Probability Primer Handout

Algorithmical and Statistical Modelling, Fall 2012.

Contents

Chapter 1

Quarks of Measure Theory

These notes are just the outline of the topics covered in the lectures (without details). Several other topics in measure theory not listed here should not be deemed to be not quintessential, but nevertheless the topics here gives a basis for a rigorous foundation for understanding probabilistic modeling and a profound theory that follows it.

The notion of a probability space is a fundamental abstract object to start with. The probability spaces comprises three entities viz., a sample space (usually denoted by Ω), all interesting subsets of the sample space called a σ-algebra ("sigma-algebra" in words), and a real valued function which defines values between 0 and 1 (inclusive of 0 and 1) for each element in the σ-algebra (note that an element in the σ-algebra is actually a subset of the sample space).

1.1 Sigma-algebra : A special subset of the powerset of a set

Definition 1.1. A collection $\mathcal{F}$ of subsets of a set Ω is said to be a σ–algebra of Ω if $\mathcal{F}$ has the following properties:

(i). $\Omega \in \mathcal{F}$.

(ii). If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, where A^c is the complement of A relative to Ω.

(iii). If $\{A_1, A_2 \ldots\}$ are any countable collection of subsets such that $A_n \in \mathcal{F}$, then their union $\bigcup_n A_n \in \mathcal{F}$.

It may be noted that as a consequence of (i) and (ii) above, $\{\emptyset, \Omega\}$ belong to any $\mathcal{F}$. In fact

$\{\emptyset, \Omega\}$ is the smallest sigma-algebra of Ω. Property (iii) is a requirement for a σ-algebra to have the closure under countable union. A consequence of (ii) and (iii) is that a σ-algebra is also closed under countable intersections (apply De Morgan's laws to achieve it).

Proposition 1.1. *Let Ω be a non-empty set, C be a nonempty collection of subsets of Ω. Let S be the set of all σ-algebras containing C and $\sigma(C) = \bigcap_{\mathcal{F} \in S} \mathcal{F}$ be its intersection. Then $\sigma(C)$ is a sigma-algebra.*

Proof. $\sigma(C)$ is a sigma-algebra since :

1. $\forall \mathcal{F} \in S, \Omega \in \mathcal{F}$, hence $\Omega \in \bigcap_{\mathcal{F} \in S} \mathcal{F} = \sigma(C)$.
2. If $A \in \sigma(C)$, then for every $\mathcal{F} \in S$, $A \in \mathcal{F}$, hence $A^C \in \mathcal{F}$, so $A^C \in \bigcap_{\mathcal{F} \in S} \mathcal{F} = \sigma(C)$.
3. Let $A_1, A_2, \cdots \in \sigma(C)$. Then, for every $\mathcal{F} \in \mathcal{S}$, $A_1, A_2, \cdots \in \mathcal{F}$. Hence $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$, for every $\mathcal{F}$. Thus $\bigcup_{i=1}^{\infty} A_i \in \sigma(C)$. ■

Remark 1.1. Given a collection of nonempty subsets C of Ω, the intersection of all sigma-algebras is called the "sigma-algebra generated by C" and it is the smallest sigma-algebra containing C. It is standard to denote such a sigma-algebra by $\sigma(C)$.

Among the many possible sigma-algebras on a set Ω, of interest is to obtain sigma-algebras generated by a topology on Ω (see Appendix for the definition of a topology). A sigma-algebra thus generated by a topology is called a Borel sigma-algebra on Ω.

When $\Omega = \mathbb{R}$ it is standard to use the sigma-algebra generated by topology obtained by the Euclidean metric (see Appendix for details). Such a Borel sigma-field is denoted by $\mathcal{B}(\mathbb{R})$ or just by $\mathcal{B}$.

Definition 1.2. A tuple $(\Omega, \mathcal{F})$, where $\mathcal{F}$ is a sigma-algebra on Ω is called a *measurable space* and the elements in $\mathcal{F}$ are called *measurable sets*.

1.2 Measure : A function on sigma-algebra

Definition 1.3. Given a measurable space $(\Omega, \mathcal{F})$, then $\mu : \mathcal{F} \to [0, \infty]$ is a "measure" if

(i). $\mu(\{\emptyset\}) = 0$.

(ii). if $A_i \in \mathcal{F}$ are any countable collection of mutually disjoint sets then $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.

Above, (ii) is know as the σ-additivity property of the measure μ.

Definition 1.4. 1. A triplet $(\Omega, \mathcal{F}, \mu)$ is called a "measure space".

2. For a measure space $(\Omega, \mathcal{F}, \mu)$ if $\mu(\Omega) < \infty$, then the triplet is called a "finite measure space".

3. A measure space $(\Omega, \mathcal{F}, \mu)$ is called a "probability measure space" if $\mu(\Omega) = 1$.

If $(\Omega, \mathcal{F}, \mu)$ is a finite measure space such that $0 < \mu(\Omega) < \infty$, then $(\Omega, \mathcal{F}, \lambda)$ is a probability measure space, where $\lambda(\cdot) := \frac{\lambda(\cdot)}{\mu(\cdot)}$. For a probability space $(\Omega, \mathcal{F}, \mu)$, the elements in $\mathcal{F}$ are called events.

A measure has few more properties one should always know:

Proposition 1.2. *Let $(\Omega, \mathcal{F}, \mu)$ be any measure space.*

1. *Let A and B be two elements in $\mathcal{F}$ such that $A \subset B$. Then $\mu(A) \leq \mu(B)$.*

2. *Let $\{E_i\}$ be a sequence in $\mathcal{F}$. Then $\mu(\bigcup_{i=1}^{\infty} E_i) \leq \sum_{i=1}^{\infty} \mu(E_i)$.*

One needs the notion of sigma-algebras not only for both mathematical technicalities but to define measures (or probabilities) of many sets which one can envisage to be useful. The power set of Ω is the largest sigma-algebra, but one lands up in technical difficulties defining a measure on it.

The following result confirms our intuition behind the notion of a measure.

Proposition 1.3. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space.*

1. *Let $A_1 \subset A_2 \subset A_3 \subset \cdots$ be a sequence of measurable sets. If $A = \bigcup_{i=1}^{\infty} A_i$ then $\lim_{n\to\infty} \mu(A_n) = \mu(A)$.*

2. *Let $A_1 \supset A_2 \supset A_3 \supset \cdots$ be a sequence of measurable sets. If $A = \bigcap_{i=1}^{\infty} A_i$ and $\mu(A_1) < \infty$, then $\lim_{n\to\infty} \mu(A_n) = \mu(A)$.*

Proof. **Proof of 1.** Let $B_1 = A_1$, and define $B_n = A_n - A_{n-1}$ for $n = 2, 3, 4, \ldots$. Then $B_n \in \mathcal{F}$, and $B_i \cap B_j = \emptyset$ if $i \neq j$, $A_n = B_1 \cup \cdots \cup B_n$, and $A = \cup_{i=1}^{\infty} B_i$. Hence

$$\mu(A_n) = \sum_{i=1}^{n} \mu(B_i) \quad \text{and} \quad \mu(A) = \sum_{i=1}^{\infty} B_i.$$

$\lim_{n\to\infty} \mu(A_n) = \mu(A)$ follows from the definition of a series.

