Random Processes
ECE537
Part II

Stationary Random Processes

Many random processes have the property that the nature of the randomness does not change with time. This randomness is exhibited in the probability law that defines the process. We have stated that a random process is completely specified in terms of the probability law if for any integer k, and any set of points in time $t_1, t_2, \cdots, t_k$, the joint CDF is specified. A process is *stationary* if these specifications are not dependent on time shifts of these points. In other words, the CDF corresponding to random variables at the points $t_1, t_2, \cdots, t_k$ is the same as the CDF corresponding to random variables at the points $t_1 + \tau, t_2 + \tau, \cdots, t_k + \tau$, where τ represents a time shift.

More precisely, a random process $X(t)$ is stationary if the joint CDFs that define the process have the following property

$$F_{X(t_1),\cdots,X(t_k)}(x_1, x_2, \cdots, x_k) = F_{X(t_1+\tau),\cdots,X(t_k+\tau)}(x_1, x_2, \cdots, x_k)$$

for all integers k, all sets of time points $\{t_1, \cdots, t_k\}$, and all time shifts τ.

There are weaker notions of stationarity. To differentiate among them the notion of stationary defined here is also referred to as *strict sense stationarity (SSS)*.

Jointly Stationary

Two processes $X(t)$ and $Y(t)$ are said to be jointly stationary if for any two integers k, n, and any two sets of points in time $t_1, \cdots, t_k$, and $t'_1, \cdots, t'_n$, the joint CDF of the random variables $X(t_1), \cdots, X(t_k), Y(t'_1), \cdots, Y(t'_n)$ is equal to the joint CDF for the random variables $X(t_1 + \tau), \cdots, X(t_k + \tau), Y(t'_1 + \tau), \cdots, Y(t'_n + \tau)$.

Discrete Time Processes

Note that the property of stationarity applies equally well to continuous or discrete time processes with the obvious restriction of the points in time, $t_1, \cdots, t_k$, to the points of the discrete process and the time shift τ also being a discrete time shift.

Example of Stationary Process

Consider a discrete time process X_n where the X_i's are independent identically distributed random variables. Then we can show that X_n is a stationary process.

For any set of integers $i_1, i_2, \cdots, i_k$, and any integer m (a discrete time shift) we have

$F_{X_{i_1+m},X_{i_2+m},\cdots,X_{i_k+m}}(j_1,j_2,\cdots,j_k) = F_{X_{i_1+m}}(j_1)F_{X_{i_2}+m}(j_2)\cdots F_{X_{i_k+m}}(j_k)$ (by independence)

$= F_{X_{i_1}}(j_1)F_{X_{i_2}}(j_2)\cdots F_{X_{i_k}}(j_k)$ (by the identical distribution assumption)

$= F_{X_{i_1},X_{i_2},\cdots,X_{i_k}}(j_1,j_2,\cdots,j_k)$ (by independence)

Example: Non-Stationary Process

The Wiener process and the Poisson process that we have discussed before are not stationary processes.

Example: Non-Stationary Process

Let $X(t) = \cos(2\pi f_0 t + \theta)$, where f_0 is a constant and θ is a random variable with uniform distribution on $\left[0, \frac{\pi}{2}\right]$. $X(t)$ is not a stationary process. Let $k = 1$ and take $t_1 = 0$. Then $X(0) = \cos\theta$ is a random variable with a density function that is non-zero in the interval [0,1], and zero elsewhere. On the other hand if we let $\tau = \frac{1}{4f_0}$, i.e. $t_1 + \tau = \frac{1}{4f_0}$, then $X(t_1 + \tau)$ has a density with non-zero values in $[-1,0]$, and zero elsewhere. Therefore the process $X(t)$ can not be stationary.

Note that it is much easier to prove that a process is non-stationary than it is to prove that the process is stationary. To prove non-stationarity all we need to do is to find one counter-example, whereas to prove stationarity we need to show that the stationarity property holds for all integers k, and all sets of points $t_1, \cdots, t_k$.

Stationary Process: First Order Distribution

According to the property of a stationary process, if $k = 1$, i.e. take one point in time

$$F_{X(t)}(x) = F_{X(t+\tau)}(x) = F_X(x)$$

for all x, τ.

Hence $m_X(t) = \mathcal{E}(X(t)) = m$.

$$\text{var}(X(t)) = \mathcal{E}((X(t)^2) - \left(\mathcal{E}(X(t))\right)^2 = \sigma^2$$

which is independent of t because the CDF for $X(t)$ is independent of t. In other words, stationary processes have constant mean and variance.

