MATH97020 STOCHASTIC DIFFERENTIAL EQUATIONS
ELEMENTS OF PROBABILITY THEORY

Professor G.A. Pavliotis

Department of Mathematics Imperial College London, UK
pavl@ic.ac.uk

TERM I 2020-2021

ELEMENTS OF PROBABILITY THEORY

1. PROBABILITY SPACES AND ELEMENTARY EXAMPLES

Definition

The set of all possible outcomes of an experiment is called the **sample space** and is denoted by Ω.

Example

- The possible outcomes of the experiment of tossing a coin are H and T. The sample space is $\Omega = \{H, T\}$.
- The possible outcomes of the experiment of throwing a die are $1, 2, 3, 4, 5$ and 6. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Definition

A collection $\mathcal{F}$ of Ω is called a **field** on Ω if

1. $\emptyset \in \mathcal{F}$;
2. if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$;
3. If $A, B \in \mathcal{F}$ then $A \cup B \in \mathcal{F}$.

From the definition of a field we immediately deduce that $\mathcal{F}$ is closed under finite unions and finite intersections:

$$A_1, \dots A_n \in \mathcal{F} \quad \Rightarrow \quad \cup_{i=1}^n A_i \in \mathcal{F}, \quad \cap_{i=1}^n A_i \in \mathcal{F}.$$

When Ω is infinite dimensional then the above definition is not appropriate since we need to consider countable unions of events.

Definition

A collection $\mathcal{F}$ of Ω is called a σ**-field** or σ**-algebra** on Ω if

1. $\emptyset \in \mathcal{F}$;
2. if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$;
3. If $A_1, A_2, \dots \in \mathcal{F}$ then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

A σ-algebra is closed under the operation of taking countable intersections.

Example

- $\mathcal{F} = \{\emptyset, \Omega\}$.
- $\mathcal{F} = \{\emptyset, A, A^c, \Omega\}$ where A is a subset of Ω.
- The **power set** of Ω, denoted by $\{0, 1\}^{\Omega}$ which contains all subsets of Ω.

- Let $\mathcal{F}$ be a collection of subsets of Ω. It can be extended to a $\sigma-$algebra (take for example the power set of Ω).
- Consider all the $\sigma-$algebras that contain $\mathcal{F}$ and take their intersection, denoted by $\sigma(\mathcal{F})$, i.e. $A \subset \Omega$ if and only if it is in every $\sigma-$algebra containing $\mathcal{F}$. $\sigma(\mathcal{F})$ is a $\sigma-$algebra. It is the smallest algebra containing $\mathcal{F}$ and it is called the $\sigma-$**algebra generated by** $\mathcal{F}$.

Example

Let $\Omega = \mathbb{R}^n$. The σ-algebra generated by the open subsets of $\mathbb{R}^n$ (or, equivalently, by the open balls of $\mathbb{R}^n$) is called the **Borel σ-algebra** of $\mathbb{R}^n$ and is denoted by $\mathcal{B}(\mathbb{R}^n)$.

- Let X be a closed subset of $\mathbb{R}^n$. Similarly, we can define the Borel σ-algebra of X, denoted by $\mathcal{B}(X)$.
- A sub-σ–algebra is a collection of subsets of a σ–algebra which satisfies the axioms of a σ–algebra.
- The $\sigma-$field $\mathcal{F}$ of a sample space Ω contains all possible outcomes of the experiment that we want to study. Intuitively, the $\sigma-$field contains all the information about the random experiment that is available to us.

Definition

A **probability measure** $\mathbb{P}$ on the **measurable space** $(\Omega, \mathcal{F})$ is a function $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$ satisfying

1. $\mathbb{P}(\emptyset) = 0, \ \mathbb{P}(\Omega) = 1;$
2. For $A_1, A_2, \dots$ with $A_i \cap A_j = \emptyset$, $i \neq j$ then

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Definition

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ comprising a set Ω, a σ-algebra $\mathcal{F}$ of subsets of Ω and a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is a called a **probability space**.

Example

A biased coin is tossed once:
$\Omega = \{H, T\},\ \mathcal{F} = \{\emptyset, H, T, \Omega\} = \{0, 1\},\ \mathbb{P} : \mathcal{F} \mapsto [0, 1]$ such that
$\mathbb{P}(\emptyset) = 0,\ \mathbb{P}(H) = p \in [0, 1],\ \mathbb{P}(T) = 1 - p,\ \mathbb{P}(\Omega) = 1$.

