Stochastic processes. Generalities

1. Discrete time stochastic process

We call *discrete time stochastic process* any sequence \( X_0, X_1, X_2, \ldots, X_n, \ldots \) of random variables defined on a probability space \((\Omega, F, P)\), taking values in \(\mathbb{R}\). This definition is not so rigid with respect to small details: the same name is given to sequences \( X_1, X_2, \ldots, X_n, \ldots \), or to the case when the r.v. \( X_n \) take values in a space different from \(\mathbb{R}\). We shall also describe below the case when the time index takes negative values.

The main objects attached to a r.v. are its law, its first and second moments (and possibly higher order moments and characteristic or generating function, and the distribution function). We do the same for a process \( (X_n)_{n \geq 0} \): the probability density of the r.v. \( X_n \), when it exists, will be denoted by \( f_n(x) \), the mean by \( \mu_n \), the standard deviation by \( \sigma_n \). Often, we shall write \( t \) in place of \( n \), but nevertheless here \( t \) will be always a non-negative integer. So, our first concepts are:

i) **mean function and variance function**:

\[
\mu_t = E[X_t], \quad \sigma_t^2 = Var[X_t], \quad t = 0, 1, 2, \ldots
\]

In addition, the time-correlation is very important. We introduce three functions:

ii) the **autocovariance function** \( C(t, s) \), \( t, s = 0, 1, 2, \ldots \):

\[
C(t, s) = Cov(X_t, X_s) = E[(X_t - \mu_t)(X_s - \mu_s)]
\]

and the function

\[
R(t, s) = E[X_t X_s]
\]

(the name will be discussed below). They are symmetric \((R(t, s) = R(s, t)\) and the same for \( C(t, s) \)) so it is sufficient to know them for \( t \geq s \). We have

\[
C(t, s) = R(t, s) - \mu_t \mu_s, \quad C(t, t) = \sigma_t^2.
\]

In particular, when \( \mu_t = 0 \) (which is often the case), \( C(t, s) = R(t, s) \). Most of the importance will be given to \( \mu_t \) and \( R(t, s) \). In addition, let us introduce:

iii) the **autocorrelation function**

\[
\rho(t, s) = Corr(X_t, X_s) = \frac{C(t, s)}{\sigma_t \sigma_s}
\]

We have

\[
\rho(t, t) = 1, \quad |\rho(t, s)| \leq 1.
\]
The functions $C(t, s)$, $R(t, s)$, $\rho(t, s)$ are used to detect repetitions in the process, self-similarities under time shift. For instance, if $(X_n)_{n\geq 0}$ is roughly periodic of period $P$, $\rho(t + P, t)$ will be significantly higher than the other values of $\rho(t, s)$ (except $\rho(t, t)$ which is always equal to 1). Also a trend is a form of repetitions, self-similarity under time shift, and indeed when there is a trend all values of $\rho(t, s)$ are quite high, compared to the cases without trend. See the numerical example below.

Other objects (when defined) related to the time structure are:

iv) the joint probability density $f_{t_1, \ldots, t_n}(x_1, \ldots, x_n), \quad t_n \geq \ldots \geq t_1$

of the vector $(X_{t_1}, \ldots, X_{t_n})$ and

v) the conditional density

$f_{t|s}(x|y) = \frac{f_{t,s}(x,y)}{f_s(y)}, \quad t > s.$

Now, a remark about the name of $R(t, s)$. In Statistics and Time Series Analysis, the name *autocorrelation function* is given to $\rho(t, s)$, as we said above. But in certain disciplines related to signal processing, $R(t, s)$ is called autocorrelation function. There is no special reason except the fact that $R(t, s)$ is the fundamental quantity to be understood and investigated, the others ($C(t, s)$ and $\rho(t, s)$) being simple transformations of $R(t, s)$. Thus $R(t, s)$ is given the name which mostly reminds the concept of self-relation between values of the process at different times. In the sequel we shall use both languages and sometimes we shall call $\rho(t, s)$ the *autocorrelation coefficient*.