Proof of 2. Define $C_n = A_1 - A_n$. Then $C_1 \subset C_2 \subset \cdots$. Hence $C = A_1 - A$ where, $C = \bigcup_n C_n$. Since $\mu(A_1) < \infty$, we can write

$$\mu(C_n) = \mu(A_1) - \mu(A_n).$$

From Proof of 1,

$$\mu(C) = \mu(A_1) - \mu(A).$$

Hence

$$\mu(C) = \lim_{n\to\infty} \mu(C_n) = \mu(A_1) - \lim_{n\to\infty} \mu(A_n).$$

This implies $\lim_{n\to\infty} \mu(A_n) = \mu(A)$ which proves 2. ■

1.3 Measurability: concept of nice functions between measurable spaces

Definition 1.5. Let $(\Omega, \mathcal{F})$ and $(S, \mathcal{B})$ be two measurable spaces. Let $f : \Omega \to S$ be any function. Then f is said to be $(\mathcal{F}, \mathcal{B})$ measurable if for every $E \in \mathcal{B}$, $f^{-1}(E) \in \mathcal{F}$.

There are examples of both measurable and non-measurable functions.

Facts about measurable functions (without worrying about their proofs):

- All continuous real valued functions on $(\mathbb{R}, \mathcal{B})$ are measurable.

  If f and g are real valued measurable functions on $(\mathbb{R}, \mathcal{B})$, then $f + g$, $f - g$, $f \cdot g$, and $\frac{f}{g}$ (when $g(\omega) \neq 0$ for any ω) are all measurable functions.

- If $\{f_n\}$ is a sequence real valued measurable functions on $(\mathbb{R}, \mathcal{B})$, then $\liminf_{n\to\infty} f_n$ and $\limsup_{n\to\infty} f_n$ are measurable functions.

Note: Two measurable functions f and g are said to be equal almost everywhere if the set

$$\{\omega : f(\omega) \neq g(\omega)\}$$

has measure zero. This is often written as $f = g$ a.e. (a.e. denotes almost everywhere).

Chapter 2

Lebesgue Integration

Definition 2.1. Consider a measurable space $(\Omega, \mathcal{F})$. If E is a measurable set in $\mathcal{F}$, then denote

$$\chi_E(\omega) := \begin{cases} 1 & : \text{ if } \omega \in E \\ 0 & : \text{ otherwise} \end{cases}$$

The function $\chi_E(\omega)$ is called the **characteristic function** of the set E. The letter χ will be reserved for characteristic functions in this notes.

In the lecture, after a review of the Riemann integral of continuous functions on the real line, the following were noted:

- Insufficiency of Riemann integral: for example, one cannot integrate the function $\chi_{\mathbb{Q}^c}(\omega)$, where $\mathbb{Q}$ is the set of all rational numbers.
- In probability theory, random variables are much complicated than continuous functions and sometimes their domain is an abstract set. We need to integrate them to find their expectation, moments etc. Hence an alternative to Riemann integral is to be found.
- French mathematician, Henri Lebesgue roughly hundred years ago provided a general theory of integration of a function with respect to a general measure.
- Lebesgue integration plays an important role in real analysis, probability theory, and many other fields in the mathematics.

To define Lebesgue integrals, special provisions are made for "measure theory". These provisions are exclusively made for "measure theory" and no further comment is attached:

- $a + \infty = \infty + a = \infty$ if $0 \leq a \leq \infty$.

- $a \cdot \infty = \infty \cdot a = 0$ if $a = 0$.
- $a \cdot \infty = \infty \cdot a = \infty$ if $a = \infty$.

Definition 2.2. A real valued (also can be a complex-valued) function s on a measurable space Ω whose range consists of only finitely many points will be called a *simple function.*

If $\alpha_1, \ldots, \alpha_n$ are the distinct values of a simple function s, and if we set $A_i = \{x : s(x) = \alpha_i\}$, then clearly

$$s = \sum_{i=1}^{n} \alpha_i \chi_{A_i},$$

where χ_{A_i} is the characteristic function of A_i. Clearly, s is measurable if and only if each of the A_is are measurable.

Definition 2.3. Consider a measurable space $(\Omega, \mathcal{F})$. If $s : \Omega \to [0, \infty)$ is a non-negative simple function given by

$$s = \sum_{i=1}^{n} \alpha_i \chi_{A_i}, \tag{2.1}$$

where $\alpha_1, \ldots, \alpha_n$ are the distinct values of s and if $E \in \mathcal{F}$, then the Lebesgue integral of s with respect to the measure μ over a set $E \in \mathcal{F}$ (denoted by $\int_E s d\mu$) is defined by [1]

$$\int_E s d\mu = \sum_{i=1}^{n} \alpha_i \mu(A_i \cap E). \tag{2.2}$$

Proposition 2.1. *Let s and h be two simple non-negative functions on $(\Omega, \mathcal{F}, \mu)$. Then*

1. **Linearity.** *For $\alpha \geq 0$ and $\beta \geq 0$, $\int(\alpha s + \beta h)d\mu = \alpha \int s d\mu + \beta \int h d\mu$.*
2. **Monotonicity.** *If $s \geq g$ a.e., $\int s d\mu \geq \int h d\mu$.*

We will see that simple functions can approximate any measurable function to an arbitrary precision in the limiting sense. First, we approximate nonnegative measurable functions by simple functions:

Proposition 2.2. *Let $f : \Omega \to [0, \infty]$ be a measurable function. Then there exist simple measurable functions s_n such that*

1. $0 \leq s_1 \leq s_2 \leq \cdots \leq f$
2. $s_n(\omega) \to f(\omega)$ *as* $n \to \infty$ *for all* $\omega \in \Omega$.

[1] Good time to notice the special provisions made for measure theory. Here, for instance, $0 \cdot \infty = 0$ is used in Measure theory. It may happen that $\alpha_i = 0$ for some i and that $\mu(A_i \cap E) = \infty$.

Definition 2.4. For a non-negative measurable function $f : X \to [0, \infty]$ and $E \in \mathcal{B}$, define

$$\int_E f d\mu = \sup \int_E s d\mu, \tag{2.3}$$

the supremum being taken over all simple measurable functions s such that $0 \leq s \leq f$. Then $\int_E f d\mu$ is called the **Lebesgue integral of f over E, with respect to the measure μ.** It is a number in $[0, \infty]$.

Observe that we have apparently two definitions of $\int_E f d\mu$ if f is simple, namely (2.2) and (2.3). However, these assign the same value to the integral as made clear in the following proposition:

Proposition 2.3. *Let $\{s_n\}$ and $\{h_n\}$ be two sequences of measurable simple nonnegative functions on a measure space $(\Omega, \mathcal{F})$ which map Ω to $[0, \infty)$ such that $s_n(\omega) \uparrow f(\omega)$ and $h_n(\omega) \uparrow f(\omega)$ for all $\omega \in \Omega$, where f is some measurable function. Then*

$$\lim_{n\to\infty} \int s_n d\mu = \lim_{n\to\infty} \int h_n d\mu.$$

The following facts are immediate consequences of Definition 2.4 (the functions and sets occuring in these are assumed to be measurable):

1. If $0 \leq f \leq g$ then $\int_E f d\mu \leq \int_E g d\mu$
2. If $A \subset B$ and $f \geq 0$, then $\int_A f d\mu \leq \int_B f d\mu$.
3. If $f \geq 0$ and c is a constant, $0 \leq c \leq \infty$ then $\int_E cf d\mu = c \int_E f d\mu$.
4. If $f(x) = 0$ for all $x \in E$ then $\int_E f d\mu = 0$ even if $\mu(E) = \infty$.
5. If $\mu(E) = 0$ then $\int_E f d\mu = 0$ even if $f(x) = \infty$ for every $x \in E$.
6. If $f \geq 0$ then $\int_E f d\mu = \int_\Omega \chi_E f d\mu$.

