Chapter 2

Random Sequences and Stochastic Processes

In this chapter we will begin with a formal definition of what a stochastic process is and how it can be characterized. We will then study certain properties related to classes of processes which have simple probabilistic characterizations both in terms of their so-called sample-path properties as well as their probabilistic behavior. Throughout we will consider several 'canonical' examples which will aid us to better understand the concepts. We will conclude the chapter by seeing some of the most important results in probability, the so-called limit theorems which are the Laws of Large Numbers (LLN's both weak and strong) and the Central Limit Theorem. These results are powerful because of their generality and allow us to characterize the behavior of sequences and processes in large time.

2.1 Definitions and examples

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let T be an *index set*. For example T could be an arbitrary interval in $\Re$ or $T = [a, b] \times [c, d]$ a rectangle in $\Re^2$ or T could be a discrete set such as the set of non-negative integers $\{0, 1, 2, ...\}$. Then the indexed family of r.v's $\{X_t(\omega)\}_{t \in T}$ is said to be a stochastic process. This means that for each fixed $t \in T$, $X_t(\omega)$ defines a random variable. If T is an interval of $\Re$ then $X_t(\omega)$ is said to be a continuous one-parameter stochastic process. Usually we think of t as time and if T is continuous then the term continuous-time stochastic process is used. If t is a point inn a rectangle i.e. $t = (t, s) \in [a, b] \times [c, d]$ then $X_t(\omega)$ is a two-parameter stochastic process which is referred to as a *random field*. This can be extended to an arbitrary n-dimensional index set. If T is discrete we usually refer to the process as a discrete-parameter(or time) stochastic process. Discrete-time processes can be thought of as sequences of r.v's. In these notes we will restrict ourselves to the study of one-parameter stochastic processes in discrete as well as continuous time. Without loss of generality we take X to be $\Re$ or the process takes real values. The extension to vector-valued or $\Re^n$ valued processes is direct but it will simplify notation to consider the processes to be real valued.

As we have seen the mapping $X_t(\omega) : \Omega \to X$ where X denotes the space in which X_t takes its values is a random variable. On the other hand for every fixed $\omega \in \Omega$ the mapping $\{X_t(\omega)\}$ as a function of t is called a realization or sample-path of the process. It is a *deterministic* function of t. The problem is that we usually do not know ω and only have information on the underlying probabilities of joint events of the type $\{X_{t_i} \in A_i\}$ available. From this we need to get a good characterization of the process. This is what we discuss below. Also we drop the argument ω as before.

Let $T_n = (t_1, t_2, ..., t_n)$ be an arbitrary partition of T i.e. T_n denotes a finite collection of points in T. Then $\{X_{t_1}, ..., X_{t_n}\}$ forms a collection of r.v's which are characterized by their joint distribution

$$F_{T_n} = F_{t_1,..,t_n}(x_1, x_2, ..., x_n)$$

The family of joint distributions $\{F_{t_n}\}$ for all finite partitions T_n of T is called the finite dimensional distributions of $\{X_t\}$. The joint distributions F_{T_n} are said to be compatible if they satisfy the property of consistency (given in Chapter 1) and symmetry (i.e. if we take any permutation of $(t_1, ..., t_n)$ then the joint distribution remains the same). Now given a family of finite dimensional distributions there is the important theorem of Kolmogorov which states that these can be associated with the finite dimensional distributions of a stochastic process i.e. finite dimensional distributions for every finite partition are sufficient to characterize a stochastic process. This is stated below without proof.

Theorem 2.1.1 *(Kolmogorov extension theorem)*

Let $\{F_{T_n}\}$ be a compatible family of finite dimensional distributions for $T_n \in T$. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a stochastic process $\{X_t\}$ defined thereon such that the finite dimensional distributions of $\{X_t\}$ coincide with F_{T_n}.

Let us study the implications of the theorem by the aid of two canonical examples. It is important to note is that what the theorem states is the existence of *a* probability space and thus, given a metric space with finite-dimensional distributions defined thereon then a probability measure need not be defined on the space without further restrictions on the distributions.

Example 1: (Continuous time 'white noise') Take as the space Ω the space $C[0, T]$ i.e. the space of continuous functions. Consider the finite dimensional distributions defined by:

$$F_{T_n}(x_1, x_2, ..., x_n) = \frac{1}{(2\pi)^{\frac{n}{2}}} \prod_{i=1}^{n} \left(\int_{-\infty}^{x_i} e^{-\frac{y_i^2}{2}} dy_i \right)$$

i.e. the joint distribution of n i.i.d. N(0,1) random variables. Then it is easy to see that F_{T_n} forms a compatible family. Hence by the Kolmogorov extension theorem there exists a probability space $\Omega', \mathcal{F}, \mathbb{P})$ and a stochastic process $\{X_t\}$ defined there on with distributions F_{T_n}. Let us show that $\Omega' \neq \Omega = C[0, T]$ i.e. X_t cannot be continuous in t and possess such finite dimensional distributions. To show this we will show that if Ω is taken as $C[0, T]$ then the probability measure will fail to be sequentially continuous (re. Section 1, Chapter 1) and thus cannot be countably additive.

Consider the events $A_n = \{\omega : X_t > \epsilon, X_{t+\frac{1}{n}} \leq -\epsilon\}$. Then:

$$\mathbb{P}(A_n) = \frac{1}{2\pi} \left(\int_{\epsilon}^{\infty} e^{-\frac{x^2}{2}} dx \right)^2$$

Now since by assumption we take $X_t \in C[0, T]$ it implies that the event $\{X_t \neq X_{t+}\}$ is null and hence:

$$0 = \lim_{n \to \infty} \mathbb{P}(A_n) = \frac{1}{2\pi} \left(\int_{\epsilon}^{\infty} e^{-\frac{x^2}{2}} dx \right)^2 > 0$$

for all $\epsilon > 0$ which leads to a contradiction and violates the sequential continuity of $\mathbb{P}$. Hence $\Omega' \neq C[0, T]$.

A natural question is what is the appropriate choice of Ω'?. The answer is that Ω' can be taken to be the space $R^{[0,T]}$. It turns out that this space is *too large* to yield any 'nice' properties for X_t. But it is beyond the scope of the course.

Example 2: (Gaussian processes) The second example we consider is a discrete-time Gaussian process. The question is can we define a discrete-time Gaussian process on the space ℓ_2 which is the space of square summable sequences i.e. $\ell_2 = \{x_n : \sum_{n=1}^{\infty} |x_n|^2 < \infty\}$. Now if X_n is said to be a discrete-time Gaussian process if the finite-dimensional distributions are Gaussian . Let $C_n(h)$ denote the characteristic functional. For simplicity we consider the r.v's to be zero mean. Then:

$$C_n(h) = \mathbf{E}[e^{\imath \sum_{i=1}^{n} h_i X_i}] = e^{-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} r_{i,j} h_i h_j}$$

where $r_{i,j} = \mathbf{E}[X_i X_j]$ and $\{h_j\}$ are arbitrary scalars. Now suppose we consider the limit of $C_n(h)$ as $n \to \infty$. If it is a characteristic functional it must satisfy the properties outlined in Chapter 1. For this it is necessary and sufficient that $\sum_i \sum_j |r_{i,j}| < \infty$ since otherwise by suitable choice of $\{h_i\}$ the sum will be infinite and the limit of the characteristic functionals will be 0 and thus we cannot define a discrete-time Gaussian process in ℓ_2. Thus for example we cannot define a discrete-time white noise process i.e. with $r_{i,j} = \delta_{i,j}$ where δ is the Kronecker delta in ℓ_2. Discrete-time white noise is defined on $\Re^{\infty}$ or the space of real valued sequences and this space is quite well defined and can be considered as the canonical space on which all sequences of discrete-time stochastic processes are defined. This basically means that we have no problem of defining the probability space for discrete-time processes with given finite dimensional distributions unlike the continuous time case.

The aim of the examples has been to show that given a family of finite dimensional distributions then the construction of the probability space imposes some restrictions on either the sample-paths if we want to define Ω or if we fix Ω then it implies restrictions on the distributions. Henceforth we will always assume that the process $\{X_t\}$ is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and we will not usually explicitly specify the space except that such a space and process could be defined by the extension theorem.

2.2 Continuity of Stochastic Processes

Let $\{X_t\}_{t\in T}$ be a continuous time stochastic process. Unlike the case of deterministic functions of t there are different types of continuity w.r.t t that can be defined. This is due to the fact that in the probabilistic context a property holding almost surely is different in general than a property holding in the mean etc. This is also the case with continuity. We need these concepts in order to differentiate between processes that might be continuous at a given point vs that which is continuous over the entire interval and the fact that jump processes (i.e. processes which have discontinuities at certain points in time) satisfy weaker forms of continuity but are actually very different in behavior from those which which have no discontinuities.

Definition 2.2.1 *(Continuity in probability) A stochastic process $\{X_t\}$ is said to be continuous in probability at t if :*

$$\mathbb{P}(|X_{t+h} - X_t| \geq \varepsilon) \stackrel{h \to 0}{\to} 0 \ \forall \varepsilon > 0 \tag{2.2. 1}$$

Definition 2.2.2 *(Continuity in p-th mean) A stochastic process $\{X_t\}$ is said to be continuous inn the p-th mean at t if*

$$\mathbf{E}[|X_{t+h} - X_t|^p] \stackrel{h \to 0}{\to} 0 \tag{2.2. 2}$$

Remark: In light of Markov's inequality if a process is continuous in the p-th mean it is continuous in probability. From Lyapunov's inequality we obtain that if a process is continuous in the p-th mean then it is continuous in the r-th mean for all $r < p$.

Definition 2.2.3 *(Almost sure continuity) A stochastic process $\{X_t\}$ is said to be almost surely continuous at t if:*

$$\mathbb{P}(\omega : \lim_{h \to 0} |X_{t+h} - X_t| = 0) = 1 \tag{2.2. 3}$$

In each of the above definitions if the property holds for all $t \in T$ then we simply state that the respective continuity holds.

A final form of continuity which is the strongest is the notion of almost sure sample continuity. In order to define this we first need to define the notion of separability of a stochastic process.

Definition 2.2.4 *(Separability) A stochastic process $\{X_t\}$ is said to be separable if there exists a countable set $S \in T$ such that for every interval $I \in T$ and every closed set $K \in \Re$ the events :*

$$A = \{\omega : X_t \in K, t \in I \bigcap T\}$$

and

$$B = \{\omega : X_t \in K,\ t \in S \bigcap I\}$$

differ by a set Λ such that $\mathbb{P}(\Lambda) = 0$.

We state without proof that on a complete probability space we can always take $\{X_t\}$ to be separable. The importance of separability is the following: suppose we want to compute the probability of $\{\omega : \sup_t X_t \in A\} = \bigcap_{t \in T} \{\omega : X_t \in A\}$ then such an set may not be measurable i.e. the resulting set may not be an event since the intersection is over an uncountable intersection but if the process is separable then the intersection is taken over a countable set S and hence the resulting set is an event due to the σ-algebra property of $\mathcal{F}$ and hence a probability can be defined.

Definition 2.2.5 *(Almost sure sample continuity) A stochastic process $\{X_t\}$ which is separable is said to be almost surely sample continuous if :*

$$\mathbb{P}(\omega : \bigcup_{t \in T} \lim_{h \to 0} |X_{t+h} - X_t| \neq 0) = 0 \tag{2.2. 4}$$

Note almost sure continuity for every t does not imply almost sure sample continuity since if it holds for every t then $\bigcup_{t \in T}$ being an uncountable union need not be an event of probability 0 even if each is. It is difficult to characterize the event related to almost sure sample continuity in terms of finite dimensional distributions of a process. However a sufficient condition for almost sure continuity was given by Kolmogorov which we state without proof.

