markov chain. The only thing that matters is the current state of the board – the next state of the board depends on the current state, and the next roll of the dice. It does not depend on how things got to their current state.

Example 4. A classic example of Markov chains are random walks. An elementary example of a random walk is the random walk on the integer number line $\mathbb{Z}$ which starts at 0, and at each step moves $+1$ or -1 with equal probability. Formally, Take independent random variables $Z_1, Z_2, \ldots$ where each variable is either 1 or -1, with probability $1/2$ for either value, and set $S_0 = 0$ and $S_n = \sum_{j=1}^{n} Z_j$. Then S_n represents position of the random walker on $\mathbb{Z}$at time n.

Example 5. The probabilities of weather conditions (modeled as either rainy or sunny), given the weather on the preceding day, can be represented by a transition matrix:

$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}$$

The matrix P represents the weather model in which a sunny day is 90% likely to be followed by another sunny day, and a rainy day is 50% likely to be followed by another rainy day. The columns can be labelled "sunny" and "rainy", and the rows can be labelled in the same order. P_{ij} is the probability that, if a given day is of type i, it will be followed by a day of type j.

For a discrete state space $\mathcal{S}$ and $i, j \in \mathcal{S}$, we define P_{ij} as $\mathbb{P}(X_1 = j \mid X_0 = i)$. The following observations can then be made:

- For each $i \in \mathcal{S}$ with $\mathbb{P}(X_0 = i) > 0$, the collection $\left\{P_{ij} : j \in \mathcal{S}\right\}$ gives a conditional probability mass function, i.e., $P_{ij} \geq 0$ and $\sum_{j \in \mathcal{S}} P_{ij} = 1$.
- As illustrated in Example 5, for a finite state space Markov chain with $|\mathcal{S}| = d$, $\left(P_{ij}\right)_{i,j \in \mathcal{S}}$ gives a matrix $\mathbf{P}_{d \times d}$ which is called the **transition probability matrix** of the Markov chain.
- For time homogeneous Markov chain $\{X_n : n \in \mathbb{N}\}$, any finite dimensional joint distributions can be expressed in terms of entries of transition probability matrix and the initial distribution of X_0.

Having defined one-step transition probabilities, in order to define n-step transition probabilities $P_{ij}^n = \mathbb{P}(X_n = j \mid X_0 = i)$, Chapman-Kolmogorov equations come to the rescue. The **Chapman-Kolmogorov equations** are

$$P_{ij}^{n+m} = \sum_{k=1}^{d} P_{ik}^n P_{kj}^m \qquad \text{for all } n, m \in \{1, \ldots, d\}, \text{ all } i, j \tag{2.1}$$

Using these equations and through induction it follows that $\mathbf{P}^{(n)}$, the n-step transition probability matrix, is given by $\mathbf{P}^n$. In other words, the n-step transition probability matrix is obtained by multiplying $\mathbf{P}$ with itself n times.

2.2 Hitting times and classifications of states of a Markov Chain

Definition 2.2 (Hitting Time)**.** Given a Markov chain $\{X_n : n \in \mathbb{N}\}$, the hitting time for a set $A \subset \mathcal{S}$ is defined to be

$$T_A := \min\{n \geq 1 : X_n \in A\}.$$

If $A = \{i\}$ the singleton set, then

$$T_i := \min\{n \geq 1 : X_n = i\}.$$

As will be seen later, the hitting time random variable happens to be a useful random variable by simplifying certain computations and providing equivalent ways of defining objects when dealing with Markov chains. An important and natural question while dealing with hitting times is whether this random variable is finite or not.

Let ρ_{ij} denote

$$\rho_{ij} = \mathbb{P}_i(T_j < \infty) = \mathbb{P}(T_j < \infty \mid X_0 = i)$$

A state j is said to be **accessible** from state i if $P_{ij}^n > 0$ for some $n \geq 1$. Equivalently, j is accessible from i if $\rho_{ij} > 0$. Two states i and j are said to **communicate** if i and j are both accessible from each other.