Example

Take $\Omega = [0, 1],\ \mathcal{F} = \mathcal{B}([0, 1]),\ \mathbb{P} =$ Leb$([0, 1])$. Then $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

An important result from measure theory that is very useful in the study of stochastic differential equations is the **Borel-Cantelli Lemma**:

Theorem

Let $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and let* $\{A_n\} \subseteq \mathcal{F}$ *an infinite sequence of events. Assume that*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < +\infty.$$

Then

$$\mathbb{P}\left(\limsup_{n \to +\infty} A_n\right) = 0, \quad \textit{where} \ \limsup A_n = \bigcap_{n=1}^{+\infty} \bigcup_{k=n}^{+\infty} A_k.$$

In other words, if the sum of the probabilities of the events $\{A_n\}$ is finite, then the set of all outcomes that are "repeated" infinitely many times must occur with probability zero.

2. RANDOM VARIABLES AND PROBABILITY DISTRIBUTION FUNCTIONS

The function of the outcome of an experiment is a **random variable**, that is, a map from Ω to $\mathbb{R}$.

Definition

A sample space Ω equipped with a $\sigma-$field of subsets $\mathcal{F}$ is called a measurable space.

Definition

Let $(\Omega, \mathcal{F})$ and $(E, \mathcal{G})$ be two measurable spaces. A function $X : \Omega \to E$ such that the *event*

$$\{\omega \in \Omega : X(\omega) \in A\} =: \{X \in A\} \tag{1}$$

belongs to $\mathcal{F}$ for arbitrary $A \in \mathcal{G}$ is called a *measurable function* or *random variable*.

When E is $\mathbb{R}$ equipped with its Borel σ-algebra, then (1) can by replaced with

$$\{X \leqslant x\} \in \mathcal{F} \quad \forall x \in \mathbb{R}.$$

Let X be a random variable (measurable function) from $(\Omega, \mathcal{F}, \mu)$ to $(E, \mathcal{G})$. If E is a metric space then we may define *expectation* with respect to the measure μ by

$$\mathbb{E}[X] = \int_{\Omega} X(\omega)\, d\mu(\omega).$$

More generally, let $f : E \mapsto \mathbb{R}$ be $\mathcal{G}$–measurable. Then,

$$\mathbb{E}[f(X)] = \int_{\Omega} f(X(\omega))\, d\mu(\omega).$$

Let U be a topological space. We will use the notation $\mathcal{B}(U)$ to denote the Borel σ–algebra of U: the smallest σ–algebra containing all open sets of U. Every random variable from a probability space $(\Omega, \mathcal{F}, \mu)$ to a measurable space $(E, \mathcal{B}(E))$ induces a probability measure on E:

$$\mu_X(B) = \mathbb{P}X^{-1}(B) = \mu(\omega \in \Omega; X(\omega) \in B), \quad B \in \mathcal{B}(E). \tag{2}$$

The measure μ_X is called the *distribution* (or sometimes the *law*) of X.

Example

Let $\mathcal{I}$ denote a subset of the positive integers. A vector $\rho_0 = \{\rho_{0,i},\ i \in \mathcal{I}\}$ is a distribution on $\mathcal{I}$ if it has nonnegative entries and its total mass equals 1: $\sum_{i \in \mathcal{I}} \rho_{0,i} = 1$.

Consider the case where $E = \mathbb{R}$ equipped with the Borel $\sigma-$algebra. In this case a random variable is defined to be a function $X : \Omega \to \mathbb{R}$ such that

$$\{\omega \in \Omega : X(\omega) \leqslant x\} \subset \mathcal{F} \quad \forall x \in \mathbb{R}.$$

We can now define the probability distribution function of X, $F_X : \mathbb{R} \to [0, 1]$ as

$$F_X(x) = \mathbb{P}\Big(\big\{\omega \in \Omega \big| X(\omega) \leqslant x\big)\big\} =: \mathbb{P}(X \leqslant x). \tag{3}$$

In this case, $(\mathbb{R}, \mathcal{B}(\mathbb{R}), F_X)$ becomes a probability space.