The last object we introduce is concerned with two processes simultaneously: $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$. It is called:

vi) *cross-correlation function*

$$C_{X,Y}(t, s) = E\left[ (X_t - E[X_t]) (Y_s - E[Y_s]) \right].$$

This function is a measure of the similarity between two processes, shifted in time. For instance, it can be used for the following purpose: one of the two processes, say $Y$, is known, has a known shape of interest for us, the other process, $X$, is the process under investigation, and we would like to detect portions of $X$ which have a shape similar to $Y$. Hence we shift $X$ in all possible ways and compute the correlation with $Y$.

When more than one process is investigated, it may be better to write $R_X(t, s)$, $C_X(t, s)$ and so on for the quantities associated to process $X$.

### 1.1. Example 1: *white noise*.

The white noise with intensity $\sigma^2$ is the process $(X_n)_{n\geq 0}$ with the following properties:

i) $X_0, X_1, X_2, \ldots, X_n, \ldots$ are independent r.v.’s

ii) $X_n \sim N(0, \sigma^2)$.

It is a very elementary process, with a trivial time-structure, but it will be used as a building block for other classes of processes, or as a comparison object to understand the features of more complex cases. The following picture has been obtained by R software by the commands `x<-rnorm(1000)`; `ts.plot(x)`.
Let us compute all its relevant quantities (the check is left as an exercise):\
\[ \mu_t = 0 \quad \sigma_t^2 = \sigma^2 \]
\[ R(t, s) = C(t, s) = \sigma^2 \cdot \delta(t - s) \]
where the symbol \( \delta(t - s) \) denotes 0 for \( t \neq s \), 1 for \( t = s \),
\[ \rho(t, s) = \delta(t - s) \]
\[ f_{t_1, \ldots, t_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i) \quad \text{where} \quad p(x) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \]
\[ f_{t|s}(x|y) = p(x). \]

1.2. Example 2: random walk. Let \((W_n)_{n \geq 0}\) be a white noise (or more generally, a process with independent identically distributed \(W_0, W_1, W_2, \ldots\)). Set
\[ X_0 = 0 \]
\[ X_{n+1} = X_n + W_n, \quad n \geq 0. \]
This is a random walk. White noise has been used as a building block: \((X_n)_{n \geq 0}\) is the solution of a recursive linear equation, \emph{driven} by white noise (we shall see more general examples later on). The following picture has been obtained by R software by the commands \(x \leftarrow \text{rnorm}(1000); y \leftarrow \text{cumsum}(x); \text{ts.plot}(y)\).

The random variables \(X_n\) are not independent (\(X_{n+1}\) obviously depends on \(X_n\)). One has
\[ X_{n+1} = \sum_{i=0}^{n} W_i. \]

We have the following facts We prove them by means of the iterative relation (this generalizes better to more complex discrete linear equations). First,
\[ \mu_0 = 0 \]
\[ \mu_{n+1} = \mu_n, \quad n \geq 0 \]
hence \(\mu_n = 0\) for every \(n \geq 0\).

By induction, \(X_n\) and \(W_n\) are independent for every \(n\), hence:
Exercise 1. Denote by $\sigma^2$ the intensity of the white noise; find a relation between $\sigma_{n+1}^2$ and $\sigma_n^2$ and prove that

$$\sigma_n = \sqrt{n} \sigma, \quad n \geq 0.$$ 

An intuitive interpretation of the result of the exercise is that $X_n$ behaves as $\sqrt{n}$, in a very rough way.

As to the time-dependent structure, $C(t, s) = R(t, s)$, and:

Exercise 2. Prove that $R(m, n) = n \sigma^2$, for all $m \geq n$ (prove it for $m = n$, $m = m+1$, $m = n+2$ and extend). Then prove that

$$\rho(m, n) = \sqrt{\frac{n}{m}}.$$ 

The result of this exercise implies that

$$\rho(m, 1) \to 0 \text{ as } m \to \infty.$$ 

We may interpret this result by saying that the random walk looses memory of the initial position.

2. Stationary processes

A process is called wide-sense stationary if $\mu_t$ and $R(t + n, t)$ are independent of $t$.