A Markov chain is said to be **irreducible** if for all $i, j \in \mathcal{S}$, state j is accessible from state i. Extending to subsets, a subset $C \subset \mathcal{S}$ of the state space, is said to be irreducible if for any $i, j \in C$, j is accesible from i. Further, $C \in \mathcal{S}$ is **closed** if $\rho_{ij} = 0$ for all $i \in C$ and $j \notin C$.

The state i is **absorbing** if $P_{ij} = 0$ for all $j \neq i$. Extending to sets, a set C is absorbing if $P\left(X_1 \notin C \mid X_0 \in C\right) = 0$.

2.3 Stopping Times

Given a stochastic process $X = \{X_n : n \geq 0\}$, a random time τ is a discrete random variable on the same probability space as X, taking values in $\mathbb{N} = \{0, 1, 2, \ldots\}$. X_τ denotes the state at the random time τ; i.e., if $\tau(\omega) = n$, then $X_{\tau(\omega)}(\omega) = X_n(\omega)$.

The essence of the stopping time is as follows – if one were to sequentially observe the values of $X_1, X_2, \ldots$ and then 'stop' observing them after some time n, basing the decision to stop observing only on what has been observed so far, then that is a stopping time. The idea is that one does not know the future hence can't base a decision to stop now on knowing the future.

Definition 2.3. A positive integer valued random variable τ defined on the probability space $(\Omega, (\mathcal{F}_n)_{n\in\mathbb{N}}, \mathbb{P})$ is said to be a **stopping time with respect to the filtration** $\mathcal{F}_n$ if the following condition holds:

$$\{\tau \leq n\} \in \mathcal{F}_n \qquad \text{for all } n.$$

Equivalently, τ is a stopping time if and only if

$$\{\tau = n\} \in \mathcal{F}_n \; \forall \, n \in \mathbb{N}.$$

The proof of equivalence follows as if $\{\tau \leq n\} \in \mathcal{F}_n$ for all n, then

$$\{T = n\} = \{T \leq n\} \backslash \{T \leq n-1\} \in \mathcal{F}_n.$$

Conversely, as $k \leq n$, $\{T = k\} \in \mathcal{F}_k \subseteq \mathcal{F}_n$ and

$$\{T \leq n\} = \bigcup_{0\leq k\leq n} \{T = k\} \in \mathcal{F}_n.$$

For a stochastic process $\{X_n : n \geq 0\}$, a random variable τ is said to be a **stopping time** (or **optional time**) if

$$\{\tau \leq n\} \in \sigma(X_1, \ldots, X_n) \qquad \text{for all } n.$$

Every constant $\tau := t_0$ is (trivially) a stopping time; it corresponds to the stopping rule "stop at time t_0". **Hitting times** mentioned in Definition 2.2 is another natural example of a stopping time. It is important to point out that all hitting times are stopping times, but not all stopping times are hitting times. One can also construct new examples from old ones. If τ_1, τ_2 are two stopping times, then $\tau_1 + \tau_2$,$\tau_1 \wedge \tau_2$ and $\tau_1 \vee \tau_2$ are also stopping times.

For any $n \in \mathbb{N}$, and for any stopping time τ the r.v. $n \wedge \tau$ also forms a stopping time. More generally, for two stopping times τ_1 and τ_2 the r.v. $\tau_1 \wedge \tau_2$ is a stopping time as well. Denote by $\mathcal{F}_\infty := \sigma((\mathcal{F}_n : n \in \mathbb{N}))$, where $\mathcal{F}_n$ is the canonical filtration that the stochastic process $\{X_n : n \in \mathbb{N}\}$ is adapted to. It is easy to observe that $\mathcal{F}_\infty$ forms a $\sigma-$field. As $\{\tau \leq n\} \in \mathcal{F}_n \subset \mathcal{F}_\infty$ for all n, it follows that τ is measurable with respect to $\mathcal{F}_\infty$. The σ-field generated by the r.v. τ, i.e., the minimal σ-field w.r.t. which τ is measurable is denoted by $\mathcal{F}_\tau$. Below we present a more detailed representation of $\mathcal{F}_\tau$:

$$\mathcal{F}_\tau := \{A \in \mathcal{F}_\tau : \text{ such that } A \cap \{\tau \leq n\} \in \mathcal{F}_n\}.$$