The distribution function $F_X(x)$ of a random variable has the properties that $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to +\infty} F(x) = 1$ and is right continuous.

Definition

A random variable X with values on $\mathbb{R}$ is called **discrete** if it takes values in some countable subset $\{x_0, x_1, x_2, \dots\}$ of $\mathbb{R}$. i.e.: $\mathbb{P}(X = x) \neq x$ only for $x = x_0, x_1, \dots$.

With a random variable we can associate the **probability mass function** $p_k = \mathbb{P}(X = x_k)$. We will consider nonnegative integer valued discrete random variables. In this case $p_k = \mathbb{P}(X = k)$, $k = 0, 1, 2, \ldots$.

Example

The Poisson random variable is the nonnegative integer valued random variable with probability mass function

$$p_k = \mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots,$$

where $\lambda > 0$.

Example

The binomial random variable is the nonnegative integer valued random variable with probability mass function

$$p_k = \mathbb{P}(X = k) = \frac{N!}{n!(N-n)!} p^n q^{N-n} \quad k = 0, 1, 2, \dots N,$$

where $p \in (0,1)$, $q = 1 - p$.

Definition

A random variable X with values on $\mathbb{R}$ is called **continuous** if $\mathbb{P}(X = x) = 0 \, \forall x \in \mathbb{R}$.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X : \Omega \to \mathbb{R}$ be a random variable with distribution F_X. This is a probability measure on $\mathcal{B}(\mathbb{R})$. We will assume that it is absolutely continuous with respect to the Lebesgue measure with density ρ_X: $F_X(dx) = \rho(x) \, dx$. We will call the density $\rho(x)$ the **probability density function** (PDF) of the random variable X.

Example

1. The exponential random variable has PDF

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0, \\ 0 & x < 0, \end{cases}$$

with $\lambda > 0$.

2. The uniform random variable has PDF

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b, \\ 0 & x \notin (a, b), \end{cases}$$

with $a < b$.

Definition

Two random variables X and Y are independent if the events $\{\omega \in \Omega \,|\, X(\omega) \leqslant x\}$ and $\{\omega \in \Omega \,|\, Y(\omega) \leqslant y\}$ are independent for all $x, \ y \in \mathbb{R}$.

Let $X, \ Y$ be two continuous random variables. We can view them as a random vector, i.e. a random variable from Ω to $\mathbb{R}^2$. We can then define the **joint distribution function**

$$F(x,y) = \mathbb{P}(X \leqslant x, \ Y \leqslant y).$$

The mixed derivative of the distribution function $f_{X,Y}(x,y) := \frac{\partial^2 F}{\partial x \partial y}(x,y)$, if it exists, is called the **joint PDF** of the random vector $\{X, \ Y\}$:

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(x,y) \, dxdy.$$

If the random variables X and Y are independent, then

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

and

$$f_{X,Y}(x,y) = f_X(x)f_Y(y).$$

The joint distribution function has the properties

$$\begin{aligned} F_{X,Y}(x,y) &= F_{Y,X}(y,x), \\ F_{X,Y}(+\infty,y) &= F_Y(y), \quad f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dx. \end{aligned}$$

We can extend the above definition to random vectors of arbitrary finite dimensions. Let X be a random variable from $(\Omega, \mathcal{F}, \mu)$ to $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The (joint) distribution function $F_X : \mathbb{R}^d \to [0,1]$ is defined as

$$F_X(\mathbf{x}) = \mathbb{P}(X \leqslant \mathbf{x}).$$

Let X be a random variable in $\mathbb{R}^d$ with distribution function $f(x_N)$ where $x_N = \{x_1, \dots x_N\}$. We define the **marginal** or **reduced distribution function** $f^{N-1}(x_{N-1})$ by

$$f^{N-1}(x_{N-1}) = \int_{\mathbb{R}} f^N(x_N)\, dx_N.$$

We can define other reduced distribution functions:

$$f^{N-2}(x_{N-2}) = \int_{\mathbb{R}} f^{N-1}(x_{N-1})\, dx_{N-1} = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x_N)\, dx_{N-1} dx_N.$$

We can use the distribution of a random variable to compute expectations and probabilities:

$$\mathbb{E}[f(X)] = \int_{\mathbb{R}} f(x)\, dF_X(x) \tag{4}$$

and

$$\mathbb{P}[X \in G] = \int_G dF_X(x), \quad G \in \mathcal{B}(E). \tag{5}$$

The above formulas apply to both discrete and continuous random variables, provided that we define the integrals in (4) and (5) appropriately.
When $E = \mathbb{R}^d$ and a PDF exists, $dF_X(x) = f_X(x)\, dx$, we have

$$F_X(x) := \mathbb{P}(X \leqslant x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f_X(x)\, dx..$$

Properties of the Expectation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let X, Y be random variables.