It follows that also $\sigma_t, C(t + n, t)$ and $\rho(t + n, t)$ are independent of $t$. Thus we speak of:

i) mean $\mu$

ii) standard deviation $\sigma$

iii) covariance function $C(n) := C(n, 0)$

iv) autocorrelation function (in the improper sense described above)

$$R(n) := R(n, 0)$$

v) autocorrelation coefficient (or also autocorrelation function, in the language of Statistics)

$$\rho(n) := \rho(n, 0).$$
A process is called strongly stationary if the law of the generic vector $(X_{n_1+t}, ..., X_{n_k+t})$ is independent of $t$. This implies wide stationarity. The converse is not true in general, but it is true for Gaussian processes (see below).

2.1. **Example: white noise.** We have

$$R(t, s) = \sigma^2 \cdot \delta(t - s)$$

hence

$$R(n) = \sigma^2 \cdot \delta(n).$$

2.2. **Example: linear equation with damping.** Consider the recurrence relation

$$X_{n+1} = \alpha X_n + W_n, \quad n \geq 0$$

where $(W_n)_{n \geq 0}$ is a white noise with intensity $\sigma^2$ and

$$\alpha \in (-1, 1).$$

The following picture has been obtained by R software by the commands ($\alpha = 0.9$, $X_0 = 0$):

```r
w <- rnorm(1000)
x <- rnorm(1000)
x[1]=0
for (i in 1:999) {
x[i+1] <- 0.9*x[i] + w[i]
}
ts.plot(x)
```

It has some features similar to white noise, but less random, more persistent in the direction where it moves.

Let $X_0$ be a r.v. independent of the white noise, with zero average and variance $\overline{\sigma}^2$. Let us show that $(X_n)_{n \geq 0}$ is stationary (in the wide sense) if $\overline{\sigma}^2$ is properly chosen with respect to $\sigma^2$. 
First we have
\[\mu_0 = 0, \quad \mu_{n+1} = \alpha \mu_n, \quad n \geq 0\]
hence \(\mu_n = 0\) for every \(n \geq 0\). The mean function is constant.

As a preliminary computation, let us impose that the variance function is constant. By induction, \(X_n\) and \(W_n\) are independent for every \(n\), hence
\[\sigma^2_{n+1} = \alpha^2 \sigma^2_n + \sigma^2, \quad n \geq 0.\]
If we want \(\sigma^2_{n+1} = \sigma^2_n\) for every \(n \geq 0\), we need
\[\sigma^2_n = \alpha^2 \sigma^2_n + \sigma^2, \quad n \geq 0\]
namely
\[\sigma^2_n = \frac{\sigma^2}{1 - \alpha^2}, \quad n \geq 0.\]
In particular, this implies the relation
\[\hat{\sigma}^2 = \frac{\sigma^2}{1 - \alpha^2}.\]
It is here that we first see the importance of the condition \(|\alpha| < 1\).

If we assume this condition on the law of \(X_0\), then we find
\[\sigma^2_1 = \alpha^2 \frac{\sigma^2}{1 - \alpha^2} + \sigma^2 = \frac{\sigma^2}{1 - \alpha^2} = \sigma^2_0\]
and so on, \(\sigma^2_{n+1} = \sigma^2_n\) for every \(n \geq 0\). Thus the variance function is constant.

Finally, we have to show that \(R(t + n, t)\) is independent of \(t\). We have
\[R(t + 1, t) = E[(\alpha X_t + W_t) X_t] = \alpha \sigma^2_n = \frac{\alpha \sigma^2}{1 - \alpha^2}\]
which is independent of \(t\);
\[R(t + 2, t) = E[(\alpha X_{t+1} + W_{t+1}) X_t] = \alpha R(t + 1, t) = \frac{\alpha^2 \sigma^2}{1 - \alpha^2}\]
and so on,
\[R(t + n, t) = E[(\alpha X_{t+n-1} + W_{t+n-1}) X_t] = \alpha R(t + n - 1, t)\]
\[= ... = \alpha^n R(t, t) = \frac{\alpha^n \sigma^2}{1 - \alpha^2}\]
which is independent of \(t\). The process is stationary. We have
\[R(n) = \frac{\alpha^n \sigma^2}{1 - \alpha^2}.\]

It also follows that
\[\rho(n) = \alpha^n.\]
The autocorrelation coefficient (as well as the autocovariance function) \textit{decays exponentially in time.}
2.3. Processes defined also for negative times. We may extend a little bit the previous definitions and call discrete time stochastic process also the two-sided sequences \( (X_n)_{n \in \mathbb{Z}} \) of random variables. Such processes are thus defined also for negative time. The idea is that the physical process they represent started in the far past and continues in the future.