Lebesgue integral of general measurable functions:

Definition 2.5. Given any real valued function f, define the two functions f_+ and f_- as follows : $f_+(x) = \max(f(x), 0)$ and $f_-(x) = \max(-f(x), 0)$, i.e., f_+ and f_- are the positive and negative parts of f. Then the **Lebesgue integral of f over E, with respect to the measure** μ exists and is equal to $\int_E f d\mu = \int_E f_+ d\mu - \int_E f_- d\mu$ provided the difference $\int_E f_+ d\mu - \int_E f_- d\mu$ is well defined. In a similar vein, the integral is defined for a complex function taking the real and complex parts separately, but we do not consider complex functions.

Remark 2.1. $\int_E f d\mu$ can be equal to ∞. However, if both $\int_E f_+ d\mu = \infty$ and $\int_E f_- d\mu = \infty$, then $\int_E f d\mu$ is not defined since $\infty - \infty$ is not defined.

Definition 2.6. The set $L^1(\mu)$ is the collection of all complex measurable functions f on Ω for which

$$\int_\Omega |f| d\mu < \infty.$$

• The members of $L^1(\mu)$ are called the **Lebesgue integrable functions with respect to μ.** (**Warning.** Note that if $f \notin L^1(\mu)$ it does not necessarily mean that its Lebesgue integral does not exist – compare this with the definition of $\int_E f d\mu$. It is a simple exercise to check that f belongs to $L^1(\mu)$ if and only if $|\int_\Omega f d\mu| < \infty$.

Chapter 3

The Lebesgue Measure

One of the most important measures considered in measure theory is the 'Lebesgue measure'.

A k-dimensional Euclidean space $\mathbb{R}^k$ is the set of all points $x = (\xi_1, \xi_2, \ldots, \xi_k)$ whose coordinates ξ_i are real numbers with the following structure:

If $x = (\xi_1, \xi_2, \ldots, \xi_k)$, $y = (\eta_1, \eta_2, \ldots, \eta_k)$ and α is a real number, $x + y$ and αx are defined by $x + y := (\xi_1 + \eta_1, \xi_2 + \eta_2, \ldots, \xi_k + \eta_k)$, $\alpha x = (\alpha\xi_1, \alpha\xi_2, \ldots, \alpha\xi_k)$.

This makes $\mathbb{R}^k$ into a vector space. If $E \subset \mathbb{R}^k$ and $x \in \mathbb{R}^k$ the translate of E by x is the set $E + x := \{y + x : y \in E\}$.

A set of the form

$$W = \{x : \alpha_i < x < \beta_i, 1 \leq i \leq k\} \tag{3.1}$$

or any set obtained by replacing any or all of the $<$ signs in (3.1) by $\leq$, is called a k-cell; its volume is defined to be

$$\mathrm{Vol}(W) = \prod_{i=1}^{k} (\beta_i - \alpha_i).$$

Note that all k-cells belong to the standard Borel sigma-algebra on $\mathbb{R}^k$.

Definition 3.1. Consider $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$. Let $A \in \mathcal{B}(\mathbb{R}^k)$. Then

$$m(A) = \inf_{A \subset \cup W_n} \Big(\sum_n \mathrm{Vol}(W_n) \Big). \tag{3.2}$$

is called the k-dimensional Lebesgue measure of the set A, where W_n are k-cells (here the infimum is taken over all possible covers $\cup W_n$ of A).

Note : The Lebesgue measure can be defined on any measure space $(\Omega, \mathcal{B}(\Omega))$, where Ω is a subset of $\mathbb{R}^k$ and $\mathcal{B}(\Omega) = \sigma(\mathcal{B}(\mathbb{R}^k) \cap \Omega)$.

Theorem 3.1. *Consider the measure space* $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k), m)$*, where* m *is defined as in (3.2). Then*

1. $m(W) = Vol(W)$ *for every* k*-cell* W*.*
2. m *is translation invariant, i.e.,* $m(A + x) = m(A)$ *for every* $A \in \mathcal{B}$ *and every* $x \in \mathbb{R}^k$*.*
3. *If* μ *is any translation invariant measure on* $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ *then for every closed and bounded set* K*, then there exists a constant* $c \geq 0$ *such that* $\mu(K) = c\, m(K)$*.*

Note : Where sets of different dimensions are in picture, m_k is used to denote the k^{th} dimensional Lebesgue measure. Also, note that $m_k(A) = 0$ whenever A is a Borel measurable subset of $\mathbb{R}^{k-1}$.

• The Lebesgue measure can be defined on a strictly bigger σ–algebra than the standard Borel σ–algebra. See [4, pp. 50-51] for details. Throughout this notes, when we denote a measure by m, it stands to denote the Lebesgue measure.

• Since the Lebesgue measure (in any dimension) of a singleton set is zero, the Lebesgue measure of a countable set is always zero. For instance, the Lebesgue measure (one-dimensional) of the set of all rationals is zero.

• There exist uncountable sets which also have Lebesgue measure zero. An example of it is the Cantor's middle third set (see Appendix for details).

Chapter 4

Distributions and densities of Random Variables

In the remainder of this notes we consider only random variables taking values in $\mathbb{R}^n$, and more often taking values in $\mathbb{R}$.

Suppose X is a random variable defined on a probability space $(\Omega, \mathcal{F}, P)$. The measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ can be endowed with a measure induced by X as follows: since X is $(\mathcal{F}, \mathcal{B}(\mathbb{R}))-$measurable, $X^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{B}$ and the function

$$\mu_X(A) = P(X^{-1}(A))$$

is a measure on $\mathcal{B}$ (verify it). Is this a probability measure on $\mathcal{B}$? The following proposition answers the question more generally:

Proposition 4.1. *Let $(\Omega_i, \mathcal{F}_i)$, $i = 1, 2$ be measurable spaces and let $X : \Omega_1 \to \Omega_2$ be a $(\mathcal{F}_1, \mathcal{F}_2)$- measurable mapping from $\Omega_1 \to \Omega_2$. Then, for any probability measure P on $(\Omega, \mathcal{F}_1)$, the function μ_X defined by*

$$\mu_X(A) = P(X^{-1}(A)), \quad A \in \mathcal{F}_2 \tag{4.1}$$

is a probability measure on $\mathcal{F}_2$.

The proof follows by an elementary verification of the properties of a measure.

Definition 4.1. The measure μ_X is called the measure induced by X (or the induced measure of X) on $\mathcal{F}_2$.

Definition 4.2. For a random variable X defined on a probability space $(\Omega, \mathcal{F}, P)$, the *probability distribution* of X is μ_X, the induced measure of X as defined in (4.1).

In introductory courses on probability and statistics, one defines probabilities of events like '$X \in [a, b]$' by using the probability mass function for random variables taking values belonging to a finite set or a countable set and distribution functions for so called random variables taking values in a continuum. The measure theoretic definition allows one to treat both the cases in a unified framework.

Definition 4.3. The *cumulative distribution function (or cdf in short)* of a random variable X given is defined as a non-negative function on $\mathbb{R}$ through the distribution μ_X:

$$F_X(x) = \mu_X((-\infty, x]), \;\; x \in \mathbb{R}.$$

Two different random variables can have the same cdf. For instance, we have seen in the lecture that the random variables $X_1(\omega) = 2\omega$ and $X_2(\omega) = 2(1 - |2\omega - 1|)$ defined on the probability space $([0, 1], \mathcal{B}([0, 1]), m)$ have the same cdf $F_X(x)$ given by

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty \leq x < 0, \\ \frac{x}{2} & \text{if } \quad 0 \leq x \leq 2, \\ 1 & \text{if } \quad x > 2. \end{cases} \tag{4.2}$$

The above example of explicitly stating the functional form of the random variable on a probability space is not a standard practice. In a real world scenario where a random variable is used to model the outcome of an experiment, one does not know explicitly the functional form. However, in many situations one can estimate its cdf. Thus the cdf of a random variable in some sense can be considered its model but it is not an attempt to model the functional form of the random variable.