Proposition 2.2.1 *(Kolmogorov criterion) Let $\{X_t\}_{t \in T}$ be a separable process and T be finite. Then a sufficient condition for $\{X_t\}$ to be almost surely sample continuous is that there exist positive constants C, α, β such that :*

$$\mathbf{E}[|X_{t+h} - X_t|^\alpha] \leq C h^{1+\beta} \tag{2.2. 5}$$

Almost sure sample continuity implies almost sure continuity for every t. As mentioned we would like to differentiate between processes which possess almost sure sample continuity from processes which are continuous everywhere except on a set of points t which are are finite but have o measure with respect to the entire interval T i.e. Lebesgue measure 0. Such discontinuous processes are said to possess discontinuities of the first kind if for every t the limits X_{t+h} and X_{t-h} exist as $h \to 0$ but are not equal. An example is right continuous processes for which $X_{t-} \neq X_t$. A sufficient condition for this is due to Cramer which we also give without proof.

Proposition 2.2.2 *Let $\{X_t\}$ be a separable process. Then with probability 1 every sample path of $\{X_t\}$ has only discontinuities of the first kind if there exist positive constants C, α, β such that :*

$$\sup_{t \leq s \leq t+h} \mathbf{E}[|X_{t+h} - X_s|^{\alpha}|X_s - X_t|^{\alpha}] \leq Ch^{1+\beta} \tag{2.2. 6}$$

Remark: The Cramer's condition is weaker than Kolmogorov's criterion since by the Cauchy-Schwarz inequality we have that if Kolmogorov's condition is satisfied with C, α, β then Cramer's condition is satisfied with $C, \frac{\alpha}{2}, \beta$.

Let us now consider some examples.

Example 1: Let $\{X_t\}_{t\in[0,T]}$ be A Gaussian process with mean 0 and :

$$\mathbf{E}[|X_{t+h} - X_t|^2] = h$$

Then it is easy to see that it is continuous in probability, in the quadratic mean (mean of order 2).

Let us show that it is also almost surely sample continuous. For this note that since $X_{t+h} - X_t$ is Gaussian we have that

$$\mathbf{E}[|X_{t+h} - X_t|^4] = 3h^2$$

Therefor the process satisfies Kolmogorov's criterion with $C = 1$, $\alpha = 4$ and $\beta = 1$ and hence is almost surely sample continuous. We shall study this process further later on. It is called a Wiener process.

Example 2: Let $\{N_t\}_{t\in[0,\infty)}$ be a *point process* i.e. $N_0 = 0$ and N_t takes values in the set of non-negative integers $\{0, 1, 2, ...\}$ with the following property : there exist a sequence of random times $\{t_n\}$ with $t_1 < t_2 < t_3 < ...$ such that at times t_n $N_{t_n} = n$ and the process remains constant on the interval $[t_i, t_{i+1})$ and for any $s < t < u$ the random variables $\{N_u - N_t\}$ and $\{N_t - N_s\}$ are independent with $\mathbb{P}(N_{t+h} - N_t = k] = \frac{(\lambda h)^k}{k!}e^{-\lambda h}$. Then it can be seen that $\{N_t\}$ is continuous in probability and the mean. The process is not almost surely sample continuous since for any finite interval $\mathbb{P}(N_{t+T} - N_t > 0) = e^{-\lambda T} > 0$ and thus there exist points of discontinuity where $N_{t-} \neq N_t$. It however satisfies the Cramer's criterion with $\alpha = 1$ and $\beta = 1$ and $C = 1$ and thus has discontinuities of the first kind almost surely on any finite interval. The process we have defined is called a Poisson process which we shall study later on.

2.3 Classification of Stochastic Processes

In this section we will study some special classes of stochastic processes some of which we will study in detail. The idea is to study processes about which we can give general properties for members of the class for their distributions and moments. Unless explicitly stated we will usually consider continuous-time processes i.e. when the index set T is a subset of or equal to $\Re$.

One of the most commonly encountered stochastic processes is the so-called Gaussian process already briefly introduced in Section 2.1.

Definition 2.3.1 *(Gaussian processes) A stochastic process $\{X_t\}$ is said to be a Gaussian process if its finite dimensional distributions are Gaussian i.e. given any finite partition $T_n\{t_1, t_2, .., t_n\}$ of T then the random variables $\{X_{t_1}, X_{t_2}, ..., X_{t_n}\}$ are jointly Gaussian with characteristic functional:*

$$C(h) = \mathbf{E}[e^{\imath[h,X]}] = e^{\imath[h,m]-\frac{1}{2}[Rh,h]} \tag{2.3. 7}$$

where m is the vector of means i.e. $m = col(m_1, ..., m_n)$ with $m_i = \mathbf{E}[X_{t_i}]$ and R the covariance matrix with entries $R_{i,j} = cov(X_{t_i}, X_{t_j})$.

Remark: It is of course of interest to know on what space such processes are defined. Let $T = [0, T]$ and now consider a partition T_n of [0,T] with n going to infinity, $t_1 = 0,\ t_n = T$ and $\sup_j |t_j - t_{j-1}| \to 0$ then it can readily be seen that the above vectorial inner-products converge to integrals and the limit of the characteristic functional is just:

$$C(h) = e^{\imath \int_0^T h_s m_s ds - \frac{1}{2}\int_0^T \int_0^T R(t,s) h_s h_t dt ds}$$

where $m_t = \mathbf{E}[X_t]$ and $R(t, s) = cov(X_t, X_s)$. It can be shown but it is beyond the scope of the course that if we require the above characteristic functional to correspond to a characteristic functional of countably additive probability measure whose finite dimensional distributions are as above then we must have $\int_0^T \int_0^T |R(t, s)| dt ds < \infty$ (as in the example in Section 2.1). Thus with this condition we have that the Gaussian process is defined on $\Omega = L_2[0, T]$ or the space of square-integrable functions on $[0, T]$. If $T = [0, \infty)$ then the space is $L_2[0, \infty)$ which is a little less interesting since on such a space all functions must go to 0 as $t \to \infty$ for the integrals to exist. We will usually consider Gaussian processes on a finite time interval.

The next important class of stochastic processes is the class of stationary processes.

2.3.1 Stationary processes

Definition 2.3.2 *A stochastic process $\{X_t\}$ is said to be (strict-sense) stationary if for any finite partition $\{t_1, t_2, ..., t_n\}$ of T the joint distributions $F_{t_1,t_2,..,t_n}(x_1, x_2, .., x_n)$ are the same as the joint distributions $F_{t_1+\tau,t_2+\tau,..,t_n+\tau}(x_1, x_2, ..., x_n)$ for any τ.*

What the above definition states is that the joint distributions of the process or its shifted version are the same. In light of the definition since the shift τ is arbitrary the index set T is taken to be $(-\infty, \infty)$. We usually use the term stationary for strict-sense stationary. For stationary processes the following properties hold:

Proposition 2.3.1 *Let $\{X_t\}$ be a stationary process. Then:*

1. *$\mathbf{E}[X_t] = m$ i.e. its mean is a constant not depending on t.*
2. *Let $R(t, s) = cov(X_t, x_s)$ then $R(t, s)$ is of the form $R(t, s) = R(|t - s|)$ i.e. it only depends on the difference between t and s.*

Proof:

1)Let $F_t(x)$ denote the one-dimensional distribution of X_t. Then,

$$\begin{aligned} m_t = \mathbf{E}[X_t] &= \int x dF_t(x) = \int x dF_{t+\tau}(x) \\ &= m_{t+\tau} \end{aligned}$$

and since by stationarity it holds for all τ it implies that m_t cannot depend on t ot it is a constant.

2) Let us for convenience take m=0. Then:

$$\begin{aligned} R(t,s) &= \mathbf{E}[X_t X_s] = \int xy dF_{t,s}(x,y) \\ &= \int xy dF_{t-s,0}(x,y) = R(t-s,0) \\ &= \int xy dF_{0,s-t}(x,y) = R(0,s-t) \end{aligned}$$

Hence we obtain that

$$R(t,s) = R(t-s,0) = R(0,s-t)$$

or $R(t,s)$ must be a function of the magnitude of the difference between t and s.

The above requirement of strict sense stationarity is usually too strong for applications. In the context of signal processing a weaker form of stationarity just based on the mean and covariance is very useful. This is the notion of *wide sense stationarity* abbreviated as *w.s.s.*

Definition 2.3.3 *A stochastic process $\{X_t\}$ with $\mathbf{E}[|X_t|^2] < \infty$ is said to be wide sense stationary or w.s.s. if the properties 1. and 2. above hold i.e.*

1. $\mathbf{E}[X_t] = m$
2. $R(t,s) = R(|t-s|)$

In a later chapter we will study further properties of w.s.s. processes. It is important to note that w.s.s. does not imply strict sense stationarity. However, if X_t is a Gaussian process, since the finite dimensional distributions are completely specified by the mean and covariance, then the reverse implication holds.

We now introduce another class of stochastic processes of importance in applications.

2.3.2 Markov processes

Definition 2.3.4 *A stochastic process $\{X_t\}$ is said to be a Markov process (or simply Markov) if for any partition $\{t_1, t_2, ..., t_n\}$ of T with $t_1 < t_2 < \ldots < t_n$ the conditional distribution satisfies:*

$$\mathbb{P}\left(X_{t_n} \leq x_n / X_{t_{n-1}} = x_{n-1}, X_{t_{n-2}} = x_{n-2}, ..., X_{t_1} = x_1\right) = \mathbb{P}\left(X_{t_n} \leq x_n / X_{t_{n-1}} = x_{n-1}\right) \quad (2.3.\ 8)$$

For simplicity we will assume that conditional densities are defined and we will denote the conditional densities by $p(x_n, t_n / x_{n-1}, t_{n-1})$. An immediate consequence of the definition of a Markov process is the property of the conditional independence of the 'future' and 'past' given the present. This can often be taken as the definition of Markov processes and is readily generalizable to random fields or multi-parameter processes where there is no-natural definition of a causal or increasing flow of 'time'.

Proposition 2.3.2 *Let $\{X_t\}$ be a Markov process. Then for any $t_0 < t_1 < t_2$ we have:*

$$p(x_2, t_2; x_0, t_0/x_1, t_1) = p(x_0, t_0/x_1, t_1)p(x_2, t_2/x_1, t_1)$$

Proof: First note that the joint density of X_{t_0}, X_{t_1} and X_{t_2} can be written as:

$$p_{t_0,t_1,t_2}(x_0, x_1, x_2) = p(x_2, t_2/x_0, t_0; x_1, t_1)p(x_0, t_0/x_1, t_1)p_{t_1}(x_1)$$

Using the Markov property and the definition of the conditional density we have:

$$\frac{p_{t_0,t_1,t_2}(x_0, x_1, x_2)}{p_{t_1}(x_1)} = p(x_0, t_0; x_2, t_2/x_1, t_1) = p(x_2, t_2/x_1, t_1)p(x_0, t_0/x_1, t_1)$$

which establishes the conditional independence property of the 'future' and 'past' given the 'present'.