Stationary Processes: Second Order Moments

The auto-correlation and auto-covariance functions for a random process involve the joint CDF at two points, t_1 and t_2

Consider any two arbitrary points t_1, t_2. Let $\tau = -t_1$. Then from the stationarity property we have

$$F_{X(t_1),X(t_2)}(x_1,x_2) = F_{(X(t_1+\tau),X(t_2+\tau))}(x_1,x_2) = F_{X(0),X(t_2-t_1)}(x_1,x_2)$$

We can see that the auto-correlation function $R_X(t_1, t_2)$ depends only on the difference $t_2 - t_1$. As a result if we let $t_1 = t$, and $t_2 = t + \tau$, then the dependence is on $t_2 - t_1 = \tau$. Assume that $t_2 > t_1$. In the above we could subtract t_2 from both t_1 and t_2 and obtain $t_1 - t_2$ and 0. Hence, the stationarity property implies that $F_{X(t_1),X(t_2)}(x_1, x_2) = F_{X(t_1-t_2),X(0)}(x_1, x_2)$. As a result both the correlation and auto-covariance functions depend only on the time difference. We usually write as follows:

$$R_X(\tau) = \mathcal{E}\big(X(t)X(t+\tau)\big)$$

$$C_X(\tau) = \mathcal{E}\{(X(t) - m)(X(t+\tau) - m)\}$$

where $m = \mathcal{E}\big(X(t)\big)$. Note that in writing the above we are re-defining the auto-correlation and auto-covariance functions to be functions of one variable. We could use $\tilde{R}$ and $\tilde{C}$, but for simplicity and assuming that the context is clear we use the same symbols.

Wide Sense Stationary Random Processes

Determining whether a process is stationary may be difficult in some cases especially because we do not have the joints distributions for all sets of time points. However, in many cases it is sufficient to work with second moments. Hence, if the mean is constant, and if we evaluate the auto-correlation function and determine that it is only a function of the time spacing then we say that the random process is *wide sense stationary.* In evaluating the auto-correlation function we usually evaluate $\mathcal{E}\{X(t)X(t+\tau)\}$, and determine if the resulting expression is only a function of τ and not t.

Example

Consider the process $X(t) = A\cos(\omega t + \theta)$, where ω is a constant, and A and θ are independent random variables. If θ has a uniform distribution on $[0,2\pi]$ then the process is wide-sense stationary. If θ has a uniform distribution on $[0, 7\pi/4\,]$ then the process $X(t)$ is not wide-sense stationary. To show this we check the mean and auto-covariance

$$\mathcal{E}\big(X(t)\big) = \mathcal{E}(A)\mathcal{E}(\cos(\omega t + \theta)) = 0$$

$$C_X(t_1, t_2) = \frac{\mathcal{E}(A^2)}{2}\mathcal{E}(\cos(\omega(t_1 + t_2) + 2\theta) + \cos\big(\omega(t_2 - t_1)\big)$$

$$= \frac{\mathcal{E}(A^2)}{2}\cos(\omega(t_2 - t_1))$$

And note that indeed the process is wide sense stationary.

Now consider the same process except that now θ has a uniform distribution in the set $\left\{0, -\frac{\pi}{2}, \frac{\pi}{2}, \pi\right\}$. Using the same evaluations as above we show that the process is also wide-sense stationary. However, in this case it is easy to show that the process is not strict-sense stationary

(SSS). To see this consider the random variables $X(0) = A\cos\theta$, and $X\left(\frac{\pi}{4\omega}\right) = A\cos\left(\frac{\pi}{4} + \theta\right)$ and shown that these do not have the same distribution. Hence the process cannot be SSS.

Example: Telegraph Signal

Previously we showed that the mean of the telegraph signal is 0 and the auto-covariance function is $C_X(t_1, t_2) = e^{-\alpha|t_2 - t_1|}$. Since the auto-covariance depends only on the difference $(t_2 - t_1)$, the telegraph signal is a wide-sense stationary process.

The Random Walk and the Poisson Processes

We have previously shown that for the random walk $R_X(t_1, t_2) = \alpha \cdot \min(t_1, t_2)$. Hence the random walk is not a wide-sense stationary process.

Similarly for the Poisson process we have shown that $R_N(t_1, t_2) = \alpha \min(t_1, t_2)$. Hence it is also not a wide-sense stationary process.

Jointly Wide-Sense Stationary

We say that the processes $X(t)$ and $Y(t)$ are *jointly wide-sense stationary* if they are both WSS and their cross-covariance depends only on $t_2 - t_1$.