Note: Two random variables X and Y are said to be identically distributed if they have the same cdf. It can also be show that given a cdf of a random variable there always exists a distribution of a random variable which would give back the cdf. Hence, we can also say two random variables X and Y are said to be identically distributed if they have the same distribution.

Proposition 4.2. *Let F be the cdf of a random variable X.*

- *For $x_1 < x_2$, $F(x_1) \leq F(x_2)$ (i.e., F is nondecreasing on $\mathbb{R}$).*
- *For $x \in \mathbb{R}$, $F(x) = \lim_{y \downarrow x} F(y)$ (i.e., F is right continuous on $\mathbb{R}$).*
- $\lim_{x \to -\infty} F(x) = 0$ *and* $\lim_{x \to +\infty} F(x) = 1$.

4.1 Probability density functions

For this course, one has to unlearn of what all has learnt about probability density functions in a basic course. A probability density function (pdf) is NOT the derivative of the cdf by definition but only happens to be so sometimes.

Definition 4.4. A probability measure μ_X on $\mathcal{B}(\mathbb{R})$ is said to be **absolutely continuous with respect to the Lebesgue measure** m, and written as

$$\mu_X \ll m$$

if $m(E) = 0$ implies that $\mu_X(E) = 0$ for every $E \in \mathcal{B}(\mathbb{R})$.

The proof of the following theorem is not part of the syllabus.

Theorem 4.1. (A simplified version of Radon-Nikodym Theorem) *Let X be a real-valued random variable with a distribution μ_X. If $\mu_X \ll m$ then there is a $h \in L^1(m)$ such that $\mu_X(E)$ of every E in $\mathcal{B}(\mathbb{R})$ can be expressed as the Lebesgue integral over E:*

$$\mu_X(E) = \int_E h dm$$

The function h above is unique up to a set of measure zero, i.e., if $\mu_X(E) = \int_E g dm$ for every $E \in \mathcal{B}$, then $g = h$ almost everywhere. Any such function h in the above theorem is called a Radon-Nikodym derivative of μ_X w.r.t to m and is denoted by $\frac{d\mu_X}{dm}$.

Definition 4.5. A random variable X has a *probability density function (pdf)* if its distribution (induced measure) μ_X is absolutely continuous with respect to the Lebesgue measure m, and a pdf of X is any function which is a Radon-Nikodym derivative of μ_X w.r.t to m.

Let $h = \frac{d\mu_X}{dm}$ be a pdf of a random variable X. If h is Riemann integrable, and $[a, b]$ is an interval of $\mathbb{R}$ then $\int_{[a,b]} h dm$ is equal to the Riemann integral $\int_a^b h(x)dx$. Hence the cdf $F_X(x)$ and a Riemann integrable pdf h are related by $F_X(x) = \int_{-\infty}^x h(t)dt$.

To consider an example of relating the cdf with a pdf, consider the example of a random variable whose cdf is given by (4.2). In principle, both the functions h_1 and h_2 below are both pdfs of X as they both can be verified to be the Radon-Nikodym derivatives of the distribution of X. However, it is a standard practice it is a convention to choose a Riemann integrable pdf (h_1 in this case) and call it "the pdf of" X. This is since Riemann integrable pdfs are useful in applications (see (6.2) and the discussion following it). Note that here h_1 and h_2 differ only on a set of Lebesgue measure zero (they differ only the set of rationals in $[0, 1]$).

$$h_1(x) = \begin{cases} 0 & \text{if } -\infty \leq x < 0, \\ \frac{1}{2} & \text{if } 0 \leq x \leq 2, \\ 0 & \text{if } x > 2. \end{cases} \tag{4.3}$$

and

$$h_2(x) = \begin{cases} \frac{1}{2} & \text{if } 0 \leq x \leq 2 \text{ and } x \text{ is an irrational}, \\ 0 & \text{elsewhere.} \end{cases} \tag{4.4}$$

Chapter 5

Random Vectors and Joint Densities

The definitions of a probability distribution, cdf and pdf for a random variable can be naturally extended to a random vector and we do that here.

Definition 5.1. Let $(\Omega, \mathcal{F}, P)$ be a probability space, $k \in \mathbb{N}$ and $X : \Omega \to \mathbb{R}^k$ be $(\mathcal{F}, \mathcal{B}(\mathbb{R}^k))$-measurable, i.e., $X^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{B}(\mathbb{R}^k)$. Then X is called a (k-dimensional) random vector defined on $(\Omega, \mathcal{F}, P)$.

Let $X = (X_1, X_2, \ldots X_k)$ be a random vector with components X_i, $i = 1, 2, \ldots, k$. Then each X_i is a random variable on $(\Omega, \mathcal{F}, P)$. this follows from the fact that the coordinate projections maps from $\mathbb{R}^k$ to $\mathbb{R}$, given by

$$\pi_i(x_1, x_2, \ldots, x_k) \equiv x_i, \quad 1 \leq i \leq k$$

are continuous and hence, are Borel measurable. Conversely, if for $1 \leq i \leq k$, X_i is a random variable on $(\Omega, \mathcal{F}, P)$, then $X = (X_1, X_2, \ldots, X_k)$ is a random vector (this is not proved here, but it is good to know this).

Definition 5.2. Let X be a k-dimensional random vector on $(\Omega, \mathcal{F}, P)$ for some $k \in \mathbb{N}$. Let

$$F_X(x) = P(\{\omega : X_1(\omega) \leq x_1, X_2(\omega) \leq x_2, \ldots, X_k(\omega) \leq x_k\})$$

for $x = (x_1, x_2, \ldots, x_k) \in \mathbb{R}^k$. Then $F_X(\cdot)$ is called the *joint cumulative distribution* function (joint cdf) of the random vector X.

Definition 5.3. Let X be a k-dimensional random vector on $(\Omega, \mathcal{F}, P)$ for some $k \in \mathbb{N}$. Let

$$\mu_X(A) = P(X^{-1}(A)) \quad \text{for all} \quad A \in \mathcal{B}(\mathbb{R}^k).$$

The probability measure μ_X is called the *(joint) probability distribution* if X.

Let $X = (X_1, X_2, \ldots X_k)$ be a random vector. Let $Y = (X_{i_1}, X_{i_2}, \ldots, X_{i_r})$ for some $1 \leq i_1 < i_2 < \cdots < i_r \leq k$, where $1 \leq r \leq k$. Then Y is also a random vector. Further, the joint cdf of Y can be obtained from F_X by setting the component x_j, $j \notin \{i_1, i_2, \ldots, i_r\}$ equal to ∞. For e.g., if $(i_1, i_2, \ldots, i_r) = (1, 2, \ldots, r)$, $r < k$ then

$$F_Y(y_1, y_2, \ldots, y_r) = F_X(y_1, \ldots, y_r, \infty, \ldots, \infty)$$

and

$$\mu_Y(A) = \mu_X(A \times \mathbb{R}^{(k-r)}), \quad A \in \mathcal{B}(\mathbb{R}^r).$$

Definition 5.4. Let $X = (X_1, X_2, \ldots X_k)$ be a random vector. Then for each $i = 1, \ldots, k$, the cdf F_{X_i}, and the probability distribution μ_{X_i} are called the *marginal* cdf and *marginal probability distribution* of X_i, respectively.