From the definition of the Markov property by repeated application it is easy to see that the joint density of $\{X_{t_0}, ..., X_{t_n}\}$ can be written as:

$$p(x_0, t_0; x_1, t_1; ...; x_n, t_n) = p(x_0, t_0) \prod_{k=1}^{n} p(x_k, t_k/x_{k-1}, t_{k-1})$$

and hence the joint distribution of a Markov process is completely specified by the initial distribution and the conditional distributions (1 step) which are obtainable from knowledge of the two-dimensional distributions of the process.

The above written in terms of the conditional distributions $F(x_n, t_n/x_{n-1}, t_{n-1}) = \mathbb{P}(X_{t_n} \leq x_n/X_{t_{n-1}} = x_{n-1})$ can be written as:

$$\begin{aligned} F(x_0, t_0; x_1, t_1; ...; x_n, t_n) &= \int_{-\infty}^{x_0} \int_{-\infty}^{x_1} .. \int_{-\infty}^{x_{n-1}} F(x_n, t_n/y_{n-1}, t_{n-1}) dF(y_{n-1}, t_{n-1}/y_{n-2}, t_{n-2}) \\ & \quad ....dF(y_1, t_1/y_0, t_0)dF(y_0, t_0) \end{aligned}$$

From the above observation that a Markov process is completely specified by its two-dimensional distributions we can pose the following question: given the two-dimensional distributions of a process what conditions must they satisfy in order that the process be Markov? The answer is that they must satisfy the following consistency properties.

1. $F(x, t) = \int_{-\infty}^{\infty} F(x, t/y, s)dF(y, s)$

2. For $t_0 < s < t$
$$F(x, t/x_0, t_0) = \int_{-\infty}^{\infty} F(x, t/y, s)dF(y, s/x_0, t_0)$$

Property 1) will be satisfied by any two-dimensional distribution by the definition of conditional distributions. Property 2 is the more important one which specifies the Markov property and is known as the Chapman-Kolmogorov equation.

Let us study the constraints imposed by the two properties through a commonly used example.

Suppose we want to define a Markov process $\{X_t\}$ which can take only two values $\{-1, 1\}$ with the following properties:

1. $\mathbb{P}(X_t = 1) = \mathbb{P}(X_t = -1) = \frac{1}{2}$ for all t.

2. For $t \geq s$

$$\begin{aligned}
\mathbb{P}(X_t = X_s) &= p(t-s) \\
\mathbb{P}(X_t = -X_s) &= 1 - p(t-s)
\end{aligned}$$

with $p(t)$ continuous and $p(0) = 1$.

Let us study the constraints it imposes on $p(.)$.
It is easy to see that Property 1 is always satisfied for all $p(t)$ since :

$$\begin{aligned}
\mathbb{P}(X_t = 1) = \frac{1}{2} &= \mathbb{P}(X_t = 1/X_s = 1)\mathbb{P}(X_s = 1) + \mathbb{P}(X_t = 1/X_s = -1)\mathbb{P}(X_s = -1) \\
&= \frac{1}{2}p(t-s) + \frac{1}{2}(1 - p(t-s)) = \frac{1}{2}
\end{aligned}$$

Property 2 or the Chapman-Kolmogorov equation specifies that $p(.)$ must satisfy:

$$p(t - t_0) = p(t-s)p(s-t_0) + (1 - p(t-s))(1 - p(s-t_0))$$

for all $t_0 < s < t$. Changing variables by setting $t_0 = 0$ and $t - s = \tau$ we obtain that $p(.)$ must satisfy:

$$p(s+\tau) = p(\tau)p(s) + (1 - p(\tau))(1 - p(s))$$

or

$$p(s+\tau) = 2p(\tau)p(s) - (p(\tau) + p(s))$$

Substituting $p(t) = \frac{1+f(t)}{2}$ we obtain that $f(t)$ must satisfy:

$$f(s+\tau) = f(s)f(\tau)$$

and the only non-trivial solution is that $f(t) = e^{-\lambda t}$ for $\lambda > 0$ and hence we obtain

$$p(t) = \frac{1 + e^{-\lambda t}}{2}$$

showing the constraints imposed by the Markov assumption.

An important sub-class of Markovian processes is the so-called processes with independent increments or simply independent increment processes.

Definition 2.3.5 *A stochastic process $\{X_t\}_{t \in [t_0, \infty)}$ is said to be an independent increment process if for any arbitrary collection of non-overlapping intervals $(s_i, t_i]; i = 1, 2, ..., n$ of $[t_0, \infty)$ the increments $\{X_{t_i} - X_{s_i}\}_{i=1}^{n}$ and X_{t_0} form a collection of independent random variables.*

Proposition 2.3.3 *Let $\{X_t\}_{t \in [0,\infty)}$ be an independent increment process. Then $\{X_t\}$ is a Markov process.*

Proof: Let $\{t_i\}$ be a partition of $[0, \infty)$ with $t_1 < t_2 \ldots < t_n$. Then to show that X_t is Markov it is sufficient to show that $\mathbb{P}(X_{t_n} \leq x / X_{t_1} = x_1, ...X_{t_{n-1}} = x_{n-1}) = \mathbb{P}(X_{t_n} \leq x / X_{t_{n-1}} = x_{n-1})$. First note that:

$$X_{t_n} = X_{t_n} - X_{t_{n-1}} + X_{t_{n-1}} = X_{t_{n-1}} + V_n$$

and by the independent increment property V_n is independent of $X_{t_i}; i = 1, 2, ..., n-1$. Therefore:

$$\begin{aligned} F(t_n, x_n / t_1, x_1; t_2, x_2, ..., t_{n-1}, x_{n-1}) &= F(V_n \leq x_n - x_{n-1} / t_1, x_1; t_2, x_2; .., t_{n-1}, x_{n-1}) \\ &= F_{V_n}(x_2 - x_1) \end{aligned}$$

by the independence of V_n and $\{X_{t_i}\}_{i=1}^{n-1}$. Hence since the distribution of X_{t_n} is completely determined from the distribution of V_n and the knowledge of $X_{t_{n-1}} = x_{n-1}$ it follows that the process is Markov.

A particular class of independent increment processes those with stationary independent increments forms an important class of processes which occur frequently in applications.

Definition 2.3.6 *A stochastic process $\{X_t\}_{t \in [0,\infty)}$ is said to be a process with stationary independent increments if the increments are independent and the distribution of $\{X_t - X_s\}$ for $t > s$ only depends on the difference $t - s$.*

The means and variances of stationary independent increment processes have a particularly simple form i.e. they are affine functions of t. We prove this result below.

Proposition 2.3.4 *Let $\{X_t\}_{t \in [0,\infty)}$ be a stationary independent increment process. Then :*

1. $\mathbf{E}[X_t] = m_0 + m_1 t$ *where* $m_0 = \mathbf{E}[X_0]$ *and* $m_1 = \mathbf{E}[X_1] - m_0$.
2. $var(X_t) = \sigma_0^2 + \sigma_1^2 t$ *where* $\sigma_0^2 = var(X_0)$ *and* $\sigma_1^2 = var(X_1) - \sigma_0^2$.

Proof: We will show 2) since the stationary independent increment property is used. The proof of 1) follows in a similar way except that only the stationary increment property is needed. To simplify the proof without loss of generality we take the mean to be 0. Note by the independent increment property $\mathbf{E}[X_t X_s] = \mathbf{E}[X_s^2]$ since

$$\begin{aligned} \mathbf{E}[X_t X_s] &= \mathbf{E}[(X_t - X_s + X_s) X_s] \\ &= \mathbf{E}[(X_t - X_s) X_s] + \mathbf{E}[X_s^2] \\ &= \mathbf{E}[X_t - X_s] \mathbf{E}[X_s] + \mathbf{E}[X_s^2] \\ &= \mathbf{E}[X_s^2] \end{aligned}$$

since the process is assumed to be zero mean. If the means are non-zero then it can be readily seen $cov(X_t X_s) = var(X_s)$.

Define: $g(t) = var(X_{t+u} - X_u)$ which does not depend on u by the stationary increment property. Then:

$$\begin{aligned} g(t+s) = var(X_{t+s} - X_0) &= var(X_{t+s} - X_s + X_s - X_0) \\ &= var(X_{t+s} - X_s) + var(X_s - X_0) \\ &= g(t) + g(s) \end{aligned}$$

Noting that
$$\frac{\partial g(t+s)}{\partial s} = g'(s) = \frac{\partial g(t+s)}{\partial t} = g'(t)$$
we obtain that $g'(t) = g'(s)$ for all s and t or $g'(t) = K$ for some constant K. Hence :
$$g(t) = Kt + K_1$$
is the general solution. Noting that from the relation above taking $t = s = 0$ we obtain $g(0) = 2g(0)$ it implies that $K_1 = 0$. Also since $g(1) = K$ we obtain that $K = var(X_1 - X_0)$. Finally noting that
$$var(X_t) = var(X_t - X_0 + X_0) = g(t) + var(X_0)$$
where we have used the independent increment property once again. The proof is then completed by noting that $var(X_1 - X_0) = \mathbf{E}[X_1^2] - 2\mathbf{E}[X_1X_0] + \mathbf{E}[X_0^2] = \mathbf{E}[X_1^2] - \mathbf{E}[X_0^2] = var(X_1) - var(X_0)$.

Remark: From the above it is evident that independent increment processes are defined on $[0, \infty)$ since if $T = (-\infty, \infty)$ then it is easy to see that such a process for any finite t is infinite and not well defined.

In signal processing (especially linear estimation theory) it is enough to require that the increments are uncorrelated and form a w.s.s. process. We give the definitions below:

Definition 2.3.7 *1. A stochastic process $\{X_t\}$ is said to be an uncorrelated increment process if the increments are uncorrelated i.e. for any $t_1 < t_2 < t_3$*
$$\mathbf{E}[(X_{t_3} - X_{t_2})(X_{t_2} - X_{t_1})] = \mathbf{E}[(X_{t_3} - X_{t_2})]\mathbf{E}[(X_{t_2} - X_{t_1}]$$

2. A stochastic process $\{X_t\}$ is said to be an orthogonal increment process if $\mathbf{E}[(X_{t_3} - X_{t_2})(X_{t_2} - X_{t_1})] = 0$.

Remark: If $\{X_t\}$ is zero mean and uncorrelated increment then it is an orthogonal increment process. If the increments in addition to being orthogonal are w.s.s. then we have a w.s.s orthogonal increment process. Such processes occur in the study of spectral theory and will be discussed in Chapter 4. Note if the process is Gaussian then the process will have stationary independent increments.

Let us now discuss some concrete examples of Markov and independent increment processes.
Example 1: Consider the following discrete-time process $\{X_n\}$ defined by:
$$X_{n+1} = f_n(X_n, W_n)$$
where $\{W_n\}$ is an independent sequence of r.v's and $f_n(.,.)$ is a deterministic function. Then $\{X_n\}$ is a Markov process. The proof is trivial.

Example 2: Let $\{Y_n\}$ be a sequence of independent identically distributed r.v's. Then:
$$X_n = \sum_{k=1}^{n} Y_k$$
is a stationary independent increment process. If the $\{Y_n\}$ are only an independent sequence but not identically distributed then $\{X_n\}$ will be an independent increment process. When the Y_n's are

i.i.d r.v's which take integer values in the set $\{-1, 1\}$, X_n is called a random walk. This model is a very important model in applications.

The next two examples we will discuss in detail since such processes are very basic processes in the study of stochastic processes.