- $\mathcal{E}(X(t)) = m_X$, i.e. independent of t
- $\mathrm{cov}(X(t_1), X(t_2)) = C_X(t_2 - t_1)$
- $\mathcal{E}(Y(t)) = m_Y$
- $\mathrm{cov}(Y(t_1), Y(t_2)) = C_Y(t_2 - t_1)$
- $\mathrm{cov}(X(t_1), Y(t_2)) = C_{XY}(t_2 - t_1)$

Wide-Sense Stationary Gaussian Process

Note that strict-sense stationary (SSS) implies wide-sense stationarity (WSS) but the converse is not necessarily true as a we have seen in an example above. However in the case of Gaussian processes WSS is equivalent to SSS. This is a result of the special properties of the Gaussian Distribution for a random vector.

Properties of the Auto-Correlation for WSS Processes

1. The auto-correlation function evaluated at zero yields the second moment,

$$R_X(0) = \mathcal{E}(X^2(t))$$

This is also called the average power of the random signal $X(t)$.

2. The auto-correlation function $R_X(\tau)$ has a maximum at 0. To show this we use the Cauchy-Schwarz Inequality.

From the Cauchy-Schwarz Inequality we have the following

$$R_X^2(\tau) = (\mathcal{E}\{X(t)X(t+\tau)\})^2 \leq \mathcal{E}\big(X^2(t)\big)\mathcal{E}\big(X^2(t+\tau)\big) = R_X^2(0)$$

Hence $|R_X(\tau)| \leq R_X(0)$.

Note that $R_X(0)$ is positive, hence the square root of $R_X^2(0)$ is $R_X(0)$. However $R_X(\tau)$ can be positive or negative for $\tau \neq 0$.

3. The auto-correlation function gives a measure of change in the process as we observe its behaviour with increasing τ. Consider the change in the process from t to $t+\tau$.

$$P(|X(t+\tau) - X(t)| > \epsilon) = P(|X(t+\tau) - X(t)|^2 > \epsilon^2) \leq \mathcal{E}\left(\frac{|X(t+\tau) - X(t)|^2}{\epsilon^2}\right)$$
$$= \frac{2\big(R_X(0) - R_X(\tau)\big)}{\epsilon^2}$$

   Where we have used the Markov inequality.

   If $R_X(0) - R_X(\tau)$ is small, i.e. the curve $R_X(\tau)$ drops off slowly then the probability of a large change in $X(t)$ in τ seconds is small.

4. The auto-correlation function is an even function of τ. To show this we write

$$R_X(-\tau) = \mathcal{E}(X(t)X(t-\tau))$$

   Since this is true for all t we may replace t with $t+\tau$, hence we obtain

$$R_X(-\tau) = \mathcal{E}\big(X(t+\tau)X(t+\tau-\tau)\big) = \mathcal{E}\big(\big(X(t)X(t+\tau)\big) = R_X(\tau)$$

5. If $R_X(d) = R_X(0)$ then $R_X(\tau)$ is periodic with period d. To show this we could use mathematical induction. First we show that $R_X(2d) = R_X(0)$. This can be proved using the triangle identity for a metric. We have an inner product for two random variables $\langle X_1, X_2\rangle = \mathcal{E}(X_1X_2)$. We can define a metric between two random variables (just like a metric between two functions) as $d(X_1, X_2) = \big||X_1 - X_2|\big|$, where $\big||X|\big| = \big(\mathcal{E}(X^2)\big)^{\frac{1}{2}}$. This is the standard metric in $\mathcal{L}_2$ spaces. Then using the triangle inequality for a metric, i.e. $d(X_1, X_2) \leq d(X_1, Y) + d(X_2, Y)$, we obtain $d\big(X(t), X(t+2d)\big) \leq d\big(X(t), X(t+d)\big) + d\big(X(t+2d), X(t+d)\big)$. Then $\big||X(t) - X(t+2d)|\big| \leq \big||X(t) - X(t+d)|\big| + ||X(t+2d) - X(t+d)||$.

   But $\big||X(t) - X(t+d)|\big|^2 = \langle X(t) - X(t+d), X(t) - X(t+d)\rangle$
   $= \langle X(t), X(t)\rangle + \langle X(t+d), X(t+d)\rangle - 2\langle X(t), X(t+d)\rangle = 2R_X(0) - 2R_X(d) = 0$, since $R_X(d) = R_X(0)$ by assumption.

The same argument can be made for $||X(t+2d) - X(t+d)||$ by replacing $t+\tau$ with t. Hence $R_X(2d) = R_X(0)$.

Note that the above argument can be repeated to shown that $R_X(\tau)$ is periodic.