Further, for every $B \in \mathcal{B}(\mathbb{R})$, $X_n^{-1}(B) \in \mathcal{F}_n \subseteq \mathcal{F}_\infty$ for all n. It then follows that $X_n^{-1} \cap \{\tau = n\} \in \mathcal{F}_n$ for all n, and so

$$X_\tau^{-1} \cap \{\tau \leq n\} = \bigcup_{i=1}^{n} (X_i^{-1} \cap \{\tau = i\}) \in \mathcal{F}_i \subseteq \mathcal{F}_n$$

which yields that X_τ is measurable with respect to $\mathcal{F}_\tau$. From this one can also easily obtain that $X_{n\wedge\tau}$ is measurable with respect to $\mathcal{F}_\tau$.

3. Martingales

3.1 Introduction

Probability theory has its roots in games of chance, and it is often profitable to interpret results in terms of a gambling situation. For example, if $X_1, X_2, \ldots$ is a sequence of random variables, one can think of X_n as the total winnings after n trials in a succession of games. Having survived the first n trials, the expected fortune after trial $n+1$ is $\mathbb{E}\left(X_{n+1} \mid X_1, \ldots, X_n\right)$. If this equals X_n, the game is "fair" since the expected gain on trial $n+1$ is $\mathbb{E}\left(X_{n+1} \ -X_n \mid X_1, \ldots, X_n\right) = X_n - X_n = 0$. If $E\left(X_{n+1} \mid X_1, \ldots, X_n\right) \geq X_n$, the game is "favorable", and if $\mathbb{E}\left(X_{n+1} \mid X_1, \ldots, X_n\right) \leq X_n$, the game is "unfavorable." The following chapter concerns itself with sequences of this type.

Definition 3.1 (Martingale). A stochastic process $\{X_n : n \in \mathbb{N}\}$ forms a **martingale** with respect to $\{\mathcal{F}_n : n \in \mathbb{N}\}$ if

(i) $\{X_n : n \in \mathbb{N}\}$ is adapted to $\{\mathcal{F}_n : n \in \mathbb{N}\}$ (this is automatically true when $\mathcal{F}_n$ is the canonical filtration).

(ii) $\mathbb{E}(|X_n|) < \infty$ for all n.

(iii) $\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) = X_n$ a.s.

If Definition 3.1 (iii) is replaced by the condition $\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) \geq X_n$ a.s, then $\{X_n : n \in \mathbb{N}\}$ forms a **submartingale**.

If Definition 3.1 (iii) is replaced by the condition $\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) \leq X_n$ a.s, then $\{X_n : n \in \mathbb{N}\}$ forms a **supermartingale**.

Clearly, any martingale is a supermartingale as well as submartingale. Following are some examples of martingales:

Example 6. The simple symmetric random walk mentioned in Example 4 is a martingale with respect to the natural filtration $\mathcal{F}_n = \sigma(S_0, Z_1, Z_2, \ldots, Z_n)$ for all n. This is true because $\mathbb{E}(S_n) = 0$ for all n, and

$$\mathbb{E}(S_{n+1} \mid \mathcal{F}_n) = \mathbb{E}(S_n \mid \mathcal{F}_n) + \mathbb{E}(Z_{n+1} \mid \mathcal{F}_n) = S_n + \mathbb{E}(Z_{n+1} \mid \mathcal{F}_n) = S_n$$

as $\mathbb{E}(Z_{n+1} \mid \mathcal{F}_n) = \mathbb{E}(Z_{n+1}) = 0$ due to the independence between Z_{n+1} and $\mathcal{F}_n$.

Example 7. Let $\{X_n : n \in \mathbb{N}\}$ be any collection of random variables such that $\mathbb{E}(X_n)$ exists and is 0 for all n. Then, $S_n := \sum_{i=1}^{n} X_n$ gives a martingale with respect to the natural filtration. The argument is similar as in Example 6.