Definition 2.3.8 *(Brownian motion or Wiener process)*
A continuous-time stochastic process $\{W_t; t \geq 0\}$ is said to be a Brownian motion or Wiener process with parameter σ^2 if:

1. $W_0 = 0$
2. $\{W_t\}$ *is a Gaussian process with mean 0 and* $\mathbf{E}[W_t W_s] = \sigma^2 \min(t, s)$.

If $\sigma^2 = 1$ then the process is said to be standard Brownian motion.

From the definition of Brownian motion the following properties can easily be shown.

Proposition 2.3.5 *Let $\{W_t\}_{t \in T}$ be a Brownian motion process. Then:*

1. *Every sample path of $\{W_t\}$ is almost surely sample continuous in t.*
2. *$\{W_t\}$ is a stationary independent increment process.*

Proof: For convenience we will take $\{W_t\}$ as a standard Brownian motion.
The proof of 1) follows from the fact that since the process is Gaussian $X_{t+h} - X_t$ is Gaussian and hence :

$$\mathbf{E}[|X_{t+h} - X_t|^4] = 3h^2$$

and hence Kolmogorov's criterion for almost sure sample continuity is satisfied with $\alpha = 4$,$\beta = 1$ and $c = 3$.
To prove the stationary independent increment property it is sufficient to that the process has orthogonal increments and then since the process is Gaussian it implies that the increments are independent. Stationary increment property follows from the fact that $\mathbf{E}[(W_t - W_s)^2] = (t - s)$. The proof of the orthogonal increment property follows by noting that for $t > s$:

$$\begin{aligned} \mathbf{E}[(W_t - W_s)(W_s)] &= \mathbf{E}[W_t W_s] - \mathbf{E}[W_s^2] \\ &= s - s = 0 \end{aligned}$$

showing that the increments are orthogonal.

The reason that the Wiener process or Brownian motion is so special is that one can show (which we will do later) that any Gauss-Markov process can be considered as a 'time-changed' Brownian motion. Gauss-Markov processes arise in the modeling of linear stochastic systems with noise and will be discussed in the next section. Actually there is an important result due to Paul Levy which states that if a process is Gaussian with stationary independent increments then it must be a Brownian motion process.
The analog of the Brownian motion process for processes with discontinuous trajectories is the Poisson process. This process plays a fundamental role in the study of processes with discontinuous trajectories the so-called jump processes. We define this below.

Definition 2.3.9 *(Poisson Process) A stochastic process $\{N_t\}_{t\in[0,\infty)}$ is said to be a Poisson process with intensity λ if N_t takes values in $\{0,1,2,\ldots\}$ and*

1. *$N_0 = 0$*
2. *For all $s,t \in [0,\infty)$ with $t > s$ then $N_t - N_s$ is a Poisson r.v. with mean $\lambda(t-s)$ i.e.*

$$\mathbb{P}(N_t - N_s = k) = \frac{(\lambda(t-s))^k}{k!}e^{-\lambda(t-s)}$$

3. *$\{N_t\}$ is an independent increment process.*

In light of property 2 above it implies that $\{N_t\}$ is a stationary, independent increment process with purely discontinuous sample-paths. Let us study some properties associated with Poisson processes.

First, note by virtue of the definition $\mathbf{E}[N_t] = \lambda t$ and $cov(N_t, N_s) = \lambda(min(t,s))$. Let $\{T_i\}_{i=1}^{\infty}$ denote the jump times of N_t with $T_0 = 0$ i.e.

$$N_t = n \qquad t \in [T_n, T_{n+1})$$

Then by definition of $\{T_n\}$ we can define N_t as:

$$N_t = \sum_{n=1}^{\infty} \mathbf{1}_{(T_n \le t)}$$

in other words we can define the Poisson process through its jump times. Such a process is referred to as a point process since the occurrence of the 'points' $\{T_n\}$ at which the process jumps by 1 allow us to completely specify the process. Let us now study some properties associated with the jump times.

First note that by definition $\{T_n \le t\} \equiv \{N_t \ge n\}$. Let $F_{T_n}(t)$ denote the distribution of T_n. Then for $n \ge 1$:

$$F_{T_n}(t) = 1 - \sum_{k=0}^{n-1} e^{-\lambda t}\frac{(\lambda t)^k}{k!}$$

and the density

$$\begin{aligned} p_{T_n}(t) &= \lambda e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}; \quad n \ge 1, \quad t \ge 0 \\ &= 0 \ \ \text{otherwise} \end{aligned}$$

In particular for $n = 1$ we obtain that $p_{T_1}(t) = \lambda e^{-\lambda t};\ \ t \ge 0$ or T_1 is exponentially distributed. From the stationary independent increment property it implies that $\{T_{n+1} - T_n\}$ have the same distribution as T_1 and thus the interarrival times $S_n = T_n - T_{n-1}$ of a Poisson process form a sequence of i.i.d. exponential mean $\frac{1}{\lambda}$ r.v's.

It can be readily seen that $\{N_t\}$ is not almost surely sample continuous although it can be shown that it is continuous in the mean square and almost surely for every t. To see this note that for any finite interval T we have:

$$\mathbb{P}(N_{t+T} \ne N_t) = \mathbb{P}(N_{t+T} - N_t > 0) = e^{-\lambda T} > 0$$

which shows that there are finite number of discontinuities in any finite interval of time with non-zero probability showing that the process is not almost-surely sample continuous. It can readily be verified that $\{N_t\}$ satisfies Cramer's criterion.

Analogous to the result of Levy for Wiener processes there is the result of Watanabe which states that if a point process has stationary independent increments then it must be Poisson.

2.4 Gauss-Markov Processes

As we have seen that by definition a Wiener process is both Gaussian and Markov. However the stationary independent increment property is special to the Wiener process. Gauss-Markov processes are very basic processes which arise in the modeling of noisy linear stochastic systems. Consider for example the following process defined by:

$$X_{k+1} = A_k X_k + F_k w_k \tag{2.4. 9}$$

where $\{w_k\}$ is a i.i.d. $N(0, I_m)$ sequence (i.e. discrete-time $\Re^m$-valued Gaussian white noise) and $X_0 \sim N(m_0, \Sigma_0)$ and A_k is a $n \times n$ matrix and F_k is a $n \times m$ matrix. It is easily seen that $\{X_k\}$ is a Gauss-Markov sequence. Such models arise in signal processing and control applications.

It turns out that the covariance of Gauss-Markov processes has a very particular structure which is both necessary and sufficient for a Gaussian process to be Markov. Let us study this issue in some detail.

Let $\{X_t\}$ be a Gauss-Markov process. Without loss of generality we take the process to be 0 mean. Let $R(t,s) = cov(X_t, X_s)$. Then we can state the following result.

Proposition 2.4.1 *Let $\{X_t\}$ be a Gaussian process with $R(t,t) > 0$ for all t. Then for $\{X_t\}$ to be Markov it is necessary that for all $t > s > t_0$*

$$R(t, t_0) = \frac{R(t,s)R(s,t_0)}{R(s,s)} \tag{2.4. 10}$$

Proof: First note that by the Gaussian property we have:

$$\mathbf{E}[X_t/X_s] = \frac{R(t,s)}{R(s,s)} X_s$$

Using the Markov property we have:

$$\begin{aligned}
\mathbf{E}[X_t/X_{t_0} = X_0] = \frac{R(t,t_0)}{R(t_0,t_0} X_0 \quad &= \int_{-\infty}^{\infty} x dF(t, x/t_0, X_0) \\
&= \int_{-\infty}^{\infty} \frac{R(t,s)}{R(s,s)} y dF(s, y/t_0, X_0) \\
&= \frac{R(t,s)}{R(s,s)} \frac{R(s,t_0)}{R(t_0,t_0)} X_0
\end{aligned}$$

where we have used the Chapman-Kolmogorov equation in the third step and the definition of the conditional mean of a Gaussian process. Hence comparing the l.h.s. and the r.h.s., since it is true for all X_0 we obtain:

$$R(t, t_0) = \frac{R(t,s)R(s,t_0)}{R(s,s)}$$

It can be shown that for the covariance to satisfy the above property the covariance $R(t,s)$ must be of the form $f(max(t,s))g(min(t,s))$. It turns out that this condition is also sufficient for a Gaussian process to be Markov. We show this below.

Proposition 2.4.2 *Let $\{X_t\}$ be a Gaussian process with $R(t,t) > 0$ for all t and $R(t,s)$ is continuous for $(s,t) \in T \times T$. Then for $\{X_t\}$ to be Markov it is necessary and sufficient that*

$$R(t,s) = f(max(t,s))g(min(t,s)) \tag{2.4. 11}$$

If in addition the process is stationary then

$$R(t) = R(0)e^{-\lambda|t-s|}; \quad \lambda \geq 0 \tag{2.4. 12}$$

and in particular $\lambda = -log(\frac{R(1)}{R(0)})$.

Proof:
Define

$$\rho(t,s) = \frac{R(t,s)}{\sqrt{R(t,t)}\sqrt{R(s,s)}}$$

Then $\rho(t,s)$ satisfies:

$$\rho(t,t_0) = \rho(t,s)\rho(s,t_0)$$

for all $t \geq s \geq t_0$. Now by the continuity of $\rho(t,s)$ and the assumption that $R(t,t) > 0$ for all $t \in T$ we see that $\rho(t,t) = 1$ and these two facts imply that $\rho(t,s) \neq 0$ for all $t, s \in int\ T$. Then since $\rho(t,s) = \frac{\rho(t,t_0)}{\rho(s,t_0)}$ for all $t > s > t_0$ we have that :

$$\rho(t,s) = \frac{\alpha(t)}{\alpha(s)}$$

for some $\alpha(t)$. If on the other hand $t < s$ then we have

$$\rho(t,s) = \frac{\alpha(s)}{\alpha(t)}$$

Hence,

$$\rho(t,s) = \frac{\alpha(max(t,s))}{\alpha(min(t,s))}$$

Hence it implies that for $t > s$ we have: $f(t) = \alpha(t)\sqrt{R(t,t)}$ and $g(s) = \frac{1}{\alpha(s)}\sqrt{R(s,s)}$ and this completes the proof of the necessity.

The proof of the sufficiency is based on the time-change property of Brownian motion. Consider the case $t > s$. Define the following time

$$\tau(t) = \frac{g(t)}{f(t)}$$

where for $t > s$ we have $R(t,s) = f(t)g(s)$. Then by the Cauchy-Schwarz inequality we obtain that :

$$R(t,s) \leq \sqrt{R(t,t)}\sqrt{R(s,s)}$$

Hence,

$$f(t)g(s) \leq \sqrt{f(t)g(t)f(s)g(s)}$$

or

$$\tau(s) \leq \tau(t)$$

implying that $\tau(t)$ is monotone non-decreasing in t. Define the process:

$$Y_t = f(t) W_{\tau(t)}$$

then Y_t is a Gauss-Markov process since $W_.$ is Gauss-Markov and $\tau(t)$ is an increasing function of t. Hence for $t > s$

$$\mathbf{E}[Y_t Y_s] = f(t) f(s) \tau(s) = f(t) g(s)$$

implying $\{Y_t\}$ is a Gauss-Markov process with the given covariance function.