- $\mathbb{E}X$ exists if and only if $E|X| < +\infty$, i.e. $X \in L^1(\Omega)$. In this case $|\mathbb{E}X| \leqslant E|X|$.
- If $a, \, b \in \mathbb{R}$ then $E(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y$.
- If $X \leqslant Y$ then $EX \leqslant EY$.
- If $g(x)$ is a convex function[1] then $g(\mathbb{E}X) \leqslant \mathbb{E}g(X)$ (**Jensen's inequality**).
- If $X \geqslant 0$, $a > 0$ then $P(X > a) \leqslant E(X)/a$ (**Markov's inequality**).
- If X, Y are multidimensional $L^2(\Omega)$ random variables, then the **covariance** of X and Y is $\mathrm{Cov}(X, Y) = \mathbb{E}\Big(\big[X - \mathbb{E}(X)\big] \big[Y - \mathbb{E}(Y)\big]^T \Big)$.
- Let $X \in L^2(\Omega)$ and nonnegative and let $\mu = \mathbb{E}X$ and $\sigma^2 = \mathbb{E}(X - \mu)^2$. Then $\mathbb{P}((X - \mu)^2 \geqslant a) \leqslant \sigma^2/a$ for $a > 0$ (**Chebyshev's inequality**).

[1] $g(\lambda x + (1-\lambda)y) \leqslant \lambda g(x) + (1-\lambda) g(y)$ for $\lambda \in (0,1)$.

Example (Normal Random Variables)

- Consider the random variable $X : \Omega \mapsto \mathbb{R}$ with pdf

$$\gamma_{\sigma,m}(x) := (2\pi\sigma)^{-\frac{1}{2}} \exp\left(-\frac{(x-m)^2}{2\sigma}\right).$$

Such an X is termed a **Gaussian** or **normal** random variable. The mean is

$$\mathbb{E}X = \int_{\mathbb{R}} x\gamma_{\sigma,m}(x)\,dx = m$$

and the variance is

$$\mathbb{E}(X-m)^2 = \int_{\mathbb{R}} (x-m)^2\gamma_{\sigma,m}(x)\,dx = \sigma.$$

Example (Normal Random Variables contd.)

- Let $m \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ be symmetric and positive definite. The random variable $X : \Omega \mapsto \mathbb{R}^d$ with pdf

$$\gamma_{\Sigma,m}(x) := \left((2\pi)^d \mathrm{det}\Sigma\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\langle \Sigma^{-1}(x-m), (x-m)\rangle\right)$$

  is termed a **multivariate Gaussian** or **normal** random variable. The mean is

$$\mathbb{E}(X) = m \tag{6}$$

  and the covariance matrix is

$$\mathbb{E}\Big((X-m)\otimes(X-m)\Big) = \Sigma. \tag{7}$$

Let X, Y be random variables we want to know whether they are correlated and, if they are, to calculate how correlated they are. We define the covariance of the two random variables as

$$\text{cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\big] = \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y.$$

The **correlation coefficient is**

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(X)}} \tag{8}$$

The Cauchy-Schwarz inequality yields that $\rho(X, Y) \in [-1, 1]$. We will say that two random variables X and Y are **uncorrelated** provided that $\rho(X, Y) = 0$. It is not true in general that two uncorrelated random variables are independent. This is true, however, for Gaussian random variables.