This notion is particularly natural in the case of stationary processes. The function \( R(n) \) (similarly for \( C(n) \) and \( \rho(n) \)) are thus defined also for negative \( n \):

\[
R(n) = E[X_nX_0], \quad n \in \mathbb{Z}.
\]

By stationarity,

\[
R(-n) = R(n)
\]

because \( R(-n) = E[X_{-n}X_0] = E[X_{-n+n}X_0] = E[X_0X_n] = R(n) \). Therefore we see that this extension does not contain so much new information; however it is useful or at least it simplifies some computation.

3. Time series and empirical quantities

A time series is a sequence or real numbers, \( x_1, \ldots, x_n \). Also empirical samples have the same form. The name time series is appropriate when the index \( i \) of \( x_i \) has the meaning of time.

A finite realization of a stochastic process is a time series. Ideally, when we have an experimental time series, we think that there is a stochastic process behind. Thus we try to apply the theory of stochastic process.

Recall from elementary statistics that empirical estimates of mean values of a single r.v. \( X \) are computed from an empirical sample \( x_1, \ldots, x_n \) of that r.v.; the higher is \( n \), the better is the estimate. A single sample \( x_1 \) is not sufficient to estimate moments of \( X \).

Similarly, we may hope to compute empirical estimates of \( R(t, s) \) etc. from time series. But here, when the stochastic process has special properties (stationary and ergodic, see below the concept of ergodicity), one sample is sufficient! By “one sample” we mean one time series (which is one realization of the r.v. \( X \)). Again, the higher is \( n \), the better is the estimate, but here \( n \) refers to the length of the time series.

Consider a time series \( x_1, \ldots, x_n \). In the sequel, \( t \) and \( n_t \) are such

\[
t + n_t = n.
\]

Let us define

\[
\bar{x}_t = \frac{1}{n_t} \sum_{i=1}^{n_t} x_{i+t}, \quad \hat{\sigma}_t^2 = \frac{1}{n_t} \sum_{i=1}^{n_t} (x_{i+t} - \bar{x}_t)^2
\]

\[
\hat{R}(t) = \frac{1}{n_t} \sum_{i=1}^{n_t} x_i x_{i+t}
\]

\[
\hat{C}(t) = \frac{1}{n_t} \sum_{i=1}^{n_t} (x_i - \bar{x}_0)(x_{i+t} - \bar{x}_t)
\]

\[
\hat{\rho}(t) = \frac{\hat{C}(t)}{\hat{\sigma}_0 \hat{\sigma}_t} = \frac{\sum_{i=1}^{n_t} (x_i - \bar{x}_0)(x_{i+t} - \bar{x}_t)}{\sqrt{\sum_{i=1}^{n_t} (x_i - \bar{x}_0)^2 \sum_{i=1}^{n_t} (x_{i+t} - \bar{x}_t)^2}}.
\]
These quantities are taken as approximations of
\[ \mu_t, \sigma_t^2, R(t, 0), C(t, 0), \rho(t, 0) \]
respectively. In the case of stationary processes, they are approximations of
\[ \mu, \sigma^2, R(t), C(t), \rho(t) . \]
In the section on ergodic theorems we shall see rigorous relations between these empirical and theoretical functions.