Finally if the process $\{X_t\}$ is stationary then $R(t,s) = R(t-s)$ and hence:

$$\rho(t+s) = \rho(t)\rho(s)$$

for $\rho(.)$ defined above. The only non-trivial solution to this equation is that :

$$\rho(t) = ce^{\lambda t}$$

Now noting that $\rho(0) = 1$ implying $c = 1$ and $\rho(1) = \frac{R(1)}{R(0)} = e^{\lambda}$ we obtain that $\lambda = log(\frac{R(1)}{R(0)})$. Once again by Cauchy-Schwarz inequality noting that $R(\tau) \leq R(0)$ for all τ we note $\frac{R(1)}{R(0)} \leq 1$ implying $\lambda \leq 0$. This completes the proof by noting that $R(t) = R(0)\rho(t)$.

Similar results can be shown in the discrete-time case. We just state the results without proof.

Proposition 2.4.3 *Let $\{X_n\}$ be a stationary Gaussian sequence. Then in order that $\{X_n\}$ be Markov it is necessary and sufficient that the covariance satisfy:*

$$R_n = R_0 \rho^{|n|}; \quad 0 < \rho < 1 \tag{2.4. 13}$$

where $\rho = e^{-log(\frac{R_1}{R_0})}$.

2.5 Convergence of random variables

In applications we are often interested in the large-time behavior of a stochastic process. The study of these issues is fundamentally related to convergence of sequences of r.v's (in discrete-time) or convergence of the process (in continuous-time). As in the case of continuity there are various forms of convergence, some stronger than others, associated with processes. In this section we discuss the convergence concepts associated with sequences of r.v's and illustrate by example that they are not equivalent. For continuous-time processes similar definitions hold and we work with partitions and then show a 'uniformity' result in going to the limit as the partitions decrease to 0. First note that for each $\omega \in \Omega$ $\{X_n(\omega)\}$ defines a deterministic sequence. Hence, we could study the convergence of the sequence for each ω as in the case of real sequences. This is called the point-wise convergence property. This is however too strong and thus in the case of random sequences we would like to study the convergence taking account of the fact that we have a probability measure defined on Ω. This leads to the following notions:

Definition 2.5.1 *(Convergence in probability) A sequence of r.v's $\{X_n\}$ is said to converge to a r.v. X in probability if :*

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0$$

for any $\varepsilon > 0$.

As in the case of real sequences we do not often know a priori the limit X and so the convergence can be studied by the Cauchy criterion or mutual convergence in probability i.e.

Proposition 2.5.1 *A sequence $\{X_n\}$ converges in probability if and only if it converges mutually in probability or the sequence is Cauchy in probability i.e.*

$$\lim_{m\to\infty} \sup_{n\geq m} \mathbb{P}(|X_n - X_m| \geq \varepsilon) \to 0$$

for $\varepsilon > 0$.

Note mutual convergence can be equivalently stated as:

$$\lim_{n\to\infty} \sup_{m} \mathbb{P}(|X_{n+m} - X_n| \geq \varepsilon) \to 0$$

Definition 2.5.2 *(Convergence in th p.th. mean) A sequence of r.v's $\{X_n\}$ is said to converge in the p-th mean (also referred to as L^P convergence) to X if :*

$$\lim_{n\to\infty} \mathbf{E}[|X_n - X|^p] = 0$$

Of particular importance in applications is the convergence in the mean of order 2 called mean square convergence or convergence in the quadratic mean (q.m)

An immediate consequence of the Markov and Lyapunov inequalities is the following result.

Proposition 2.5.2 *Let $\{X_n\}$ be a sequence of r.v's.*

a) A necessary and sufficient condition for $\{X_n\}$ to converge in the p-th mean is that it be Cauchy in the p-th mean.

b) If $\{X_n\}$ converges in the p-th. mean then $\{X_n\}$ converges in the r-th mean for $1 \leq r \leq p$.

Definition 2.5.3 *(Almost sure convergence)*

A sequence $\{X_n\}$ is said to converge to X almost surely if

$$\mathbb{P}(\omega : \lim_{n\to\infty} |X_n - X| \neq 0) = 0$$

Once again if we do not know the limit a priori a necessary and sufficient condition for a.s. convergence is that the sequence be Cauchy a.s.

A final and weak form of convergence is the notion of convergence in distribution.

Definition 2.5.4 *(Convergence in distribution) A sequence $\{X_n\}$ converges to X in distribution if for every bounded continuous function $f(.)$ the following holds:*

$$\mathbf{E}[f(X_n)] \ \to \mathbf{E}[f(X)] \ \textit{as } n \to \infty$$

Convergence in distribution is equivalent to the convergence of the characteristic function. In particular the above condition implies that the distribution $F_n(x)$ of X_n converges to the distribution $F(x)$ of X for each x which is a point of continuity of $F(.)$.

In most applications we either require q.m. convergence or a.s. convergence. Before giving some examples we first state the following result which establishes the relationships between the different forms of convergence. We will use the notation $X_n \stackrel{a.s}{\to}$ to denote convergence a.s. and so on.

Proposition 2.5.3 *Let $\{X_n\}$ be a sequence of r.v's. Then the following relationships hold:*

1. $X_n \stackrel{a.s}{\to} X \Rightarrow X_n \stackrel{P}{\to} X$

2. $X_n \stackrel{L^P}{\to} X \Rightarrow X_n \stackrel{P}{\to} X$

3. $X_n \stackrel{P}{\to} X \Rightarrow X_n \stackrel{d}{\to} X$

4. $X_n \stackrel{d}{\to} C(a\ constant) \Rightarrow X_n \stackrel{P}{\to} C$

5. $X_n \stackrel{P}{\to} X \Rightarrow \exists\ a\ \text{subsequence} X_{n_k} \stackrel{a.s}{\to} X$

Proof: The implication 1 follows directly from the definitions while 2 follows from Lyapunov's inequality. We will prove 3 and 4 only. The proof of 5 is technical and we will omit it.

Proof of 3. Let $f(.)$ be bounded and continuous, let $|f(x)| \leq c$ and let $\varepsilon > 0$ and N be such that $\Pr(|X| > N) \leq \frac{\varepsilon}{4c}$. Choose δ such that $|f(x) - f(y) \leq \frac{\varepsilon}{2c}$ for $X| < N$ and $|x - y| < \delta$. Then:

$$\begin{aligned}\mathbf{E}[|f(X_n) - f(X)|] &= \mathbf{E}[|f(X_n) - f(X)|\mathbf{1}_{[|X_n - X| < \delta, |X| \leq N]}] \\ &\quad + \mathbf{E}[|f(X_n) - f(X)|\mathbf{1}_{[|X_n - X| < \delta, |X| > N]}] \\ &\quad + \mathbf{E}[|f(X_n) - f(X)|\mathbf{1}_{|X_n - X| \geq \delta]}] \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} + 2c\mathbb{P}(|X_n - X| \geq \delta)\end{aligned}$$

and the third term on the r.h.s. above goes to 0 as $n \to \infty$ by convergence in probability, hence for n sufficiently large we can make the r.h.s. smaller that 2ε establishing the convergence in distribution.

The proof of 4 follows from the fact that for n sufficiently large the probability distribution is concentrated around a ball of radius ε around C. Therefore for n sufficiently large :

$$\mathbb{P}(|X_n - C| \leq \varepsilon) \geq 1 - \delta$$

or:

$$\mathbb{P}(|X_n - X| \geq \varepsilon) \leq \delta$$

establishing the convergence in probability.

We now consider some examples to show that the various forms of convergence are not equivalent. In the next section we will study applications of convergence in terms of the so-called weak and strong laws of large numbers and ergodic theorems.

Example 1: (L^p convergence does not imply a.s. convergence) Let $\{X_n\}$ be a sequence of independent $\{0,1\}$ valued r.v's with $\mathbb{P}(X_n = 1) = \frac{1}{n}$ and $\mathbb{P}(X_n = 0) = 1 - \frac{1}{n}$.

Then $\{X_n\}$ converges to 0 in L^p for all $p > 0$ since : $\mathbf{E}[|X_n|^p] = \frac{1}{n}$. However, it does not converge $a.s$ since taking the events $A_n = \{X_n = 1\}$ gives:

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \sum_{n=1}^{\infty} p_n = \infty$$

(by Borel-Cantelli lemma).

Example 2: (a.s. convergence does not imply L^2 convergence). Let X be a r.v. uniformly distributed in [0,1]. Define:

$$\begin{aligned} Y_n &= n \quad if\ 0 \le Y \le \frac{1}{n} \\ &= 0 \quad otherwise \end{aligned}$$

Then clearly as $n \to \infty$ $Y_n \to 0$ a.s. but : $\mathbf{E}[Y_n^2] = n$ which goes to ∞.

Let us conclude with a result concerning a.s. convergence of sequences. This result is the key to establishing important ergodic theorems discussed in the next section.

Proposition 2.5.4 *Let $\{X_n\}$ be a sequence of r.v's with $\mathbf{E}[|X_n|^2] < \infty$. If $\sum_{n=1}^{\infty} \mathbf{E}[|X_n|^2] < \infty$ then $X_n \to 0$ almost surely as $n \to \infty$.*

Proof: Define the event :

$$A_n = \{\omega : |X_n| \ge \varepsilon\}$$

Then, by the Chebychev inequality:

$$\mathbb{P}(A_n) \le \frac{\mathbf{E}[|X_n|^2}{\varepsilon^2}$$

Therefore,

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) \le \frac{1}{\varepsilon^2} \sum_{n=1}^{\infty} \mathbf{E}[|X_n|^2] < \infty$$

Hence, by the Borel-Cantelli lemma, the probability that $\{A_n\ i.o.\}$ is 0 or $|X_n| \ge \varepsilon$ only for a finite number of n. Since $\varepsilon > 0$ is arbitrary it implies that $X_n \to 0$ a.s.

2.6 Laws of large numbers and the Central Limit Theorem

Strong and weak laws and their generalization, the so called ergodic theorem for stationary random processes are concerned with the problem of when we can infer the statistics of a process (or an appropriate function of it) by observing a single realization of the process. Typically the quantities we wish to estimate are means and variances. This is the basis for practical algorithms where we only observe a single realization or sample-path and we do not have the opportunity to observe numerous independent realizations (i.e. the basis of Monte-Carlo methods) to then calculate *ensemble averages*. The difference between weak and strong laws is that for the former the conclusion is in terms of a result holding in probability while the latter is an *almost sure* property. An intermediate result corresponds to the property holding in the mean squared sense.

Before we prove the results we first point out some facts about characteristic functions which will be used in some of the proofs.

Note by definition, the characteristic function of a r.v. is the Fourier transform of the density function if it exists. Hence, a natural question is can we know when a density exists if are given the characteristic function? The answer is, yes. It really is just a consequence of the Fourier inversion theorem and we state the result along with other results about characteristic functions in the following proposition.

Proposition 2.6.1 *(More about characteristic functions)*
Let $\phi(h)$ denote the characteristic function associated with a probability distribution $F(\cdot)$ defined on $\Re$ i.e.

$$\phi(h) = \int_{\Re} e^{\imath hx} dF(x)$$

Then the following results hold:

a) For any points $a, b \in \Re$ with $a < b$ at which $F(x)$ is continuous:

$$F(b) - F(a) = \lim_{c\to\infty} \frac{1}{2\pi} \int_{-c}^{c} \frac{e^{-\imath ta} - e^{-\imath tb}}{\imath t} \phi(t)dt$$

b) If $\int_{-\infty}^{\infty} |\phi(t)|dt < \infty$, the distribution function $F(x)$ possesses a density

$$F(x) = \int_{-\infty}^{x} p(y)dy$$

and

$$p(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\imath tx} \phi(t)dt$$

c) If $\phi^{(k)}(0)$ exists then:

$$\begin{aligned} \mathbf{E}|X|^k < \infty \qquad & if\ k\ is\ even \\ \mathbf{E}|X|^{k-1} < \infty \qquad & if\ k\ is\ odd \end{aligned}$$

d) If $X \sim F(.)$ and $\mathbf{E}[|X|^k] < \infty$ then

$$\phi(h) = \sum_{j=0}^{k} \frac{\mathbf{E}[X^j]}{j!} = o(t^k)$$

The proofs of these results follow from standard Fourier theory and the use of the Taylor expansion.