The **characteristic function** of a random variable X is

$$\phi(t) = \int_{\mathbb{R}} e^{it\lambda}\, dF(\lambda) = \mathbb{E}(e^{itX}). \tag{9}$$

For a continuous random variable for which the distribution function F has a density, $dF(\lambda) = p(\lambda) d\lambda$, (9) gives

$$\phi(t) = \int_{\mathbb{R}} e^{it\lambda} p(\lambda)\, d\lambda.$$

For a discrete random variable for which $\mathbb{P}(X = \lambda_k) = \alpha_k$, (9) gives

$$\phi(t) = \sum_{k=0}^{\infty} e^{it\lambda_k} a_k.$$

The characteristic function determines uniquely the distribution function of the random variable, in the sense that there is a one-to-one correspondance between $F(\lambda)$ and $\phi(t)$.

Lemma

Let $\{X_1, X_2, \dots X_n\}$ be independent random variables with characteristic functions $\phi_j(t)$, $j = 1, \dots n$ and let $Y = \sum_{j=1}^n X_j$ with characteristic function $\phi_Y(t)$. Then

$$\phi_Y(t) = \Pi_{j=1}^n \phi_j(t).$$

Lemma

Let X be a random variable with characteristic function $\phi(t)$ and assume that it has finite moments. Then

$$E(X^k) = \frac{1}{i^k}\phi^{(k)}(0).$$

Theorem

Let $\mathbf{b} \in \mathbb{R}^n$ *and* $\Sigma \in \mathbb{R}^{n\times n}$ *a symmetric and positive definite matrix. Let* $\mathbf{X}$ *be the multivariate Gaussian random variable with probability density function*

$$\gamma(\mathbf{x}) = \frac{1}{Z}\exp\left(-\frac{1}{2}\langle \Sigma^{-1}(\mathbf{x}-\mathbf{b}), \mathbf{x}-\mathbf{b}\rangle\right).$$

Then

1. *The normalization constant is* $Z = (2\pi)^{n/2}\sqrt{\det(\Sigma)}$.
2. *The mean vector and covariance matrix of* $\mathbf{X}$ *are given by* $\mathbb{E}\mathbf{X} = \mathbf{b}$ *and* $\mathbb{E}((\mathbf{X}-\mathbb{E}\mathbf{X})\otimes(\mathbf{X}-\mathbb{E}\mathbf{X})) = \Sigma$.
3. *The characteristic function of* $\mathbf{X}$ *is*

$$\phi(\mathbf{t}) = e^{i\langle \mathbf{b},\mathbf{t}\rangle - \frac{1}{2}\langle \mathbf{t},\Sigma\mathbf{t}\rangle}.$$

3. INDEPENDENCE AND CONDITIONAL EXPECTATIONS

- A family $\{A_i \,:\, i \in I\}$ of events is called independent if

$$\mathbb{P}\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} \mathbb{P}(A_j) \tag{10}$$

  for all finite subsets J of I.
- The sub-σ-algebras $\mathcal{F}_1, \dots \mathcal{F}_n$ of $\mathcal{F}$ are said to be independent if Eqn. (10) holds for any choice of events $A_i \in \mathcal{F}_i$, $i = 1, \dots n$.
- The random variables $X_1, \dots X_n$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ are independent if the corresponding $\sigma-$algebras $\mathcal{F}(X_1), \dots \mathcal{F}(X_n)$ generated by the random vectors are independent.

- Let F and $F_1, \ldots, F_n$ denote the joint and marginal distributions of the random variables, respectively. Then $X_1, \ldots X_n$ are independent if and only if

$$F(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_i(x_i), \tag{11}$$

- and similarly for the probability density functions, if they exist.
- If $X_1, \ldots X_n$ are independent r.v. then

$$\mathbb{E}\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} \mathbb{E}X_i.$$

- Furthermore, if the X_i's are in $L^2(\Omega)$, then (with $V(X) = \text{Cov}(X, X)$),

$$V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i).$$

- When two events A, B are dependent it is important to know the probability that the event A will occur, given that B has already happened. We define this to be **conditional probability**, denoted by $\mathbb{P}(A|B)$. We know from elementary probability that

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

- More generally:

$$\begin{aligned}\mathbb{P}(A_1 \cap \cdots A_n) &= \mathbb{P}(A_1 \cap \cdots A_n \,|\, A_1 \cap \cdots A_{n-1}) \\ &\quad \times \mathbb{P}(A_1 \cap \cdots A_{n-1} \,|\, A_1 \cap \cdots A_{n-2}) \\ &\quad \cdots \times \mathbb{P}(A_1 \cap A_2 | A_1) \mathbb{P}(A_1).\end{aligned}$$