The empirical correlation coefficient
\[ \hat{\rho}_{X,Y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \]
between two sequences \( x_1, \ldots, x_n \) and \( y_1, \ldots, y_n \) is a measure of their linear similarity. If the there are coefficients \( a \) and \( b \) such that the residuals
\[ \varepsilon_i = y_i - (ax_i + b) \]
are small, then \( |\hat{\rho}_{X,Y}| \) is close to 1; precisely, \( \hat{\rho}_{X,Y} \) is close to 1 if \( a > 0 \), close to -1 if \( a < 0 \). A value of \( \hat{\rho}_{X,Y} \) close to 0 means that no such linear relation is really good (in the sense of small residuals). Precisely, smallness of residuals must be understood compared to the empirical variance \( \hat{\sigma}_Y^2 \) of \( y_1, \ldots, y_n \):
one can prove that
\[ \hat{\rho}_{X,Y}^2 = 1 - \frac{\hat{\sigma}_Y^2}{\hat{\sigma}_Y^2} \]
(the so called explained variance, the proportion of variance which has been explained by the linear model). After these remarks, the intuitive meaning of \( \hat{R}(t), \hat{C}(t) \) and \( \hat{\rho}(t) \) should be clear: they measure the linear similarity between the time series and its \( t \)-translation. It is useful to detect repetitions, periodicity, trend.

**Example 1.** Consider the following time series, taken form EUROSTAT database. It collects export data concerning motor vehicles accessories, since January 1995 to December 2008.

![Image of time series](image.png)

*Its autocorrelation function \( \hat{\rho}(t) \) is given by*
4. GAUSSIAN PROCESSES

We see high values (the values of $\hat{\rho}(t)$ are always smaller than 1 in absolute value) for all time lag $t$. The reason is the trend of the original time series (highly non stationary).

**Example 2.** If we consider only the last few years of the same time series, precisely January 2005 - December 2008, the data are much more stationary, the trend is less strong. The autocorrelation function $\hat{\rho}(t)$ is now given by

![Graph of autocorrelation function]

where we notice a moderate annual periodicity.

4. Gaussian processes

If the generic vector $(X_{t_1}, ..., X_{t_n})$ is jointly Gaussian, we say that the process is Gaussian. The law of a Gaussian vector is determined by the mean vector and the covariance matrix. Hence the law of the marginals of a Gaussian process are determined by the mean function $\mu_t$ and the autocorrelation function $R(t,s)$.

**Proposition 1.** For Gaussian processes, stationarity in the wide and strong sense are equivalent.

**Proof.** Given a Gaussian process $(X_n)_{n \in \mathbb{N}}$, the generic vector $(X_{t_1+s}, ..., X_{t_n+s})$ is Gaussian, hence with law determined by the mean vector of components

$$E[X_{t_1+s}] = \mu_{t+s}$$

and the covariance matrix of components

$$\text{Cov}(X_{t_1+s}, X_{t_j+s}) = R(t_i + s, t_j + s) - \mu_{t_i+s}\mu_{t_j+s}.$$ 

If the process is stationary in the wide sense, then $\mu_{t_i+s} = \mu$ and

$$R(t_i + s, t_j + s) - \mu_{t_i+s}\mu_{t_j+s} = R(t_i - t_j) - \mu^2$$

do not depend on $s$. Then the law of $(X_{t_1+s}, ..., X_{t_n+s})$ does not depend on $s$. This means that the process is stationary in the strict sense. The converse is a general fact. The proof is complete. \(\square\)

Most of the models in these notes are obtained by linear transformations of white noise. White noise is a Gaussian process. Linear transformations preserve gaussianity. Hence the resulting processes are
Gaussian. Since we deal very often with stationary processes in the wide sense, being them Gaussian they also are strictly stationary.

5. An ergodic theorem

There exist several versions of ergodic theorems. The simplest one is the Law of Large Numbers. Let us recall it in its simplest version, with convergence in mean square.

**Proposition 2.** If \((X_n)_{n \geq 1}\) is a sequence of uncorrelated r.v. \((\text{Cov}(X_i, X_j) = 0\) for all \(i \neq j)\), with finite and equal mean \(\mu\) and variance \(\sigma^2\), then \(\frac{1}{n} \sum_{i=1}^{n} X_i\) converges to \(\mu\) in mean square:

\[
\lim_{n \to \infty} E \left[ \left( \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right)^2 \right] = 0.
\]

It also converges in probability.