Proposition 2.6.2 *(Weak law of large numbers (WLLN)) Let $\{X_n\}$ be a sequence of i.i.d r.v.'s with $\mathbf{E}[|X_1|] < \infty$ and $\mathbf{E}[X_1] = m$. Then*

$$\frac{1}{n} \sum_{k=1}^{n} X_k \xrightarrow{P} m$$

Proof: Let $C(h) = \mathbf{E}[e^{\imath hX_1}]$ denote the characteristic function of X_i. Then denoting $S_n = \sum_1^n X_k$:

$$C_n(\frac{h}{n}) = \mathbf{E}[e^{\imath h\frac{S_n}{n}}] = \left[C(\frac{h}{n})\right]^n$$

But for each h:

$$C(h) = 1 + \imath hm + o(h) \quad h \to 0$$

Hence:

$$C_n(\frac{h}{n}) = \left(1 + \imath\frac{h}{n}m + o(\frac{1}{n})\right)^n$$

Hence as $n \to \infty$ we have :

$$\lim_{n\to\infty} C_n(\frac{h}{n}) \to e^{\imath hm}$$

which corresponds to the characteristic function of a probability distribution concentrated at m or $\frac{S_n}{n}$ converges in distribution to a constant m which in light of the previous section implies that $\frac{S_n}{n}$ converges to m in probability.

In the above result we assumed that the sequence was i.i.d. We can remove the i.i.d. restriction provided that we impose further conditions such as the existence of the second moments. In this case we do not even need stationarity of the sequence. We give the form of the WLLN below.

Proposition 2.6.3 *(WLLN for dependent 2nd. order sequences)*
Let $\{X_n\}$ be a sequence of r.v's with $\mathbf{E}[|X_n|^2] < \infty$ for all n. Let $m_k = \mathbf{E}[X_k]$ and $r(j,k) = cov(X_j, X_k)$. Suppose that $|r(j,k)| \leq Cg(|j-k|)$ with $g(k) \to 0$ as $k \to \infty$ then :

$$\frac{1}{n}\sum_{k=1}^{n}(X_k - m_k) \xrightarrow{P} 0$$

Proof: Define $S_n = \sum_{k=1}^{n}(X_k - m_k)$. Then using the Chebychev inequality:

$$\begin{aligned}
\mathbb{P}(|\frac{S_n}{n}| \geq \varepsilon) &\leq \frac{\sum_{j=1}^{n}\sum_{k=1}^{n}|r(j,k)|}{n^2\varepsilon^2} \\
&\leq C\frac{\sum_{|k|\leq n-1} g(k)(1-\frac{|k|}{n})}{n\varepsilon^2} \\
&\leq 2C\frac{\sum_{|k|\leq n-1} g(k)}{n\varepsilon^2}
\end{aligned}$$

Now since $g(k) \to 0$ as $k \to \infty$ for all $n \geq N(\delta\varepsilon^2 C^{-1})$ we have $g(n) \leq \delta\varepsilon^2 C^{-1}$ and hence for all $n \geq N(\delta\varepsilon^2)$:

$$\mathbb{P}(|\frac{S_n}{n}| \geq \varepsilon) \leq 2C\frac{\sum_{k=1}^{N(\delta\varepsilon^2)} g(k)}{n\varepsilon^2} + 2\delta$$

The first term on the r.h.s. can be made as small as possible since it is a finite sum divided by $n \to \infty$ hence for $n \geq N(\delta\varepsilon^2 C^{-1})$ large we can make:

$$\mathbb{P}(|\frac{S_n}{n}| \geq \varepsilon) \leq 2\delta$$

and δ can be chosen arbitrarily small and so we have established convergence in probability.

Let us note some implications of the WLLN. Consider a collection of i.i.d. r.v's $\{X_i\}_{i=1}^{n}$ with $\mathbf{E}[x_i] = m$ and $var(X_i) = \sigma^2$. Let us define: $M_n = \frac{S_n}{n}$. Then by the WLLN $M_n \to m$ and $var(M_n) \to 0$. However we obtain very little information about the behavior of S_n. For example, the weak law states that S_n is about nm where m is the mean when n is large. How good is this estimate on how large S_n is? We need much more precision. It turns out that we can say quite a lot about the limiting distribution of $S_n - nm$ for large n under the rather weak assumption that the X_n's have finite variance. This is the essence of the powerful Central Limit Theorem (CLT) which we state and prove below.

Proposition 2.6.4 *(Central Limit Theorem) Let $\{X_i\}_{i=1}^n$ be a sequence of i.i.d. r.v's with $\mathbf{E}[x_i^2] < \infty$. Define $S_n = \sum_{i=1}^n X_i$. Then as $n \to \infty$:*

$$\mathbb{P}\left(\frac{S_n - nm}{\sigma\sqrt{n}} \leq x\right) \to \Phi(x)$$

where $\mathbf{E}[X_1] = m$ and $var(X_1) = \sigma^2$ and $\Phi(x)$ denotes the standard normal distribution given by:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy$$

Proof: Let $\phi(t) = \mathbf{E}[e^{it(X_1 - m)}]$

Then:

$$\begin{aligned}\phi_n(t) &= \mathbf{E}[e^{it(\frac{S_n - nm)}{\sigma\sqrt{n}})}] \\ &= \left[\phi(\frac{t}{\sigma\sqrt{n}})\right]^n\end{aligned}$$

But from the expansion for $\phi(t)$ we have:

$$\phi(t) = 1 - \frac{\sigma^2 t^2}{2} + o(t) \ as \ t \to \ 0$$

Hence:

$$\phi_n(t) = [1 - \frac{\sigma^2 t^2}{2\sigma^2 n} + o(\frac{1}{n})]^n \to e^{-\frac{t^2}{2}}$$

as $n \to \infty$ for fixed t.

But the rhs is just the characteristic function of a $N(0,1)$ r.v. Hence the result is established since the above result implies convergence in distribution.

Remarks: The above version of the CLT assumes that the r.v's are i.i.d. There are much more general versions of the CLT which can be shown in the absence of the i.i.d. hypothesis. However, there are important observations to be made even in the i.i.d. case. First note that no matter what the distribution is or whether the r.v's are continuous or discrete, the limiting distribution of $\frac{S_n - nm}{\sigma\sqrt{n}}$ is Gaussian. The CLT thus provides the justification in stating that the scaled (by the factor $\sqrt{n}$) convolution of copies of a probability distribution (measure) can be approximated by a Gaussian distribution. However, is this approximation valid at the level of densities? It is well known that just because two functions are "close" their derivatives need not be and since the density is the derivative of the distribution function it should not necessarily follow that density of the scaled (by $\sqrt{N}$) sum should also be close to a Gaussian density.

We can indeed state such a result if we impose some more conditions on the random variables. This result is often referred to as a Local Limit Theorem which have been developed by Gnadenko, Korolyuk, Petrov, etc. We omit the proof. It relies more on Fourier analysis rather than probability.

Proposition 2.6.5 *(Local limit theorem)*

Let $\{X_i\}_{i=1}^n$ be a collection of i.i.d. mean 0 and variance 1 random variables. Suppose that their common characteristic function $\phi(.)$ satisfies:

$$\int_{-\infty}^{\infty} |\phi(t)|^r dt < \infty$$

for some integer $r \geq 3$. Then a density $p_n(x)$ exists for the normalized sum $\frac{S_n}{\sqrt{n}}$ for $n \geq r$ and furthermore:

$$p_n(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \left(1 + O(\frac{1}{n^{\frac{r-1}{2}}})\right)$$

Remark: Actually the error term can be made more precise given that moments of order $r \geq 2$ exist under the assumption on the characteristic function.

The next ergodic theorem concerns the ergodic property holding in the mean squared sense. We first state the result for w.s.s. processes.

Proposition 2.6.6 *a) Let $\{X_n\}$ be a w.s.s. sequence of r.v's with mean $\mathbf{E}[X_n] = m$ and covariance $R(k) = cov(X_{n+k}, X_n)$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} R(k) = 0 \Leftrightarrow \frac{\sum_{k=0}^{n-1} X_k}{n} \stackrel{L^2}{\to} m$$

b) If $\{X_t\}$ is q.m. continuous w.s.s. process with $\mathbf{E}[X_t] = m$ and covariance $R(t) = cov(X_{s+t}, X_s)$. Then

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T R(t)dt = 0 \Leftrightarrow \frac{1}{T} \int_0^T X_s ds \stackrel{L^2}{\to} m$$

Proof: We prove a). The proof of b) is analogous. Proof of sufficiency: Define $S_n = \sum_{k=0}^{n-1}(X_k - m)$. Then:

$$\begin{aligned} \mathbf{E}[|\frac{S_n}{n}|^2] &= \frac{1}{n^2} \sum_{k=0}^{n-1} \sum_{m=0}^{n-1} R(k-m) \\ &= \frac{1}{n} \sum_{|k| \leq n-1} R(k)(1 - \frac{|k|}{n}) \end{aligned}$$

Now from the non-negative definiteness of $R(k)$ (see Chapter 3) it implies that $\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} R(i-j) = \sum_{|k|\leq n-1} R(k)(n - |k|) \geq 0$ implying that $\sum_{|k|\leq n-1} nR(k) \geq \sum_{|k|\leq n-1} |k|R(k)$ we obtain

$$\frac{1}{n} \sum_{|k|\leq n-1} R(k)(1 - \frac{|k|}{n}) \stackrel{n\to\infty}{\to} 0$$

from which the sufficiency follows.

The proof of the necessity follows from the fact that by the w.s.s. property:

$$\begin{aligned} |\frac{1}{n} \sum_{k=0}^{n-1} R(k)| &= |\mathbf{E}[\frac{1}{n} \sum_{k=0}^{n-1} (X_k - m)(X_0 - m)]| \\ &\leq \sqrt{\mathbf{E}[(\frac{1}{n} \sum_{k=0}^{n} (X_k - m))^2]} \sqrt{\mathbf{E}[(X_0 - m)^2]} \\ &\stackrel{n\to\infty}{\to} 0 \end{aligned}$$

Remark: If we remove the hypothesis that the process is w.s.s. but only retain the fact that it is a second order process then a sufficient condition for $\frac{1}{n}\sum_{k=0}^{n-1}(X_k - m_k) \stackrel{L^2}{\to} 0$ is that

$\frac{1}{n^2}\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} R(i,j) \to 0$ as $n \to \infty$. For the analogous result in for q.m. continuous processes is that $\frac{1}{T^2}\int_0^T\int_0^T R(t,s)dtds \to 0$ as $T \to \infty$.

We now discuss the SLLN (Strong Law of Large Numbers) and its generalization the individual ergodic theorem for strictly stationary processes due to Birkhoff. We will omit the proofs since they are technical but just indicate the fact that the SLLN follows from the Borel-Cantelli lemma. As above we state the SLLN for the case of non-stationary second-order processes and sequences due to Cramer.