CAUTION It is clear that the distribution of X determines the marginal distribution μ_{X_i} of X_i, for all $i = 1, 2, \ldots, k$. However, the marginal distributions $\{\mu_{X_i} : i = 1, 2, \ldots, k\}$ do not determine the joint distribution without additional conditions such as independence (see next Chapter 7).

One can define probability density function for a random vector in the same way as it was done in the one-dimensional case earlier. First we define absolutely continuity with respect to the k-dimensional Lebesgue measure m_k similar to Definition 4.4:

A probability measure μ_X on $\mathcal{B}(\mathbb{R}^k)$ is said to be **absolutely continuous with respect to the k-dimensional Lebesgue measure** m_k, and written as

$$\mu_X \ll m_k$$

if $m_k(E) = 0$ implies that $\mu_X(E) = 0$ for every $E \in \mathcal{B}(\mathbb{R})$.

The Radon Nikodym Theorem (Theorem 4.1) can be stated for any finite dimensional Lebesgue measure m_k and probability measure on $\mathcal{B}(\mathbb{R})$ (just replace m by m_k & $\mathcal{B}(\mathbb{R})$ by $\mathcal{B}(\mathbb{R}^k)$ in Theorem 4.1.

Definition 5.5. A random vector $X = (X_1, X_2, \ldots X_k)$ has a *probability density function (pdf)* if its induced measure or joint distribution μ_X is absolutely continuous with respect to the Lebesgue measure m_k, and a pdf of X is any function which is a Radon-Nikodym derivative of μ_X w.r.t to m_k.

Chapter 6

Moments

Definition 6.1. Let X be random variable on $(\Omega, \mathcal{F}, P)$. The expected value of X, denoted by $E(X)$ is defined as

$$E(X) = \int X dP. \tag{6.1}$$

The expected value of a random variable is often called the mean of the random variable.

Definition 6.2. The variance of a random variable X is defined as

$$\text{Var}(X) = E((X - E(X))^2$$

provided $E(X^2) < \infty$.

It is not convenient to evaluate Lebesgue integrals for computing the mean and variance. When density functions exist, they can have a convenient form using Riemann integrals as in (6.2). We do not prove Theorem 6.1 in the lecture since other background material is not covered. However, note (6.2) and see how mean and variance of X can be computed via Riemmann integrals when $E(|X|) < \infty$ and $E(X^2) < \infty$ are satisfied.

Theorem 6.1. Let X be a random variable on $(\Omega, \mathcal{F}, P)$. Let $g : \mathbb{R} \to \mathbb{R}$ be a Riemann integrable function on $\mathbb{R}$. Suppose that X has a probability density function $h_X(x)$ then

$$E(|g(X)|) = \int_{-\infty}^{\infty} |g(x)| h_X(x) dx.$$

If $E(|g(X)|) < \infty$, then

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) h_X(x) dx. \tag{6.2}$$

The usefulness of equation (6.2) does not diminish the usefulness of defining expectation as in (6.1) using a Lebesgue integral. For instance, to derive $E(X+Y) = E(X) + E(Y)$ using only the Riemann integral in (6.2) is tedious.

The following inequalities (proven in the lecture) are frequently employed in proving other results such as the "weak law of large numbers" (see Chapter 10). It is important that you know how to prove them.

Proposition 6.1. (Markov's inequality) Let X be random variable on $(\Omega, \mathcal{F}, P)$. Then for any $t > 0$

$$P(|X| \geq t) \leq \frac{E(|X|)}{t}.$$

Proposition 6.2. (Chebyshev's inequality) Let X be random variable on $(\Omega, \mathcal{F}, P)$ whose variance is well defined. Then for any $\epsilon > 0$,

$$P(|X - E(X)| \geq \epsilon) \leq \frac{\mathrm{Var}(X)}{\epsilon^2}.$$

6.1 Covariance and auto-correlation

So far we considered expectations and moments of individual random variables. We can also take expectations on joint densities of two random variables — this is very useful as we shall see in the later chapters.

Definition 6.3. The correlation between two random variables X and Y is defined as $E(XY)$; here $E(\cdot)$ is either the Lebesgue integral $\int XY dP$ or the Riemann Integral $\int_{-\infty}^{\infty} xy h_{XY}(x,y) dx dy$, where h_{XY} is the joint pdf of X and Y.

Definition 6.4. The *covariance* between two random variables X and Y is defined by

$$\mathrm{Cov}(X,Y) = E((X - E(X))(Y - E(Y)).$$

Definition 6.5. Given a random vector $X = (X_1, X_2, \ldots X_k)$, the covariance matrix C of X is given by its elements C_{ij} (element positioned in the i^{th} row and j^{th} column), where $C_{ij} := \mathrm{Cov}(X_i, X_j)$.

Remark 6.1. The diagonal elements of a covariance matrix are the variances of the individual random variables in the random vector. A covariance matrix is positive semi-definite (google it).

Chapter 7

Independence

In this chapter we will define the concept of independence of random variables. Although a probability space is nothing more than a measure space space with the measure of the whole space equal to one, probability theory is not just a subset of measure theory. A distinct feature of measure theory is the notion of independence. Here, you will find more than one definition of independence and you may adopt anyone that appeals to you:

Definition 7.1. Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\mathcal{G}$ and $\mathcal{H}$ be sigma-algebras such that $\mathcal{G} \subset \mathcal{F}$ and $\mathcal{H} \subset \mathcal{F}$ (i.e., the sets in $\mathcal{G}$ and $\mathcal{H}$ are also in $\mathcal{F}$). We say $\mathcal{G}$ and $\mathcal{H}$ are independent if

$$P(A \cap B) = P(A) \cdot P(B) \text{ for all } A \in \mathcal{G}, B \in \mathcal{H}.$$

Definition 7.2. Consider a real valued random variable X defined on $(\Omega, \mathcal{F}, P)$. Then then the set $\sigma(X) = \{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}$ is called the σ-algebra generated by X. (Note : $\sigma(X)$ has nothing to do with P)

Definition 7.3. Let X and Y be two real valued random variables defined on $(\Omega, \mathcal{F}, P)$. Then the following definitions are equivalent:

1. The random variables X and Y are said to be independent if $\sigma(X)$ and $\sigma(Y)$ are independent.
2. X and Y are said to be independent if the joint distribution μ_{XY} factors:

   $$\mu_{XY}(A \times B) = \mu_X(A) \cdot \mu_Y(B)$$

   for all subsets A and B belonging to $\mathcal{B}(\mathbb{R})$.
3. X and Y are said to be independent if the joint CDF F_{XY} factors:

   $$F_{XY}(a, b) = F_X(a) \cdot F_Y(b) \text{ for all } a \in \mathbb{R}, b \in \mathbb{R}$$

4. If the joint density h_{XY} is defined, then X and Y are said to be independent if

$$h_{XY}(a,b) = h_X(a) \cdot h_Y(b) \text{ for almost all } a \in \mathbb{R}, b \in \mathbb{R}$$

where almost all is with respect to the Lebesgue measure.

Note that the independence of two random variables X and Y not only depends on them, but also the measure P. That is to say, that if X and Y are independent on $(\Omega, \mathcal{F}, P)$ they may not be independent on $(\Omega, \mathcal{F}, \hat{P})$, where $P \neq \hat{P}$ (look back at the examples in the lecture).

Independence is a very strong notion of measuring the uncommonness between two random variables. A weaker notion is of uncorrelatedness:

Definition 7.4. Two random variables X and Y are *uncorrelated* if $E(XY) = E(X)E(Y)$

Note : Two random variables can be uncorrelated but not independent, but independent random variables are always uncorrelated. Two uncorrelated random variables have zero covariance.