**Proof.**

\[
\frac{1}{n} \sum_{i=1}^{n} X_i - \mu = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)
\]

hence

\[
\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right|^2 = \frac{1}{n^2} \sum_{i,j=1}^{n} (X_i - \mu)(X_j - \mu)
\]

\[
E \left[ \left( \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right)^2 \right] = \frac{1}{n^2} \sum_{i,j=1}^{n} E[(X_i - \mu)(X_j - \mu)] = \frac{1}{n^2} \sum_{i,j=1}^{n} \text{Cov}(X_i, X_j)
\]

\[
= \frac{1}{n^2} \sum_{i,j=1}^{n} \sigma^2 \delta_{ij} = \frac{\sigma^2}{n} \to 0.
\]

Recall that Chebyshev inequality states (in this particular case)

\[
P \left( \left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| > \varepsilon \right) \leq \frac{E \left[ \left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right|^2 \right]}{\varepsilon^2}
\]

for every \(\varepsilon > 0\). Hence, from the computation of the previous proof we deduce

\[
P \left( \left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| > \varepsilon \right) \leq \frac{\sigma^2}{\varepsilon^2 \cdot n}.
\]

In itself, this is an interesting estimate on the probability that the sample average \(\frac{1}{n} \sum_{i=1}^{n} X_i\) differs from \(\mu\) more than \(\varepsilon\). It follows that

\[
\lim_{n \to \infty} P \left( \left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| > \varepsilon \right) = 0
\]
for every $\varepsilon > 0$. This is the convergence in probability of $\frac{1}{n} \sum_{i=1}^{n} X_i$ to $\mu$.

**Remark 1.** Often this theorem is stated only in the particular case when the r.v. $X_i$ are independent and identically distributed, with finite second moment. We see that the proof is very easy under much more general assumptions.

**Remark 2.** It can be generalized to the following set of assumptions: $(X_n)_{n \geq 1}$ is a sequence of uncorrelated r.v.; the moments $\mu_n = E[X_n]$ and $\sigma_n^2 = \text{Var}[X_n]$ satisfy

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mu_i = \mu, \quad \sigma_n^2 \leq \sigma^2 \text{ for every } n \in \mathbb{N}$$

for some finite constants $\mu$ and $\sigma^2$. Under these assumptions, the same results hold.

We have written the proof, very classical, so that the proof of the following lemma is obvious.

**Lemma 1.** Let $(X_n)_{n \geq 1}$ be a sequence of r.v. with finite second moments and equal mean $\mu$. Assume that

$$\lim_{n \to \infty} \frac{1}{n^2} \sum_{i,j=1}^{n} \text{Cov}(X_i, X_j) = 0. \tag{5.1}$$

Then $\frac{1}{n} \sum_{i=1}^{n} X_i$ converges to $\mu$ in mean square and in probability.

The lemma will be useful if we detect interesting sufficient conditions for (5.1). Here is our main ergodic theorem. Usually by the name ergodic theorem one means a theorem which states that the time-averages of a process converge to a deterministic value (the mean of the process, in the stationary case).

**Theorem 1.** Assume that $(X_n)_{n \geq 1}$ is a wide sense stationary process (this ensures in particular that $(X_n)_{n \geq 1}$ is a sequence of r.v. with finite second moments and equal mean $\mu$). If

$$\lim_{n \to \infty} C(n) = 0$$

then $\frac{1}{n} \sum_{i=1}^{n} X_i$ converges to $\mu$ in mean square and in probability.