Proposition 2.6.7 *(Strong Law of Large Numbers (SLLN))*

a. *Let $\{X_n\}$ be a second order sequence with mean $\mathbf{E}[X_k] = m_k$ and covariance $R(j,k)$. Suppose $R(j,k)$ satisfies for $|j-k|$ large:*

$$|R(j,k)| \leq Cg(|j-k|)$$

such that

$$\frac{1}{n}\sum_{|k|\leq n-1} g(k)(1-\frac{|k|}{n}) \leq \frac{C}{n^\alpha} \quad ;\alpha > 0$$

Then,

$$\frac{1}{n}\sum_{k=0}^{n-1}(X_k - m_k) \stackrel{a.s}{\to} 0$$

b. *Let $\{X_t\}$ be a q.m. continuous process with $\mathbf{E}[X_t] = m_t$ and covariance $R(t,s)$. If for large $|t-s|$*

$$|R(t,s)| \leq Cg(|t-s|)$$

with

$$\frac{1}{T}\int_0^T g(t)(1-\frac{|t|}{T})dt \leq \frac{c}{T^\gamma} \quad ;\gamma > 0$$

Then:

$$\frac{1}{T}\int_0^T (X_s - m_s)ds \stackrel{a.s.}{\to} 0$$

Proof: We will prove this result under a more restrictive assumption, namely that $\{X_n\}$ is a w.s.s. sequence with $\sum_{k=0}^\infty |R(k)| < \infty$. This satisfies hypothesis a) with $\alpha = 1$ and also the hypothesis of the Proposition 2.6.6. Hence it follows that $\frac{S_n}{n} \to m$ in q.m. where $\mathbf{E}[X_n] = m$ and $S_n = \sum_{k=1}^n X_k$. Since it converges in L_2 it follows from Proposition 2.5.3 that $\exists$ a subsequence S_{n_i} such that $\frac{S_{n_i}}{n_i} \to m$ a.s.

Indeed let us choose $n_i = i^2$, then:

$$\begin{aligned}
\Pr\{|\frac{S_{i^2}}{i^2} - m| > \varepsilon\} &\leq \frac{\sum_1^{i^2}\sum_1^{i^2} R(|j-k|)}{i^4\varepsilon^2} \\
&\leq \frac{2}{i^2\varepsilon^2}\sum_{k=0}^{i^2-1}|R(k)| \leq \frac{2}{i^2\varepsilon^2}\sum_{k=1}^\infty |R(k)|
\end{aligned}$$

Therefore:

$$\sum_{i=1}^\infty \Pr\{|\frac{S_{i^2}}{i^2} - m| > \varepsilon\} \leq \frac{2\sum_{k=0}^\infty |R(k)|}{\varepsilon^2}\sum_{i=1}^\infty \frac{1}{i^2} < \infty$$

Hence, by the Borel-Cantelli lemma, $\frac{S_{i^2}}{i^2} \to m$ a.s.

Now, let us assume that $\{X_k\}$ is a non-negative sequence. This means that S_n increases with n. Hence, for $i^2 < n < (i+1)^2$ we have:

$$\frac{S_{i^2}}{(i+1)^2} \leq \frac{S_n}{n} \leq \frac{S_{(i+1)^2}}{i^2}$$

and noting that $\frac{(i+1)^2}{i^2} \to 1$ as $i \to \infty$ and the observation about $\frac{S_{i^2}}{i^2}$ above we have

$$m \leq \lim_{n\to\infty} \frac{S_n}{n} \leq m$$

and hence $\frac{S_n}{n} \to m$ a.s. as $n \to \infty$.

We extend the result to general X_n by noting that:

$$X_n = X_n^+ - X_n^-$$

where $X_n^= = \max(X_n, 0)$ and $X_n^- = -\min(X_n, 0)$ and so both $X_n^=$ and X_n^- are non-negative with $\mathbf{E}[X_n] = \mathbf{E}[X_n^+] - \mathbf{E}[X_n^-]$. By definition of X_n^+ and X_n^- we have $X_n^+ \leq |X_n|$ and hence $\mathbf{E}|X_n^+|^2 \leq \mathbf{E}|X_n|^2$. Similarly for X_n^-. Therefore $\mathbf{E}|X_n^=|^2 < \infty$ and $\mathbf{E}|X_n^-|^2 < \infty$ if $\mathbf{E}|X_n|^2 < \infty$. Applying the result for non-negative X_n to X_n^+ and X_n^- we have:

$$\frac{S_n}{n} = \frac{\sum_{k=1}^n X_k^+}{n} - \frac{\sum_{k=1}^n X_k^-}{n}$$

and the result follows.

Remark: Condition b) holds if

$$|R(t,s)| < \frac{K}{1 + |t-s|^\gamma}$$

for some constant K.

Let us conclude this section with a statement of the ergodic theorem for strictly stationary processes. We will only discuss the continuous time case. The discrete-time case follows mutatis mutandis.

Definition 2.6.1 *A set A is said to be invariant for a process $\{X_t\}$ if $\{X_s \in A; s \in \Re\}$ implies that $\{X_{t+s} \in A; s \in \Re\}$ for all $t \in \Re$. In other words the shifted version of the process remains in A.*

An example of an invariant set is :

$$A = \{X_. : \lim_{T\to\infty} \frac{1}{T}\int_0^T f(X_{s+t})dt = 0\}$$

Another example of a set which is not invariant the following:

$$A = \{X_. : X_s \in (a,b) \ for \ some \ s \in [0,T]\}$$

Then:

Definition 2.6.2 *A stochastic process $\{X_t\}$ is said to be ergodic if every invariant set A of realizations is such that $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$.*

Remark: An ergodic process need not be stationary. For example a deterministic process can be ergodic but is not stationary.

If the process is strictly stationary the following theorem can be stated.

Proposition 2.6.8 *Let $\{X_t\}$ be a strictly stationary process and f(.) be a measurable function such that $\mathbf{E}[|f(X_t)|] < \infty$. Then*

a) The following limit :

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T f(X_s)ds$$

exists almost surely and is a random variable whose mean is $\mathbf{E}[f(X_0)]$.

b) If in addition the process is ergodic then:

$$\frac{1}{T}\int_0^T f(X_s)ds \overset{a.s.}{\to} \mathbf{E}[f(X_0)]$$

Let us study the issue of ergodicity a bit further via an example. The ergodic theorems (WLLN, SLLN) state that time averages of the form either $\frac{S_n}{n}$ or $\frac{\int_0^T X_s ds}{T}$ converge to their means or *ensemble* averages $E[X_0]$ under appropriate conditions.

Ergodicity is much stronger since it states that $\lim_{N\to\infty}\sum_{k=0}^{N-1} f(X_k) \to \mathbf{E}[f(X_0)]$ for all "nice" functions. In other words, the SLLN applies to the sequence $Y_n = f(X_n)$. If we do not know that $\{X_k\}$ is ergodic then we would have to check that the SLLN applies to $\{Y_k\}$.

To do so let us consider the following example in continuous time for simplicity.

Consider the process:

$$X_t(\omega) = A(\omega)\cos(2\pi t + \theta(\omega))$$

where A and θ are independent r.v's with θ uniformly distributed in $[0, 2\pi]$.

Then one can show that $\{X_t\}$ is strictly stationary (see Problem 6 at the end of the chapter). A natural question is whether $\{X_t\}$ is ergodic.

First note that:

$$\begin{aligned}\mathbf{E}[X_t] &= \mathbf{E}[A]\left(\frac{1}{2\pi}\int_0^{2\pi}\cos(2\pi t + x)dx\right)\\ &= 0\\ &= \mathbf{E}[X_0]\end{aligned}$$

since the integral of a cosine function over a period is 0.

Also:

$$\begin{aligned}\lim_{T\to\infty}\frac{1}{T}\int_0^T X_t dt &= \lim_{T\to\infty}\frac{1}{T}\int_0^T A\cos(2\pi t + \theta)dt\\ &= 0\end{aligned}$$

Note, in the above $A(\omega)$ and $\theta(\omega)$ are just treated as some constants. Hence the SLLN holds for $\{X_t(\omega)\}$.

Now let us see if the result holds for any arbitrary but "nice" function. Let $f(.)$ be a continuous bounded function.

$$\begin{aligned}
\mathbf{E}[f(X_0)] &= \mathbf{E}[Af(\cos(\theta))] \\
&= \frac{1}{2\pi}\mathbf{E}_A[\int_0^{2\pi} f(A\cos x)dx \\
&= \mathbf{E}_A[\int_0^1 f(A\cos(2\pi t)dt]
\end{aligned}$$

where $\mathbf{E}_A$ denotes taking the average over the distribution of A since A and θ are independent.

Now let us compute $\frac{1}{T}\int_0^T f(X_t)dt$.

$$\begin{aligned}
\lim_{T\to\infty}\frac{1}{T}\int_0^T f(X_t)dt &= \lim_{N\to\infty}\frac{1}{N}\int_0^N f(A\cos(2\pi t+\theta))dt \\
&= \lim_{N\to\infty}\frac{1}{N}\sum_1^N\int_0^1 f(A\cos(2\pi t+\theta))dt \\
&= \lim_{N\to\infty}\frac{1}{N}\sum_1^N\int_\theta^{1+\theta} f(A\cos(2\pi s))ds \\
&= \lim_{N\to\infty}\frac{1}{N}\sum_1^N\int_0^1 f(A\cos(2\pi s))ds \\
&= \int_0^1 f(A\cos(2\pi s))ds
\end{aligned}$$

Comparing the two expressions we see that they are equal if and only if A is non-random. Thus, in general, the process is not ergodic.

Indeed it can be shown that the only invariant r.v. corresponding to X_t is a r.v. which depends on $A(\omega)$., and so if it is a constant, then X_t will be ergodic.

In general there are no easy conditions to know when a process is ergodic (except in the Gaussian case discussed in the next chapter) and thus proving an ergodicity result usually involves using the SLLN applied to the process $Y_t = f(X_t)$. The second important point to note that even if $\{X_t\}$ is stationary , $f(X_t)$ need not be stationary and thus the above theorem states that ergodicity is maintained under nonlinear transformation but not stationarity. In many engineering books ergodic processes are taken to be a synonymous of stationary processes but as the above results show they are not necessarily linked concepts.

2.7 Discussion : Classical limit theorems

Let us conclude with a discussion of the classical limit theorems due to de Moivre-Laplace and Poisson limit theorem which were the pre-cursors of the Law of Large Numbers and the CLT.

Let $\{X_i\}$ be a sequence of i.i.d. $\{0,1\}$ random variables where $\{X_i = 1\}$ is identified as a "success". Let $\Pr(X_i = 1) = p$, $\Pr(X_i = 0) = q = 1-p$ and $S_n = \sum_{i=1}^n X_i$. Then S_n counts the number of successes in n independent observations or trials. The distribution of S_n is the well known Binomial distribution denoted by $B(n,p,k) = \Pr(S_n = k) = \binom{n}{k}p^k q^{n-k}$. It is easy to see that $\mathbf{E}[S_n] = np$ and $var(S_n) = npq$.

The classical de Moivre-Laplace limit theorem states that for large n $\frac{S_n - np}{\sqrt{npq}} \sim N(0,1)$.