Example 8. Let $\{X_n : n \in \mathbb{N}\}$ be independent random variables such that $\mathbb{E}(X_n) = 1$ for all n. Define $Y_n := \prod_{i=1}^{n} X_i$. Then, $\{Y_n : n \in \mathbb{N}\}$ forms a martingale as

$$\mathbb{E}(Y_{n+1} \mid \mathcal{F}_n) = \mathbb{E}(X_{n+1} \mid \mathcal{F}_n)\mathbb{E}(Y_n \mid \mathcal{F}_n) = 1 \cdot \mathbb{E}(Y_n \mid \mathcal{F}_n) = Y_n$$

due to the independence of X_{n+1} and $\mathcal{F}_n$.

Example 9 (Doob's Martingale)**.** Let Y be any random variable with $\mathbb{E}(|Y|) < \infty$. Suppose $\{\mathcal{F}_0, \mathcal{F}_1, \ldots\}$ is a filtration, i.e. $\mathcal{F}_s \subset \mathcal{F}_n$ when $s < n$. Define

$$Z_n = \mathbb{E}(Y \mid \mathcal{F}_n)$$

then $\{Z_0, Z_1, \ldots\}$ form a martingale as

- $\mathbb{E}(|Z_n|) = \mathbb{E}(|\mathbb{E}(Y \mid \mathcal{F}_n)|) \leq \mathbb{E}(\mathbb{E}(|Y| \mid \mathcal{F}_n)) = \mathbb{E}(|Y|) < \infty$;
- $\mathbb{E}(Z_n \mid \mathcal{F}_{n-1}) = \mathbb{E}(\mathbb{E}(Y \mid \mathcal{F}_n) \mid \mathcal{F}_{n-1}) = \mathbb{E}(Y \mid \mathcal{F}_{n-1}) = Z_{n-1}$ as $\mathcal{F}_{n-1} \subset \mathcal{F}_n$.

3.2 Martingale Convergence Theorems

Joseph L. Doob was a famous American mathematician who contributed significantly to the theory of martingales, and Doob's martingale convergence theorems are a collection of results on the limits of supermartingales. Informally, the martingale convergence theorem typically refers to the result that any supermartingale satisfying a certain boundedness condition must converge. One may think of supermartingales as the random variable analogues of non-increasing sequences; from this perspective, the martingale convergence theorem is a random variable analogue of the monotone convergence theorem, which states that any bounded monotone sequence converges. There are symmetric results for submartingales, which are analogous to non-decreasing sequences.

As mentioned in Section 2.3, the stopped process $X_{\tau\wedge n}$ where τ is a stopping time is $\mathcal{F}_\tau$ measureable. One can in fact prove the following theorem:

Theorem 3.1. *(a) If X is a supermartingale and τ is a stopping time (w.r.t. the same filtration), then the stopped process $X^\tau = \left(X_{\tau\wedge n} : n \in \mathbb{Z}^+\right)$is a supermartingale, so that in particular,*

$$\mathbb{E}\left(X_{\tau\wedge n}\right) \leq \mathbb{E}\left(X_0\right), \qquad \textit{for all } n.$$

(b) If X is a martingale and τ is a stopping time (w.r.t. the same filtration), then X^τ is a martingale, so that in particular,

$$\mathbb{E}\left(X_{\tau\wedge n}\right) = \mathbb{E}\left(X_0\right), \qquad \textit{for all } n.$$

It is important to notice that this theorem imposes no extra integrability conditions whatsoever (except of course for those implicit in the definition of supermartingale and martingale).

Theorem 3.1 clearly implies that

$$\mathbb{E}\left(X_{\tau\wedge n}\right) = \mathbb{E}\left(X_0\right) \text{ for every } n.$$

Now, it is natural to ask that under what conditions we can have $\mathbb{E}\left(X_\tau\right) = \mathbb{E}\left(X_0\right)$ too. We need to assume that τ is a.s. finite to make sure that the r.v. X_τ is defined throughout. The following example illustrates that this is not automatic.