The following result is important:

Theorem 7.1. *Let X and Y be independent random variables and let $f : \mathbb{R} \to \mathbb{R}$ be Borel-measurable functions on $\mathbb{R}$. Then $f(X)$ and $g(Y)$ are independent random variables.*

Proof. It follows from the fact that $\sigma(f(X)) \subset \sigma(X)$ and $\sigma(g(Y)) \subset \sigma(Y)$.

The rest of this chapter is out of the syllabus. It is for generalizing the above to define define independence of a sequence of random variables.

Definition 7.5. Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\{B_1, B_2, \ldots, B_n\} \subset \mathcal{F}$ be a finite collection of events.

1. $B_1, B_2, \ldots B_n$ are called independent (w.r.t. P) if

$$P(\bigcap_{j=1}^{k} B_{i_j}) = \prod_{j=1}^{k} P(B_{i_j})$$

for all $\{i_1, i_2, \ldots i_k\} \subset \{1, 2, \ldots, n\}$, $1 \leq k \leq n$.

2. $B_1, B_2, \ldots B_n$ are called pairwise independent (w.r.t. P) if $P(B_i \cap B_j) = P(B_i)P(B_j)$, for all i, j.

Definition 7.6. Let $(\Omega, \mathcal{F}, P)$ be a probability space. A (finite of infinite) collection of events $\{B_\alpha : \alpha \in I\} \subset \mathcal{F}$ is called independent w.r.t. P if for every finite sub-collection $\{\alpha_1, \alpha_2, \ldots \alpha_k\} \subset I$,

$$P(\bigcap_{i=1}^{k} B_{\alpha_i}) = \prod_{j=1}^{k} P(B_{\alpha_i}).$$

Definition 7.7. Consider two measurable spaces $(\Omega, \mathcal{F})$ and $(\mathbb{R}, \mathcal{B})$ and let $X : \Omega \to \mathbb{R}$, then the set $\sigma(X) = \{X^{-1}(B) : B \in \mathcal{B}\}$ is called the σ-algebra generated by X.

Definition 7.8. . (Independent random variables) Let $\{X_\alpha : \alpha \in I\}$ be a collection of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$. Then the $\{X_\alpha : \alpha \in I\}$ is called an independent collection of random variables w.r.t P if the family of σ-algebras $\sigma(X_\alpha) : \alpha \in I\}$ is independent w.r.t P.

Remark 7.1. Note that $\{X_1, X_2, \ldots, X_k\}$ of random variables are independent if and only if their joint cdf $F(x_1, x_2, \ldots x_n) = \prod_{i=1}^{k} F_{X_i}(x_i)$, where F_{X_i} are the individual cdfs of X_i. This is since by definition $F(x_1, x_2, \ldots x_n) = P(X_i \leq x_i, i = 1, 2, \ldots, k) = \prod_{i=1}^{k} P(X_i \leq x_i)$.

Chapter 8

Gaussian Random Variables

A real valued random variable X is said be Gaussian if its cdf is given by (the Riemann integral)

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{2\pi\sigma^2} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt,$$

where μ and σ^2 are the mean and variance of X.

Thus if

$$f_X(x) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

is a pdf of a random variable X, where μ and σ^2 are the mean and variance of X, then X is Gaussian.

More generally, $\mathbb{R}^n$ for all $1 \leq n < \infty$, a random vector X is (jointly) Gaussian whenever its pdf is given by

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n \det(\mathrm{C})}} \exp -\frac{1}{2}(x-m)C^{-1}(x-m)^T$$

where m is the mean vector of X and $\det(C)$ denotes the determinant of the covariance matrix of X.

A Gaussian distributed random variable is also called a "normally distributed" random variable. Special properties of Gaussian Random variables (also see Chapter 3 of Prof. Herbert's notes):

- A Gaussian random variable has a special property: the mean and the variance (or its covariance matrix) determine its cdf completely. It is standard to denote the pdf of a Gaussian random variable with mean m and variance σ^2 by $\mathcal{N}(m, \sigma^2)$; $\mathcal{N}$ is meant to identify the phrase "normal".
- If two Gaussian random variables are uncorrelated then they are independent.

- Linear combinations of jointly normal Gaussian random variables is again jointly Gaussian.
- If a random variable is jointly normal then it can be expressed as the linear combinations of independent Gaussian random variables (a proof is given in the lecture when two randomly variables are jointly normal).

Despite all the special properties Gaussian random variables enjoy, it does not necessarily mean that any two Gaussian random variables may not have their joint density to be normal. For example, let (X, Y) be a pair of random variables with the non-Guassian (joint) pdf:

$$f_{X,Y}(x,y) = \begin{cases} \frac{2|x|+y}{\sqrt{2\pi}} \exp -\frac{(2|x|+y)^2}{2} & \text{if} \quad y \geq -|x|, \\ 0 & \text{if} \quad y < -|x|. \end{cases}$$

One may verify here that the marginal pdfs of $f_{X,Y}$ are Gaussian.

Chapter 9

Stochastic Processes

The conception of a finite dimensional random vector were discussed earlier. Infinite families of random variables such as a random sequence $\{X_n\}_{n\geq 1}$ or a random function $\{X(t) : 0 \leq t < T\}$, $0 \leq T \leq \infty$ are also important notions that can be considered for dealing with applications. For example, X_n could the population size of the nth generation of a randomly evolving biological population, and $X(t)$ could be the temperature at time t in a chemical reaction over a period $[0, T]$.

A **stochastic process** or a **random process** is a collection of random variables , $\{g_\alpha : \alpha \in I\}$ on a common probability space, where I is some index set. In practice, I is usually a subset of the integers or of the real line, with $\mathbb{N}$ and $[0, \infty)$ being particularly common choices in which the index represents, respectively, the discrete or the continuous flow of time. In this notes, the index set is either the set of all integers $\mathbb{Z}$ or $\mathbb{N}$. For ease of notation we consider the index set to be $\mathbb{N}$ in this notes, however all concepts have a natural extension when the index set is $\mathbb{Z}$ as well.

The material in next two paragraphs (till the end of Theorem 9.1) are not part of the syllabus — it is for those who may wish to know the existence of a stochastic process.

In the previous chapters we have seen that we can define a finite-dimensional random vectors induce a probability measure. One may wonder if we can extend the dimension of the random vectors to infinity, whether there would be a probability space on the infinite dimensional space which is consistent with its finite dimensional distributions? The answer was provided by Kolmogorov in a remarkable result which is only stated here in Theorem 9.1 (again, its discussion or proof is not part of the syllabus) in the particular case when the index set $I = \mathbb{N}$. This celebrated consistency theorem which establishes the existence of a stochastic process provided the finite-dimensional descriptions of the random phenomenon have no inconsistencies. In other words, given a consistent family of finite dimensional distributions, one can construct a unique probability measure in an infinite-dimensional space. For this reason, the consistency theorem is also known as Kolmogorovs extension theorem. This

single result showed the existence of a variety of processes such as Brownian motion, Poisson process, Markov processes with regular transition probabilities, (which are not studied here).

Theorem 9.1. *Let $\{\mu_n\}_{n\geq 1}$ be a sequence of probability measures such that*

1. *for each $n \in \mathbb{N}$, μ_n is a probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$,*
2. *for each $n \in \mathbb{N}$, $\mu_{n+1}(B \times \mathbb{R}) = \mu_n(B)$ for all $B \in \mathcal{B}(\mathbb{R}^n)$.*

Then there exists a stochastic process $\{X_n : n \geq 1\}$ on some probability space $(\Omega, \mathcal{F}, P)$ such that for each $n \geq 1$, the distribution $\mu_{(X_1,X_2,\ldots,X_n)}$ (as defined in Definition 5.3) of the random vector $(X_1, X_2, \ldots, X_n)$ is μ_n.