Let us also note that the above result implies the LLN. First note that:

$$\Pr\{|\frac{S_n}{n} - p| \leq \varepsilon\} = \Pr\{|\frac{S_n - np}{\sqrt{npq}}| \leq \varepsilon\sqrt{\frac{n}{pq}}\}$$

Now from the de Moivre-Laplace theorem:

$$\Pr\{|\frac{S_n - np}{\sqrt{npq}}| \leq \varepsilon\sqrt{\frac{n}{pq}}\} - \frac{1}{\sqrt{2\pi}}\int_{-\varepsilon\sqrt{\frac{n}{pq}}}^{\varepsilon\sqrt{\frac{n}{pq}}} e^{-\frac{x^2}{2}}dx \to 0 \ as \ n \to \infty$$

Noting that $\lim_{n\to\infty}\int_{-\varepsilon\sqrt{\frac{n}{pq}}}^{\varepsilon\sqrt{\frac{n}{pq}}} e^{-\frac{x^2}{2}}dx = 1$ we have : $\Pr\{|\frac{S_n}{n} - p| \leq \varepsilon\} \to 1$ which is the conclusion of the WLLN.

Note that while the Gaussian (or Normal approximation) works for fixed p and large n there is also another approximation which is of interest when the probability of success is small. This is the so-called Poisson convergence theorem which we state and prove below.

Proposition 2.7.1 *(Poisson approximation of Binomial probabilities)*
Let $B(n, p(n), k)$ be the Binomial distribution with the probability of success $p(n)$ such that : $p(n) \to 0$ as $n \to \infty$ and $\lim_{n\to\infty} np(n) \to \lambda > 0$.
Then for $k = 0, 1, \ldots$

$$|B(n, p(n), k) - \pi_k| \to 0 \ as \ n \ \to \ \infty$$

where $\{\pi_k\}$ denotes the Poisson distribution given by:

$$\pi_k = \frac{\lambda^k}{k!}e^{-\lambda}$$

Proof: Since by assumption $p(n) = \frac{\lambda}{n} + o(\frac{1}{n})$ we have:

$$B(n, p(n), k) = \frac{n(n-1)\cdots(n-k+1)}{k!}[\frac{\lambda}{n} + o(\frac{1}{n})]^k[1 - \frac{\lambda}{n} + o(\frac{1}{n})]^{n-k}$$

Now as $n \to \infty$,

$$\frac{n(n-1)\cdots(n-k+1)}{k!}[\frac{\lambda}{n} + o(\frac{1}{n})]^k \to \frac{\lambda^k}{k!}$$

and

$$[1 - \frac{\lambda}{n} + o(\frac{1}{n})]^{n-k} \to e^{-\lambda}$$

Hence $B(n, p(n), k) \to \pi_k$ as $n \to \infty$.

While there are other limit theorems of interest, in most applications the SLLN and the CLT turn out to be the most useful.

Concluding remarks

In this chapter we defined the notion of a stochastic process and studied their characterization in terms of distributions and sample-paths. We delineated several important classes of processes which can be characterized in terms of their moments. Finally we have discussed how empirical means are related to the means under the probability measure (the ensemble average) through the use of ergodic theorems. In the sequel we will study in detail w.s.s. processes and Markov processes which take values in a discrete set i.e. Markov chains.

Exercises

1. Let $\{X_t\}$ be a random telegraph process defined as follows.

   (a) X_t takes values $\{-1, 1\}$.
   (b) $\mathbb{P}(X_0 = 1) = \mathbb{P}(X_0 = -1) = \frac{1}{2}$
   (c) $X_t = (X_0)(-1)^{N_t}$ where $\{N_t\}$ is a Poisson process with intensity $\lambda > 0$.

   Find $\mathbf{E}[X_t]$, $var(X_t)$ and $cov(X_t, X_s)$. Show that it is an independent increment process.

2. Let $\{X_t\}$ be a 0 mean Gaussian process with covariance

$$R(t, s) = \exp\{-2\lambda|t - s|\}$$

   Find the conditional density $p_{X_t/X_s}(x/y)$. Show that it is not an independent increment process.

3. Let $\{X_t\}$ be the stochastic process defined below:

$$X_t = \sqrt{2}cos(2\pi f t + \phi)$$

   where f and ϕ are random variables with the following densities:

$$\begin{aligned} p_\phi(\phi) &= \frac{1}{2\pi} \quad 0 \leq \phi \leq 2\pi \\ p_f(f) &= \frac{2\lambda}{\lambda^2 + \pi^2 f^2} \quad -\infty < f < \infty \end{aligned}$$

   Find $\mathbf{E}[X_t]$ and $cov(X_t, X_s)$.

   The above problems show that there is very little information about a stochastic process obtained by knowing the mean and covariance without specifying the sample-paths.

4. Show that a Poisson process is mean square continuous and almost surely continuous at every t.

5. Let $\{X_t\}$ be a Gaussian process for $t \in [0, T]$. Let $\mathbf{E}[|X_{t+h} - X_t|^2] = h^\alpha$ for some $\alpha > 0$. Show that $\{X_t\}$ is almost surely sample continuous no matter how small α can be.

6. Let $\{X_t;\ -\infty < t < \infty\}$ be the stochastic process defined by:

$$X_t = A\cos(2\pi t + \theta)$$

   where A is a non-negative r.v. independent of θ which is uniformly distributed in $[0, 2\pi)$.

   (a) Show that $\{X_t\}$ is a stationary process.
   (b) Show that $\mathbf{E}[X_t] = 0$ provided $\mathbf{E}[A] < \infty$.
   (c) If A has the Rayleigh distribution given by:

$$p_A(x) = \frac{x}{\sigma^2}\exp\{-\frac{x^2}{2\sigma^2}\} \quad ; x \geq 0$$

   Show that $\{X_t\}$ is a Gaussian process. Is $\{X_t\}$ Markov?

7. a) Show $\sqrt{\lambda} W_{\frac{t}{\lambda}}$ is a standard Brownian motion process.

   b) Let $\{X_t;\ -\infty < t < \infty\}$ be a 0 mean Gaussian process with $\mathbf{E}[X_t X_s] = e^{-\lambda|t-s|}$. Express X_t for $t \geq 0$ in the form

$$X_t = f(t) W_{\frac{g(t)}{f(t)}}$$

   where $\{W_t; t \geq 0\}$ is a standard Brownian motion.

8. Consider the finite sequence of mean 0 jointly Gaussian r.v's $\{X_i\}_{i=1}^{10}$ with:

$$\mathbf{E}[X_i X_j] = 2^{-|i-j|}$$

   Find:

   a) $\mathbf{E}[X_5 / X_4, X_3]$

   b) $\mathbf{E}[X_7 / X_8, X_9, X_{10}]$

   c) $\mathbf{E}[X_6 X_9 / X_7, X_8]$

9. Let $\{N_t; t \geq 0\}$ be a non-homogeneous Poisson process with time-dependent intensity $\lambda_t > 0$ for all t defined as follows:

   a) $N_0 = 0$

   b) For $t > s$, $\mathbb{P}(N_t - N_s = k) = \frac{(\int_s^t \lambda_u du)^k}{k!} \exp\{-\int_s^t \lambda_u du\}$

   Show that N_t has independent increments. Let $\{T_n\}$ be the points of N_t i.e. $N_t = n;\ t \in [T_n, T_{n+1})$. Find the probability density $p_{T_n}(t)$ and show that $\sum_{n=1}^{\infty} p_{T_n}(t) = \lambda_t$. Show that $\mathbb{P}(T_1 < \infty)) = 1 \Rightarrow \int_0^{\infty} \lambda_u du = \infty$.

10. Let $\{X_n\}$ be the discrete-time Gauss Markov process defined by:

$$X_{n+1} = a X_n + b w_n$$

   where $X_0 \sim N(0, \sigma^2)$ and $\{w_n\}$ is an i.i.d. $N(0,1)$ sequence independent of X_0. Define $R_k = E[X_k^2]$.

   a) Show that $\{X_n\}$ converges asymptotically to a stationary process if $R_k \rightarrow R_\infty$ as $n \rightarrow \infty$.

   b) Show that $R_n \rightarrow R_\infty$ if $|a| < 1$. Find R_∞.

   c) Under the condition above, find what σ^2 must be in order that $\{X_n\}$ is a stationary process.

   d) If $a > 1$ show that $\frac{X_n}{a^n}$ converges in the mean square.

11. Let $\{X_n\}$ be a sequence of non-negative identically distributed r.v's with $\mathbf{E}[X_n] < \infty$. Then show that :

$$\frac{X_n}{n} \overset{a.s}{\rightarrow} 0$$

12. Let N_t be a Poisson process with intensity λ. Show that

$$\frac{N_t}{t} \overset{a.s}{\rightarrow} \lambda$$

13. Let $\{X_n\}$ be i.i.d. r.v's with mean m and finite variance. Define $S_n = \sum_1^n X_k$. Then show that:

$$\mathbf{E}[X_1/S_n, S_{n+1}, ...] \stackrel{a.s}{\to} m$$

14. Let X_t be a zero mean w.s.s. process with covariance $R(t)$. Suppose that for some $T > 0$, $R(T) = R(0)$. Then :

   a) Show that $\mathbb{P}(X_{t+T} = X_t) = 1$ for every t.

   b) Show that $R(t + KT) = R(t)$ for every t and not just for t=0.

   c) Find the best linear mean squared estimate of X_{t+KT} given X_t i.e. find the constant C which minimizes $\mathbf{E}[(X_{t+KT} - CX_t)^2]$.

15. Let $\{X_n\}$ be a second order sequence (mean 0). Show that a necessary and sufficient condition for $\{X_n\}$ to converge in the mean square is that:

$$\mathbf{E}[X_n X_m] \to C$$

   as $n, m \to \infty$ independently and C is a constant. Using this condition repeat problem 10.

16. Let $\{X_n\}$ be a sequence of independent $\{0, 1\}$ r.v's with $\Pr(X_n = 1) = p_n$ and $\Pr(X_n = 0) = 1 - p_n$. Show that as $n \to \infty$

   (a) $X_n \stackrel{P}{\to} 0 \iff p_n \to 0$

   (b) $X_n \stackrel{L_p}{\to} 0 \iff p_n \to 0$

   (c) $X_n \stackrel{a.s.}{\to} 0 \iff \sum_{n=1}^{\infty} p_n < \infty$

17. Let us now consider some applications of the Chebychev inequality and the CLT in obtaining so-called sample size estimates in random sampling situations.

   Let p be the fraction of a population who have a preference for a given product (or candidate). We choose n randomly sampled members of the population. Let M_n be the fraction of the sampled population who have the preference. M_n is thus an estimate of p.

   a) Using Chebychev's inequality and the fact that $p(1-p) \leq 0.25$ find the size of the population n if we want to obtain an estimate of p within 0.01 with 95% confidence.

   b) Via the CLT show that a better estimate (smaller n) of the sample size is possible for the same level of confidence.

Bibliography

The material for this section can be found in any standard text on stochastic processes. However, the following list of books may be consulted for more details (particularly 1 and 2).

1. E. Wong and B. Hajek; *Stochastic processes in engineering systems*, Springer-Verlag, N.Y., 1985

2. H. Cramer and M. R. Leadbetter; *Stationary and related stochastic processes*, J. Wiley and Sons, 1967

3. W. B. Davenport and W. L. Root; *An introduction to the theory of random signals*, Mc Graw-Hill, 1958

4. A. Papoulis; *Probability, random variables and stochastic processes*, McGraw-Hill, 1984