Consider the the random walk X defined in Example 4. X is a martingale as proved in Example 6. Define $\tau := \min\{n \geq 0 : X_n = 1\}$. It is well known that $P(\tau < \infty) = 1$. However, even though Theorem 3.1 implies that

$$\mathbb{E}\left(X_{\tau\wedge n}\right) = \mathbb{E}\left(X_0\right) \text{ for every } n,$$

we have

$$1 = \mathbb{E}\left(X_\tau\right) \neq \mathbb{E}\left(X_0\right) = 0.$$

Thus, one would like to identify conditions when one can claim that

$$\mathbb{E}\left(X_\tau\right) = \mathbb{E}\left(X_0\right)$$

for a martingale $X_n : n \in \mathbb{N}$. The following theorem gives some sufficient conditions.

Theorem 3.2 (Doob's Optional-Stopping Theorem). *(a) Let τ be a stopping time which is a.s. finite. Let X be a supermartingale. Then X_τ is integrable and*

$$\mathbb{E}\left(X_\tau\right) \leq \mathbb{E}\left(X_0\right)$$

in each of the following situations:

(i) τ is bounded (for some N in $\mathbb{N}, \tau(\omega) \leq N, \forall \omega$);

(ii) X is bounded (for some K in $\mathbb{R}^+, |X_n(\omega)| \leq K$ for every n and every ω);

(iii) $\mathbb{E}(\tau) < \infty$, and, for some K in $\mathbb{R}^+$,

$$|X_n(\omega) - X_{n-1}(\omega)| \leq K \qquad \text{for all } (n, \omega).$$

(b) If any of the conditions (i)-(iii) holds and X is a martingale, then

$$\mathbb{E}\left(X_\tau\right) = \mathbb{E}\left(X_0\right).$$

A common idea used to prove the martingale convergence theorems is that of upcrossings. Let N be a natural number. Let $(X_n)_{n \in \mathbf{N}}$ be a supermartingale with respect to a filtration $(\mathcal{F}_n)_{n \in \mathbf{N}}$. Let a, b be two real numbers with $a < b$. Define the random variables $(U_n)_{n \in \mathbf{N}}$ so that U_n is the maximum number of disjoint intervals $\left[n_{i_1}, n_{i_2}\right]$ with $n_{i_2} \leq n$, such that $X_{n_{i_1}} < a < b < X_{n_{i_2}}$. These are called **upcrossings** with respect to interval $[a, b]$. A diagram is provided below to further illustrate the definition.

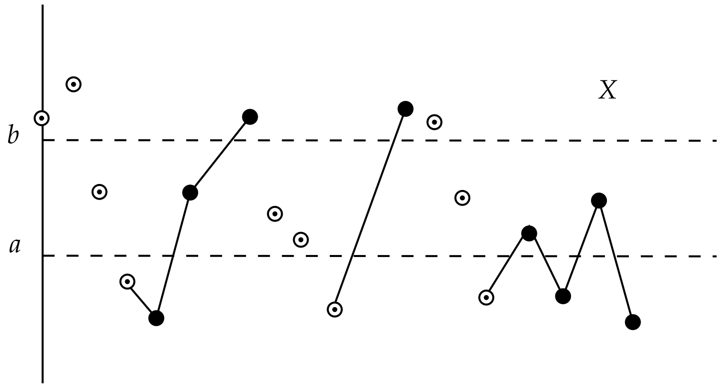


Figure 3.1: Upcrossings with respect to an interval $[a, b]$ are illustrated.

Theorem 3.3 (Doob's Upcrossing Lemma). *Let X be a supermartingale and let $U_N[a, b]$ denote the number of upcrossings of the interval $[a, b]$ by the process X in time N. Then*

$$(b - a)\mathbb{E}\left(U_n\right) \leq \mathbb{E}\left((X_n - a)^-\right)$$

where $X^-$is the negative part of X, defined by $X^- = -\min(X, 0)$.

Martingale convergence is a consequence of the upcrossing lemma. The following lemma states that and $\mathcal{L}^1$ bounded martingale X_n in discrete time converges almost surely.