9.1 Stationary Process

Consider a stochastic process $\{X_n\}$. A particular class of processes where the random variables are identically distributed makes life little easier for statisticians. For such identically distributed processes, the mean and variance and all moments do not change with time. However, more desirable for statisticians is to assume consistent dependence (in terms of their joint density) on adjacent random variables — it is also an important hypothesis in many real world processes such dependence exists. For instance, in a process $\{X_n\}$, the random variables $X_{100}, X_{101}, X_{101}$ may have a similar dependence among them as the random variables X_1, X_2, X_3 have in them. To capture this idea, a special subclass of identically distributed processes called "stationary processes" is defined:

Definition 9.1. Let $\{X_n : n \geq 1\}$ be a stochastic process defined on a probability space. The process is $\{X_n\}$ is said to be stationary if the joint cdf of $X_1, X_2, \ldots, X_k$ is equal to that of $X_n, X_{n+1}, \ldots, X_{n+k-1}$ for every $n, k \in \mathbb{N}$, i.e., if for every $(x_1, x_2, \ldots, x_k)$,

$$F_{X_1,\ldots,X_k}(x_1, x_2, \ldots, x_k) = F_{X_n,\ldots,X_{n+k-1}}(x_1, x_2, \ldots, x_k) \ \forall n, k, \in \mathbb{N},$$

where $F_{X_j,\ldots,X_{j+r}}$ denotes the joint density of the random variables $X_j, X_{j+1}, \ldots, X_{j+r}$.

The simplest example of a stationary process is a process in which all random variables are independent and identically distributed, i.e., an iid process (it is a good exercise to check this).

Another example is $\{X_n = \cos(n + Y)\}_{n\geq 1}$ where Y is an uniformly distributed random variable taking values in $[0, 2\pi]$. You may not worry about examples of stationary processes, but it is important to clearly understand its definition.

Chapter 10

Stochastic Convergence

Definition 10.1. Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$.

1. If $X(\omega) = \lim_{n\to\infty} X_n(\omega)$ for all $\omega \in \Omega$, then X_n is said to converge to X *pointwise*.
2. If $X(\omega) = \lim_{n\to\infty} X_n(\omega)$ for every $\omega \in A$, where $P(A) = 1$, then X_n is said to converge to X *almost everywhere* or *almost surely* or with *probability* 1.
3. If for every $\epsilon > 0$, $\lim_{n\to\infty} P(\{|X_n - X| > \epsilon\}) = 0$, then X_n is said to converge to X in *probability*.
4. If $\lim_{n\to\infty} F_n(x) = F(x)$ for every $x \in \mathbb{R}$ where $F(x)$ is continuous, then X_n is said to converge in *distribution* to X; here F_n is the cdf of X_n and F that of X.

Remark 10.1. Pointwise convergence $\Longrightarrow$ almost sure convergence $\Longrightarrow$ convergence in probability $\Longrightarrow$ convergence in distributions. The first $\Longrightarrow$ is straightforward. The other two implications are proved in the lectures.

In the case of an infinite coin tossing experiment, let $X_i(\omega) = 0$ or 1 for each i (here each ω denotes an individual infinite tossing experiment) depending on whether the i^{th} coin is tails or heads. If the coin is assumed to be unbiased, intuitively one expects the fraction of heads gets closer to $\frac{1}{2}$ (as the number of tosses increases) with high probability. This is since the outcome of X_i does not influence the outcome of any X_j $(j \neq i)$). Informally, this notion of independence is thought out in our intuition while expecting the fraction of heads converges to $\frac{1}{2}$.

The weak law of large numbers proved by Kolmogorov gives a formal treatment of such an intuitive feeling. It is enough that the random variables are uncorrelated instead of demanding independence.

Theorem 10.1. (Weak Law of Large numbers) *Let $\{X_i\}$ be a sequence of uncorrelated random variables with identical mean $E(x)$ and variance σ^2. Then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \text{ converges in probability to } E(X).$$

Proof Outline. Use the fact the variance of a finite sum of uncorrelated random variables is equal to the sum of their individual variances and apply Chebyschev's inequality (proved in the lecture).

Note that the average in the weak law of large numbers $\frac{1}{n}\sum_{i=1}^{n} X_i$ has a variance which tends to zero as $n \to \infty$. As the variance tends to zero the behavior of the average tends to that of a degenerate random variable (a constant random variable). Often, in practice, one considers (types of) averages where the variance does not tend to zero, in particular where the variance can be kept a constant. For instance, the variance of the average $\frac{1}{\sqrt{n\sigma^2}}\sum_{i=1}^{n} X_i$ is independent of n if X_i are iid. One may wonder, what would happen to such an average as $n \to \infty$. In fact the average of this type may not converge to any given ω, but informally it behaves like a random variable as $n \to \infty$. Interestingly, the behavior of the average tends to that of a Gaussian random variable and rather shockingly, it is independent of the distribution of X_i. This result is known as the Central limit theorem.

Theorem 10.2. (Central limit theorem) *Let $X_1, X_2, \ldots$ be an iid process such that $E(|X_i|^r) < \infty$ for every $r \in \mathbb{N}$. Then*

$$Z_n = \frac{1}{\sqrt{n\sigma^2}}\sum_{i=1}^{n} X_i - m \text{ converges in distribution to } \mathcal{N}(0,1)$$

as $n \to \infty$, where m and σ^2 are the mean and variance of X_i.

Proof Outline. One can show that the characteristic function of Z_n converges to a Gaussian function as $n \to \infty$. The characteristic function of a random variable is Gaussian if and only if it has a Gaussian distribution. Hence, Z_n tends to a Gaussian : the characteristic function of Z_n has a valid Taylor series expansion thanks to $E(|X_i|^r) < \infty$ and from this one can show that the higher terms in the Taylor series as $n \to \infty$ tend to zero. The non vanishing terms of the Taylor series tend to that of a Gaussian function.

Appendix A

Appendix : Elementary Mathematical Background

This appendix may be a bit superfluous, but it is worth having it in the notes (especially to those who believe they can love math if the story is complete.

In this chapter, we catalogue some basic definitions and results from topology and analysis which is for completeness of the earlier chapters. The material is in no sense detailed or exhaustive, but many standard references are available for understanding.

A.1 Topological spaces and metric spaces

In this section, we only look at formal definitions of basic notions of point set topology and metric spaces. Topological spaces are more general than metric spaces. Apart from the definition of "topology", a few more definitions of classical notions are given to make a non-abrupt transition to metric spaces. We assume the reader is aware of the basic operations of set theory and the notion of a metric.

Definition A.1. A **topology** on a set X is a collection $\mathcal{T}$ of subsets of X, called open sets, satisfying three conditions :

(i) The emptyset $\emptyset$ and the set X are in $\mathcal{T}$.
(ii) The union of the elements of any arbitrary subcollection of $\mathcal{T}$ is in $\mathcal{T}$.
(iii) The intersection of the elements of any finite subcollection of $\mathcal{T}$ is in $\mathcal{T}$.

A set X for which a topology $\mathcal{T}$ has been specified is called a *topological space*. A topological space is represented as a pair $(X, \mathcal{T})$ consisting of a set X and a topology $\mathcal{T}$ on X. Often for brevity, the specific mention of $\mathcal{T}$ is omitted and the phrase 'X is a topological space' is used.