Definition

A family of events $\{B_i \, : \, i \in I\}$ is called a partition of Ω if

$$B_i \cap B_j = \emptyset, \quad i \neq j \quad \text{and} \quad \cup_{i \in I} B_i = \Omega.$$

Theorem

Law of total probability. *For any event A and any partition $\{B_i \, : \, i \in I\}$ we have*

$$\mathbb{P}(A) = \sum_{i \in I} \mathbb{P}(A | B_i) \mathbb{P}(B_i).$$

- Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and fix $B \in \mathcal{F}$. Then $\mathbb{P}(\cdot|B)$ defines a probability measure on $\mathcal{F}$:

$$\mathbb{P}(\emptyset|B) = 0, \quad \mathbb{P}(\Omega|B) = 1$$

- and (since $A_i \cap A_j = \emptyset$ implies that $(A_i \cap B) \cap (A_j \cap B) = \emptyset$)

$$P\left(\bigcup_{j=1}^{\infty} A_i | B\right) = \sum_{j=1}^{\infty} \mathbb{P}(A_i|B),$$

- for a countable family of pairwise disjoint sets $\{A_j\}_{j=1}^{+\infty}$. Consequently, $(\Omega, \mathcal{F}, \mathbb{P}(\cdot|B))$ is a probability space for every $B \in \mathcal{F}$.

Conditional Expectation

- Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{G}$ be a sub–σ–algebra of $\mathcal{F}$. The **conditional expectation** of X with respect to $\mathcal{G}$ is defined to be the random variable $\mathbb{E}[X|\mathcal{G}]$ on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

$$\int_G \mathbb{E}[X|\mathcal{G}] \, d\mu = \int_G X \, d\mu \quad \forall \, G \in \mathcal{G}.$$

- The existence of the conditional expectation is guaranteed by the Radon-Nikodym theorem. We can define $\mathbb{E}[f(X)|\mathcal{G}]$ and the conditional probability $\mathbb{P}[X \in F|\mathcal{G}] = \mathbb{E}[I_F(X)|\mathcal{G}]$, where I_F is the indicator function of F, in a similar manner.
- The conditional expectation has the same properties as the expectation (linearity, Jensen's inequality etc).

- If X is $\mathcal{G}-$measurable, then

$$\mathbb{E}(X|\mathcal{G}) = X.$$

- If X and $\mathcal{G}$ are independent, then

$$\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X).$$

- The conditional expectation has the same properties as the expectation (linearity, Jensen's inequality etc).
- For sub-σ-algebras $G_1 \subseteq G_2 \subseteq \mathcal{F}$,

$$\mathbb{E}(\mathbb{E}(X|G_1)|G_2)) = \mathbb{E}(\mathbb{E}(X|G_2)|G_1)) = \mathbb{E}(X|G_1).$$

- Furthermore:

$$\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}X.$$

- Assume that the r.v. X and Y have joint density $f(x,y)$ and let $f_2(y)$ denote the marginal density of Y

$$f_2(y) = \int f(x,y)\,dx.$$

- The conditional density $f(x|y)$ of X given $Y = y$ is

$$f(x|y) = \frac{f(x,y)}{f_2(y)}, \quad f_2(y) \neq 0,$$

- and satisfies

$$\mathbb{E}(h(X)|Y = y) = \int h(x) f(x|y)\,dx,$$

- for any integrable function h.

Example

Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\mathbf{X} = (X, Y)$ be a two-dimensional Gaussian random variable. Then, the conditional distribution of X given $Y = y$ is Gaussian with mean $\mu_X + \rho \frac{\sigma_X}{\sigma_Y}(y - \mu_Y)$ and variance $s_X^2(1 - \rho^2)$ where ρ denotes the correlation coefficient. Let $Z = \mathbb{E}(X \,|\, Y)$. Then

$$Z = \mu_X + \mu_X + \rho \frac{\sigma_X}{\sigma_Y}(Y - \mu_Y).$$

4. LIMIT THEOREMS OF PROBABILITY THEORY

- One of the most important aspects of the theory of random variables is the study of limit theorems for sums of random variables.
- The most well known limit theorems in probability theory are the **law of large numbers** and the **central limit theorem**.
- There are various different types of convergence for sequences or random variables.