Theorem 3.4 (Doob's Forward Convergence Theorem). *Let X be a $\mathcal{L}^1$ bounded supermartingale i.e., $\sup \mathbb{E}\left(|X_n|\right) < \infty$. Then, $X_\infty := \lim X_n$ exists a.s. and is finite, i.e.,*

$$\mathbb{P}(X_\infty \in \mathbb{R}) = 1 \qquad a.s.$$

Corollary 1. *If X is a non-negative super martingale, then $X_\infty := \lim X_n$ exists a.s.*

Proof. X is bounded as $\mathbb{E}(|X_n|) = \mathbb{E}(X_n) = E(X_0)$. □

Corollary 2. *Let X_n be a non-negative Martingale and a Markov chain taking values in the set of non-negative integers with an absorbing state 0. Further $\mathbb{P}(X_{n+1} = 0 \mid X_n = k) = p_k > 0$ for all $k \in \mathbb{N}$. Then $\lim_{n\to\infty} X_n = X_\infty$ exists and is 0 with probability 1.*

Proof. Firstly notice that from Doob's forward convergence theorem and Corollary 1, it follows that X_∞ exists and is finite. In order to show that it is 0, proceed by contradiction. Assume that the event $B := \{X_\infty = k\}$ is of positive probability for some $k \in \mathbb{N}$. By the property of almost sure convergence on the discrete space of $\mathbb{N} \cup \{0\}$ it follows that for $\omega \in B$, there exists $n_0 = n_0(\omega) \in \mathbb{N}$ such that for all $m \geq n_0, X_m(\omega) = k$. This contradicts the fact that $\mathbb{P}(X_{n+1} = 0 \mid X_n = k) = p_k > 0$. □

4. Poisson Point Processes

4.1 Counting Processes

The **counting process** $\{N(t); t > 0\}$, is family of random variables $\{N(t); t > 0\}$ where $N(t)$, for each $t > 0$, is the number of arrivals in the interval $(0, t]$. More formally, a stochastic process $\{N(t); t \geq 0\}$ is said to be a counting process if the following hold:

(i) $N(t) \geq 0$

(ii) $N(t) \in \mathbb{Z}^+$

(iii) If $s < t$, then $N(s) < N(t)$

(iv) For $s < t$, $N(t) - N(s)$ denotes the number of events in the interval $(s, t]$.

Many everyday examples of counting processes include keeping track of the number of people arriving at a bus stop at or prior to time t or the number of people born in a hospital at or prior to time t.

If the number of events that occur in disjoint intervals are independent, then the counting process is said to possess *independent increments*. This could be a fair assumption to make in certain models, for example arrivals at a bus-stop. For situations like the number of births at or prior to time t, this could be unreasonable, for if $N(s)$ is small at some s, then for time $t + s$, it would be only fair to estimate that $N(t + s)$ would be small too; in other words, it may not be appropriate to suppose that $N(t)$ is independent of $N(t + s) - N(s)$.

A counting process for which the distribution of the number of events in an interval depends only on the length of the interval and not its location is said to have *stationary increments*. A Poisson Point Process (elaborated on in the next section) is one such example.

The n^{th} arrival epoch $\{S_n \leq t\}$ denotes the event that the n^{th} arrival occurs by time t. This event implies that $N(t)$, the number of arrivals by time t, must be at least n, i.e., it implies the event $\{N(t) \geq n\}$. Similarly, $\{N(t) \geq n\}$ implies $\{S_n \leq t\}$, yielding

$$\{S_n \leq t\} = \{N(t) \geq n\}.$$

By taking complements, one can observe that

$$\{S_n > t\} = \{N(t) < n\}.$$

4.2 Poisson Point Processes

One such model of a counting process with independent increments can be given by the Poisson Point process, defined below:

Definition 4.1 (Poisson Point Process)**.** A counting process $\{N(t); t > 0\}$ is said to be a Poisson process on $\mathbb{R}$ if the following axioms hold:

(i) $N(0) = 0$

(ii) $\{N(t), t \geq 0\}$ has independent increments

(iii) $\mathbb{P}(N(t+h)) - N(t) = 1) = \lambda h + o(h)$

(iv) $\mathbb{P}(N(t+h) - N(t) \geq 2) = o(h)$

Let T_1 denote the first arrival of a Poisson process $\{N(t); t > 0\}$. That is,

$$T_1 := \min\{t \geq 0 : N(t) = 1\}$$

Lemma 4.1. *If T_1 denotes the first time of arrival of a Poisson Point process, then*

$$\mathbb{P}(T_1 > t) = P(N(t) = 0) = e^{-\lambda t}$$

Proof. Let $\mathbb{P}_0(t) = P(N(t) = 0)$. Then

$$\begin{aligned} \mathbb{P}_0(t+h) &= P(N(t+h) = 0) \\ &= \mathbb{P}(N(t) = 0, N(t+h) - N(t) = 0) \\ &= \mathbb{P}(N(t) = 0) P(N(t+h) - N(t) = 0) && \text{by Axiom (ii)} \\ &= \mathbb{P}_0(t)(1 - \lambda h + o(h)) && \text{by Axioms (iii) and (iv)} \end{aligned}$$

Hence,

$$\mathbb{P}_0(t+h) - \mathbb{P}_0(t) = -\lambda h \mathbb{P}_0(t) + o(h)$$

Dividing by h and then letting $h \to 0$, gives that

$$\mathbb{P}_0'(t) = -\lambda \mathbb{P}_0(t)$$

or, equivalently

$$\frac{\mathbb{P}_0'(t)}{\mathbb{P}_0(t)} = -\lambda$$

Integration yields

$$\log(\mathbb{P}_0(t)) = -\lambda t + C$$

or

$$\mathbb{P}_0(t) = K e^{-\lambda t}$$

Using that $1 = \mathbb{P}_0(0)$ gives that $K = 1$. Because the time of the first event exceeds t if and only if $N(t) = 0$, we see that $\mathbb{P}(T_1 > t) = \mathbb{P}(N(t) = 0) = e^{-\lambda t}$. □

The sequence $\{T_n : n \in \mathbb{N}\}$ is called the *sequence of interarrival times*. Using the property of independent increments, it can further be shown that $T_1, T_2, \ldots$ are all i.i.d exponential random variables. Thus, one can formulate an equivalent definition of a Poisson Process as below:

Definition 4.2. A Poisson Point Process is a sequence of increasing random variables in which the interarrival times are a sequence of i.i.d random variables that have exponential distribution.

In order to find out the distribution of $N(t)$, the following fact will be used to derive the same. Firstly, notice that from Lemma 4.1, it follows that $f_{T_1}(t) = \lambda e^{-\lambda t}$. If $S_n = \sum_{i=1}^n T_i$, then S_n is the sum of n many i.i.d random variables with exponential distribution, from which it can be obtained that S_n follows the Erlang distribution with parameters (n, λ), i.e.,

$$f_{S_n}(s) = \frac{\lambda(\lambda s)^{n-1} e^{-\lambda s}}{(n-1)!}$$

.

Lemma 4.2. *If $\{N(t); t \geq 0\}$ is a Poisson Process with rate λ, then $N(t)$ is a Poisson random variable with parameter λt i.e.,*

$$\mathbb{P}(N(t) = n) = e^{-\lambda t}(\lambda t)^n / n!. \tag{4.1}$$

Proof. In Lemma 4.1 it was shown that $\mathbb{P}(N(t)=0)=e^{-\lambda t}$. For $n>0$, conditioning on S_n, one can compute $\mathbb{P}(N(t)=n)$ by

$$\begin{aligned}\mathbb{P}(N(t)=n) &= \int_0^t \mathbb{P}(N(t)=n \mid S_n=s) f_{S_n}(s)ds \\ &= \int_0^t \mathbb{P}\left(N(t)=n \mid S_n=s\right)\lambda e^{-\lambda s}\frac{(\lambda s)^{n-1}}{(n-1)!}ds \end{aligned} \tag{4.2}$$

where the preceding used that $\mathbb{P}\left(N(t)=n \mid S_n=s\right)=0$ when $s>t$. Now, for $0<s<t$, given that the nth event occurs at time s, there will be a total of n events by time t if the next interarrival time exceeds $t-s$. Hence,

$$\begin{aligned}\mathbb{P}\left(N(t)=n \mid S_n=s\right) &= \mathbb{P}\left(T_{n+1}>t-s \mid T_1+\ldots T_n=s\right) \\ &= \mathbb{P}\left(T_{n+1}>t-s\right) \\ &= e^{-\lambda(t-s)}\end{aligned}$$