Looking at the definition above, an immediate question that arises is 'why does anyone *care* about open sets?' The answer is : concepts like continuity of functions, compactness and connectedness of sets etc. can all be defined with the aid of open sets without mention to any metric or distance. Even in the case where a metric is used to define these notions, once the collection of all open sets is determined, there will be no need to refer to distance again. Further explanation can be found in books like [3][2].

Definition A.2. If X is a set, a *basis* for a topology on X is a collection $\mathfrak{B}$ of subsets of X (called **basic elements**) such that

(i) For each $x \in X$, there is at least one basis element B containing x.
(ii) If x belongs to the intersection of two basis elements B_1 and B_2, then there is a basis element B_3 containing x such that $B_3 \subset B_1 \cap B_2$.

Definition A.3. *If $\mathfrak{B} = B_\alpha$ is a basis for a topology on X, then the* **topology generated** *by $\mathfrak{B}$ is defined as the collection of all sets, each of which is given by a union of elements in $\mathfrak{B}$.*

The above definition makes sense only if the collection of the union of sets in $\mathfrak{B}$ satisfies the three axioms for a topology. It indeed satisfies the three axioms and the proof can be found in any introductory book on Topology. Returning to the basis for a topology, we have not said as to how the basis can be got. Though it can be got by different means, one of the most important and frequently used ways of imposing a topology on a set is to define the topology in terms of a metric on the set. Topologies given in this way lie at the heart of modern analysis and for any further understanding of topology in this notes, it is sufficient to consider only such topologies.

Given a metric d on X, with the positive real numbers $d(x, y)$ and ϵ, the set

$$B(x, \epsilon) = \{y \ : \ d(x, y) < \epsilon\}$$

is referred to as the open ball or just the ball at center y with radius ϵ. Now, it can be easily verified that the collection of all open balls centered at each $y \in X$ with all possible different radii satisfies the definition of a basis for a topology on X.

Definition A.4. If d is a metric on the set X, the collection of all ϵ–balls $B(x, \epsilon)$, for $x \in X$ and $\epsilon > 0$, is a basis for a topology on X called the *metric topology* induced by d.

We will use the term "metric space" to mean a topological space such that there is an underlying metric through which open balls are obtained and hence the basis generating the topology. The following definitions are made where the setting is a metric space. The definitions also can be extended for an arbitrary topological space (i.e. topologies which cannot be obtained by a metric). We make no reference to non-metrizable topologies (nor do we encounter them) and since the definitions in a metric space are more familiar to the

non-mathematician, the following definitions are defined on metric spaces though it could have been more general. It is to be kept in mind that a metric space is also a topological space and any set in the topology is called as an open set. With this idea behind, several topological notions concerning sets are defined.

A subset A of a metric space X is said to be *closed* if the complement of the set A relative to X is open. Therefore, trivially the whole space is both open and closed. We next look at the union and intersection of open and closed sets as we encounter them often. The union and intersection of open sets and closed sets can be characterized by the following Theorem([3, pp. 92-93]) :

Theorem A.1. *Let X be a metric space. Then the following conditions hold : (i) The empty set $\emptyset$ and X are closed. (ii) Arbitrary intersections of closed sets are closed. (iii) Finite unions of closed sets are closed.*

Definition A.5. Let X be a topological space with topology $\mathcal{T}$. If Y is a subset of X, the collection

$$\mathcal{T}_Y = \{Y \cap U \, : \, U \in \mathcal{T}\}$$

is a topology on Y, called the **subspace topology**.

Definition A.6. Given a subset A of a metric space X, the *interior* of A is defined as the union of all open sets contained in A, and the *closure* is defined as the intersection of all closed sets containing A.

• The interior of A is denoted by A^o and the closure of A by $\bar{A}$. A set $A \subseteq$ of a metric space X is said to be dense in X if $\bar{A} = X$.

The following notion of compactness is frequently used in point set topology:

• By an *open cover* of a set E in a metric space X we mean a collection $\{G_\alpha\}$ of open subsets of X such that $E \subseteq \cup_\alpha G_\alpha$.

Definition A.7. A subset K of a metric space X is said to be *compact* if every open cover of K contains a *finite* subcover.

We give one example to illustrate its definition. To appreciate the notion of compactness, it would be helpful to look at more examples in textbooks.

Example A.1. Consider the space $(0, 1)$ with standard metric $d(x, y) : |x-y|$. Now consider the collection of the open sets $\{(\frac{1}{n}, 1), n = 1, 2, \ldots\}$. Clearly this collection is an open cover of the entire space. However it should be verified that there is no finite open cover that can be extracted from this cover. Hence the space $(0, 1)$ is not compact even though the whole space is closed.

For metric spaces, compactness can be understood through sequences. A metric space is compact if and only if every sequence has a convergent subsequence [3, Chapter 3]. This is

an important result as in many mathematical frameworks or models since it is a desirable property that the limits of every convergent sequence exists. We shall see later in this notes that compactness is a useful property in getting to several results.

One of the natural questions that may arise whether there are uncountable sets with zero Lebesgue measure. One such construction can be got by what is called the Cantor's middle third set (google it). The following definitions help in defining a general Cantor set.

Definition A.8. A set S is called *disconnected* if there are two open, non-empty sets U and V such that and $U \cap S \neq \emptyset$ and $V \cap S \neq \emptyset$ satisfying $U \cap V = \emptyset$ and $U \cup V \supseteq S$. A set is said to be *connected* if it is not disconnected.

• A point $x \in S$ is called *accumulation point* or a *limit point* of S, if every neighborhood of x contains infinitely many distinct points of S.

• A set S is said to be *perfect* if every point of the set is an accumulation point of S.

• A totally disconnected set is a disconnected set whose interior is empty.

• A set is said to be a *Cantor set* if it is closed, perfect and totally disconnected.

We have encountered the definition of an infinite set which is uncountable and totally disconnected like the one of a Cantor set and also a set like an open ball in $\mathbb{R}$ which is a connected set. Though both sets are uncountable, geometrically one gets a feeling that the connected set is a larger set than the disconnected one.

Limits

We do not review all the definitions pertaining to limits as some of them are elementary and expected to be known to the reader. However, two definitions need mention as they are referred to extensively in this notes. In the following definitions, a sequence means a sequence of real numbers of some metric space.

Definition A.9. Let E be the set of all subsequential limits of a sequence $\{a_n\}$. Define $a_* = \inf E$ and $a^* = \sup E$. The numbers a_* and a^* are called as the lower and upper limits and represented as $\liminf_{n\to\infty} a_n$ and $\limsup_{n\to\infty} a_n$.

Note that regardless of whether the limit of a sequence exists or not, the lower and upper limits always exists. Two inequalities concerning the lim inf and lim sup of any two arbitrary real sequence $\{a_n\}$ and $\{b_n\}$ are

$$\liminf_{n\to\infty}(a_n + b_n) \geq \liminf_{n\to\infty}(a_n) + \liminf_{n\to\infty}(b_n)$$

and

$$\limsup_{n\to\infty}(a_n + b_n) \geq \limsup_{n\to\infty}(a_n) + \limsup_{n\to\infty}(b_n).$$

Equality holds in both the cases when one of the limits on the right hand side exists, i.e., either $\lim_{n\to\infty} a_n$ or $\lim_{n\to\infty} b_n$ exists.

Bibliography

[1] , Y.S. Chow and H. Teicher, "Probability Theory", 3^{rd}-edition, Springer, Berlin, 1997.

[2] J. G. Hocking and G.S. Young, "Topology", Dover publications, 1988.

[3] J. Munkres, "Topology", 2^{nd}-edition, Prentice Hall, 2000.

[4] W. Rudin, "Real and complex analysis", Mc-Graw Hill, 3^{rd}-edition, 1986.