Definition

Let $\{Z_n\}_{n=1}^{\infty}$ be a sequence of random variables. We will say that

(a) Z_n converges to Z with probability one if $\mathbb{P}\big(\lim_{n\to+\infty} Z_n = Z\big) = 1$.

(b) Z_n converges to Z in probability if for every $\varepsilon > 0$ $\lim_{n\to+\infty} \mathbb{P}\big(|Z_n - Z| > \varepsilon\big) = 0$.

(c) Z_n converges to Z in L^p if $\lim_{n\to+\infty} \mathbb{E}\big[\big|Z_n - Z\big|^p\big] = 0$.

(d) Let $F_n(\lambda), n = 1, \cdots + \infty$, $F(\lambda)$ be the distribution functions of Z_n $n = 1, \cdots + \infty$ and Z, respectively. Then Z_n converges to Z in distribution if $\lim_{n\to+\infty} F_n(\lambda) = F(\lambda)$ for all $\lambda \in \mathbb{R}$ at which F is continuous.

The distribution function F_X of a random variable from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $\mathbb{R}$ induces a probability measure on $\mathbb{R}$ and that $(\mathbb{R}, \mathcal{B}(\mathbb{R}), F_X)$ is a probability space. We can show that the convergence in distribution is equivalent to the weak convergence of the probability measures induced by the distribution functions.

Definition

Let (E, d) be a metric space, $\mathcal{B}(E)$ the $\sigma-$algebra of its Borel sets, P_n a sequence of probability measures on $(E, \mathcal{B}(E))$ and let $C_b(E)$ denote the space of bounded continuous functions on E. We will say that the sequence of P_n converges weakly to the probability measure P if, for each $f \in C_b(E)$,

$$\lim_{n \to +\infty} \int_E f(x)\, dP_n(x) = \int_E f(x)\, dP(x).$$

Theorem

Let $F_n(\lambda), n = 1, \cdots + \infty$, $F(\lambda)$ *be the distribution functions of* Z_n $n = 1, \cdots + \infty$ *and* Z, *respectively. Then* Z_n *converges to* Z *in distribution if and only if, for all* $g \in C_b(\mathbb{R})$

$$\lim_{n \to +\infty} \int_X g(x)\, dF_n(x) = \int_X g(x)\, dF(x). \tag{12}$$

- (12) is equivalent to

$$\mathbb{E}_n(g) = \mathbb{E}(g),$$

where E_n and E denote the expectations with respect to F_n and F, respectively.

- When the sequence of random variables whose convergence we are interested in takes values in $\mathbb{R}^n$ or, more generally, a metric space space (E, d) then we can use weak convergence of the sequence of probability measures induced by the sequence of random variables to define convergence in distribution.

Definition

A sequence of real valued random variables X_n defined on a probability spaces $(\Omega_n, \mathcal{F}_n, P_n)$ and taking values on a metric space (E, d) is said to converge in distribution if the indued measures $F_n(B) = P_n(X_n \in B)$ for $B \in \mathcal{B}(E)$ converge weakly to a probability measure P.

- Let $\{X_n\}_{n=1}^{\infty}$ be iid random variables with $\mathbb{E}X_n = V$. Then, the **strong law of large numbers** states that average of the sum of the iid converges to V with probability one:

$$\mathbb{P}\left(\lim_{n \to +\infty} \frac{1}{N} \sum_{n=1}^{N} X_n = V \right) = 1.$$

- The strong law of large numbers provides us with information about the behavior of a sum of random variables (or, a large number or repetitions of the same experiment) on average.
- We can also study fluctuations around the average behavior.

- let $\mathbb{E}(X_n - V)^2 = \sigma^2$. Define the centered iid random variables $Y_n = X_n - V$. Then, the sequence of random variables $\frac{1}{\sigma\sqrt{N}}\sum_{n=1}^{N} Y_n$ converges in distribution to a $\mathcal{N}(0,1)$ random variable:

$$\lim_{n \to +\infty} \mathbb{P}\left(\frac{1}{\sigma\sqrt{N}}\sum_{n=1}^{N} Y_n \leqslant a\right) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\, dx.$$

- This is the **central limit theorem.**