where the last two equalities both used the fact that $T_1, T_2, \ldots$ are i.i.d exponentially distributed. Substituting this back into Eq (4.2) yields that

$$\begin{aligned}\mathbb{P}(N(t)=n) &= \int_0^t e^{-\lambda(t-s)}\lambda e^{-\lambda s}\frac{(\lambda s)^{n-1}}{(n-1)!}ds \\ &= e^{-\lambda t}\lambda^n \int_0^t \frac{s^{n-1}}{(n-1)!}ds \\ &= e^{-\lambda t}(\lambda t)^n/n!\end{aligned}$$

□

A random variable X is said to be *without memory*, or *memoryless*, if

$$\mathbb{P}\{X>s+t \mid X>t\}=\mathbb{P}\{X>s\} \quad \text{for all } s,t\geq 0 \tag{4.3}$$

The condition in Eq (4.3) is equivalent to

$$\begin{aligned} &\frac{\mathbb{P}(X>s+t, X>t)}{\mathbb{P}(X>t)} = \mathbb{P}(X>s) \\ \Rightarrow \quad & \mathbb{P}(X>s+t)=\mathbb{P}(X>s)\mathbb{P}(X>t)\end{aligned} \tag{4.4}$$

Since Eq (4.4) is satisfied when X is exponentially distributed (for $e^{-\lambda(s+t)}=e^{-\lambda s}e^{-\lambda t}$), it follows that exponentially distributed random variables are memoryless. This implies that the interarrival times for a Poisson process are memoryless.

Suppose exactly one arrival of a Poisson process has taken place by time t, then how does one determine the distribution of the time at which the event occurred? Since a Poisson process possesses stationary and independent increments it seems reasonable that each interval in $[0,t]$ of equal length should have the same probability of containing the event. In other words, the time of the event should be uniformly distributed over $[0,t]$. To confirm, for $s\leq t$,

$$\begin{aligned}\mathbb{P}\left(T_1<s \mid N(t)=1\right) &= \frac{\mathbb{P}\left(T_1<s, N(t)=1\right)}{\mathbb{P}\left(N(t)=1\right)} \\ &= \frac{\mathbb{P}\{1 \text{ event in } [0,s), 0 \text{ events in } [s,t]\}}{\mathbb{P}\left(N(t)=1\right)} \\ &= \frac{\mathbb{P}\left(1 \text{ event in } [0,s)\right)\mathbb{P}\left(0 \text{ events in } [s,t]\right)}{\mathbb{P}\left(N(t)=1\right)} \\ &= \frac{\lambda s e^{-\lambda s}e^{-\lambda(t-s)}}{\lambda t e^{-\lambda t}} \\ &= \frac{s}{t}.\end{aligned}$$

4.2.1 Poisson Point Process in $\mathbb{R}^2$

Keeping this motivation in mind, one can extend Poisson Processes from $\mathbb{R}$ to higher dimensions. One first considers a Borel measureable region A of the plane. The number of points of a point process N existing in this region $A \subset \mathbb{R}^2$ is a random variable, denoted by $N(A)$. If the points belong to a homogeneous Poisson process with parameter $\lambda > 0$, then the probability of n points existing in A is given by the Poisson distribution. Formally, we define a Poisson Point Process on $\mathbb{R}^2$ as the following:

Definition 4.3. A counting process $\mathcal{N} := \{N(A) : A \text{ an open bounded subset of } \mathbb{R}^2\}$ is said to be a Poisson Point Process on $\mathbb{R}^2$ with intensity $\lambda > 0$ if

(i) For a Borel measureable region $A \subset \mathbb{R}^2$, $N(A) \sim \text{Poi}(\lambda l(A))$, where $l(A)$ is the Lebesgue measure on $\mathbb{R}^2$.

(ii) For disjoint open subsets $A_1, \ldots, A_k \subset \mathbb{R}^2$, $N(A_1), \ldots, N(A_k)$ are independent random variables

Properties about memorylessness and uniform distributions can analogously be extended to $\mathbb{R}^2$ as well.