**Proof.** Since $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$, we have

$$\left| \sum_{i,j=1}^{n} \text{Cov}(X_i, X_j) \right| \leq \sum_{i,j=1}^{n} |\text{Cov}(X_i, X_j)| \leq 2 \sum_{i=1}^{n} \sum_{j=1}^{n} |\text{Cov}(X_i, X_j)|$$

so it is sufficient to prove that

$$\lim_{n \to \infty} \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |\text{Cov}(X_i, X_j)| = 0.$$ 

Since the process is stationary, $\text{Cov}(X_i, X_j) = C(i-j)$ so we have to prove $\lim_{n \to \infty} \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |C(i-j)| = 0$. But

$$\sum_{i=1}^{n} \sum_{j=1}^{n} |C(i-j)| = \sum_{i=1}^{n} \sum_{k=0}^{i-1} |C(k)|$$

The proof is completed.
\begin{align*}
&= |C(0)| + (|C(0)| + |C(1)|) + (|C(0)| + |C(1)| + |C(2)|) + \ldots + (|C(0)| + \ldots + |C(n-1)|) \\
&= n |C(0)| + (n-1) |C(1)| + (n-2) |C(2)| + \ldots + |C(n-1)| \\
&= \sum_{k=0}^{n-1} (n-k) |C(k)| \leq n \sum_{k=0}^{n-1} |C(k)|.
\end{align*}

Therefore it is sufficient to prove \( \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} |C(k)| = 0 \). If \( \lim_{n \to \infty} C(n) = 0 \), for every \( \varepsilon > 0 \) there is \( n_0 \) such that for all \( n \geq n_0 \) we have \( |C(n)| \leq \varepsilon \). Hence, for \( n \geq n_0 \),
\[
\frac{1}{n} \sum_{k=0}^{n-1} |C(k)| \leq \frac{1}{n} \sum_{k=0}^{n_0-1} |C(k)| + \frac{1}{n} \sum_{k=n_0}^{n-1} \varepsilon \leq \frac{1}{n} \sum_{k=0}^{n_0-1} |C(k)| + \varepsilon.
\]

Since \( \sum_{k=0}^{n_0-1} |C(k)| \) is independent of \( n \), there is \( n_1 \geq n_0 \) such that for all \( n \geq n_1 \)
\[
\frac{1}{n} \sum_{k=0}^{n_0-1} |C(k)| \leq \varepsilon.
\]

Therefore, for all \( n \geq n_1 \),
\[
\frac{1}{n} \sum_{k=0}^{n-1} |C(k)| \leq 2\varepsilon.
\]

This means that \( \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} |C(k)| = 0 \). The proof is complete. \( \square \)

5.1. Rate of convergence. Concerning the rate of convergence, recall from the proof of the LLG that
\[
E \left[ \left( \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right)^2 \right] \leq \frac{\sigma^2}{n}.
\]

We can reach the same result in the case of the ergodic theorem, under a suitable assumption.

**Proposition 3.** If \( (X_n)_{n \geq 1} \) is a wide sense stationary process such that
\[
\alpha := \sum_{k=0}^{\infty} |C(k)| < \infty
\]
(this implies \( \lim_{n \to \infty} C(n) = 0 \)) then
\[
E \left[ \left( \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right)^2 \right] \leq \frac{2\alpha}{n}.
\]
Proof. It is sufficient to put together several steps of the previous proof:

\[
E \left[ \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right]^2 = \frac{1}{n^2} \sum_{i,j=1}^{n} \text{Cov} (X_i, X_j) \leq \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |\text{Cov} (X_i, X_j)| \\
\leq \frac{2}{n} \sum_{k=0}^{n-1} |C(k)| \leq \frac{2\alpha}{n}.
\]

The proof is complete. \qed

Notice that the assumptions of these two ergodic results (especially the ergodic theorem) are very general and always satisfied in our examples.

5.2. Empirical autocorrelation function. Very often we need the convergence of time averages of certain functions of the process: we would like to have

\[
\frac{1}{n} \sum_{i=1}^{n} g(X_i) \rightarrow \eta_g
\]

in mean square, for certain functions \( g \). We need to check the assumptions of the ergodic theorem for the sequence \( (g(X_n))_{n \geq 1} \). Here is a simple example.

Proposition 4. Let \( (X_n)_{n \geq 0} \) be a wide sense stationary process, with finite fourth moments, such that \( E[X_n^2 X_{n+k}^2] \) is independent of \( n \) and

\[
\lim_{k \to \infty} E[X_0^2 X_k^2] = E[X_0^2]^2.
\]

In other words, we assume that

\[
\lim_{k \to \infty} \text{Cov} (X_0^2, X_k^2) = 0.
\]

Then \( \frac{1}{n} \sum_{i=1}^{n} X_i^2 \) converges to \( E[X_1^2] \) in mean square and in probability.

Proof. Consider the process \( Y_n = X_n^2 \). The mean function of \( (Y_n) \) is \( E[X_n^2] \) which is independent of \( n \) by the wide-sense stationarity of \( (X_n) \). For the autocorrelation function

\[
C(n, n+k) = E[Y_n Y_{n+k}] - E[Y_n^2] = E[X_n^2 X_{n+k}^2] - E[X_n^2]^2
\]

we need the new assumption of the proposition. Thus \( (Y_n) \) is wide-sense stationary. Finally, from the assumption \( \lim_{k \to \infty} E[X_0^2 X_k^2] = E[X_0^2]^2 \), which means \( \lim_{k \to \infty} C_Y(k) = 0 \) where \( C_Y(k) \) is the autocorrelation function of \( (Y_n) \), we can apply the ergodic theorem. The proof is complete. \qed

More remarkable is the following result, related to the estimation of \( R(n) \) by sample path autocorrelation function. Given a process \( (X_n)_{n \geq 1} \), call sample path (or empirical) autocorrelation function the process

\[
\frac{1}{n} \sum_{i=1}^{n} X_i X_{i+k}.
\]
Theorem 2. Let \((X_n)_{n \geq 0}\) be a wide sense stationary process, with finite fourth moments, such that \(E[X_nX_{n+k}X_{n+j}X_{n+j+k}]\) is independent of \(n\) and

\[
\lim_{j \to \infty} E[X_0X_kX_jX_{j+k}] = E[X_0X_k]^2 \quad \text{for every } k = 0, 1, \ldots
\]

In other words, we assume that

\[
\lim_{j \to \infty} \text{Cov}(X_0X_k, X_jX_{j+k}) = 0.
\]

Then the sample path autocorrelation function \(\frac{1}{n} \sum_{i=1}^{n} X_iX_{i+k}\) converges to \(R(k)\) as \(n \to \infty\) in mean square and in probability. Precisely, for every \(k \in \mathbb{N}\), we have

\[
\lim_{n \to \infty} E \left[ \left( \frac{1}{n} \sum_{i=1}^{n} X_iX_{i+k} - R(k) \right)^2 \right] = 0
\]

and similarly for the convergence in probability.

Proof. Given \(k \in \mathbb{N}\), consider the new process \(Y_n = X_nX_{n+k}\). Its mean function is constant in \(n\) because of the wide-sense stationarity of \((X_n)\). For the autocorrelation function,

\[
R_Y(n, n+j) = E[Y_nY_{n+j}] = E[X_nX_{n+k}X_{n+j}X_{n+j+k}]
\]

it is independent of \(n\) by assumption. Moreover, \(C_Y(j)\) converges to zero. Thus it is sufficient to apply the ergodic theorem, with the remark that \(E[Y_0] = R(k)\). The proof is complete.

With similar proof one can obtain other results of the same type. Notice that the additional assumptions that \(E[X_nX_{n+k}]^2\) and \(E[X_nX_{n+k}X_{n+j}X_{n+j+k}]\) are independent of \(n\) are a consequence of the assumption that \((X_n)\) is stationary in the strict sense. Thus, stationarity in the strict sense can be put as an assumption for several ergodic statements. Recall that wide-sense stationarity plus gaussianity implies strict-sense stationarity.

Remark 3. There is no way to check ergodic assumptions in applications. Thus one should believe they hold true. The intuition behind this faith is that they are true when the random variables \(X_n\), for very large \(n\), become roughly independent of the random variables \(X_i\) with small \(i\). In the ideal (never true) case when \(X_n\) would be exactly independent of \(X_i\), we would have \(\text{Cov}(X_0X_k, X_nX_{n+k}) = 0\) for given \(k\) and very large \(n\), and so on. Therefore, when we think that the stochastic process of a certain application looses memory, tend to assume values independent from those at the beginning, as time goes to infinity, then we believe the assumptions of the ergodic theorems are satisfied.