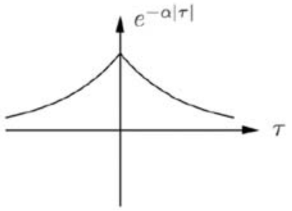
6. Let $X(t) = m + N(t)$, where $N(t)$ is a zero-mean process for which $R_N(\tau) \to 0$ as $\tau \to \infty$, Then $R_X(\tau) = \mathcal{E}[\big(m + N(t)\big)(m + N(t+\tau))] = m^2 + m\mathcal{E}\big(N(t)\big) + m\mathcal{E}\big(m(t+\tau)\big) + R_N(\tau)$

$$= m^2 + R_N(\tau)$$
$$\to m^2, \quad \text{as } \tau \to \infty$$

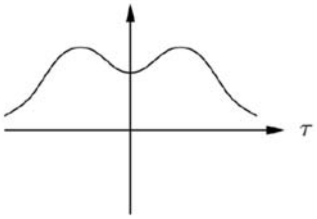
$R_X(\tau)$ approaches the square of the mean as $\tau \to \infty$.

Example of Possible Autocorrelation Functions

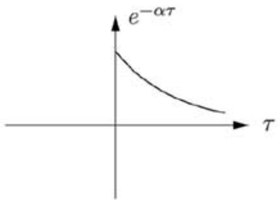
Which of these functions cannot represent the auto-correlation function for a WSS process.



(a)



(b)



(c)

a) This is a possible auto-correlation function, and in fact processes exist with this autocorrelation function.
b) This is not a possible auto-correlation function because it violates the requirement that the maximum occurs at $\tau = 0$.
c) This is not a possible auto-correlation function because it is not an even function.

Wide Sense Stationary Gaussian Processes

If a Gaussian random process is WSS then it is also Strict Sense Stationary (SSS). This follows because WSS is a condition on 2nd moments. For SSS we need a condition on the PDF. However for a Gaussian vector the second moments along with the mean completely specify the PDF.

Proof:

Assume that $X(t)$ is a WSS Gaussian process. Now consider the PDF for the vector

$\boldsymbol{X} = \big(X(t_1+\tau), \cdots, X(t_k+\tau)\big)$. Strict sense stationarity means that this PDF is independent of τ.

The mean of $\boldsymbol{X}$ is a constant m because of the WSS assumption. The covariance matrix for $\boldsymbol{X}$ is $C = (c_{ij}) = \left(\mathcal{E}\left(X(t_i+\tau)X(t_j+\tau)\right)\right) = \left(\mathcal{E}\left(X(t_i)X(t_j)\right)\right)$, where the last equality follows from the WSS property.

Now due to the Gaussian assumption the PDF is

$$f_{X(t_1+\tau),\cdots,X(t_k+\tau)}(x_1,\cdots,x_k) = \frac{e^{-\frac{1}{2}(x-m)^T C^{-1}(x-m)}}{(2\pi)^{\frac{k}{2}}|C|^{\frac{1}{2}}}$$

Thus,

$$f_{X(t_1),\cdots,X(t_k)}(x_1,\cdots,x_k) = \frac{e^{-\frac{1}{2}(x-m)^T C^{-1}(x-m)}}{(2\pi)^{\frac{k}{2}}|C|^{\frac{1}{2}}} = f_{X(t_1+\tau),\cdots,X(t_k+\tau)}(x_1,\cdots,x_k)$$

This proves that the process is SSS.

Note that we have assumed that the covariance matrix C is non-singular, i.e. the inverse exists. However we can show that the same result follows if the covariance matrix is singular, i.e. has determinant equal to 0.

Cyclo-Stationary Random Processes

A random process $X(t)$ is said to be cyclo-stationary if the joint CDF is invariant with respect to time shifts of the origin by multiples of some constant T. In other words $X(t_1), X(t_2), \cdots, X(t_k)$ has the same joint CDF as $X(t_1+nT), X(t_2+nT), \cdots, X(t_k+nT)$ for some parameter T, all integers k, all sets of time points $t_1, \cdots, t_k$, and all integers n.

Wide-Sense Cyclo-Stationary

A process $X(t)$ is wide sense cyclo-stationary if the cyclo-stationary property, as above, applies to the mean and second order moments

$$m_X(t+nT) = m_X(t)$$

$$C_X(t_1+nT, t_2+nT) = C_X(t_1, t_2)$$

Example

In a binary digital communication system a pulse $g(t)$, is transmitted for each data symbol modulated by the data value, e.g. $d_k = \pm 1$. We assume that the d_k are independent i.i.d. random variables. The transmitted signal is

$$X(t) = \textstyle\sum_{i=-\infty}^{\infty} d_i g(t-iT) \qquad ****$$

This process is wide sense cyclo-stationary with the periodicity parameter T. Note that if a process is WSS then it is also Wide Sense Cyclo-Stationary. The converse is not true in general. In the case of $X(t)$, above, for some specific choices of $g(t)$, $X(t)$ may also be WSS.

To check that *** is wide sense stationary we do the following

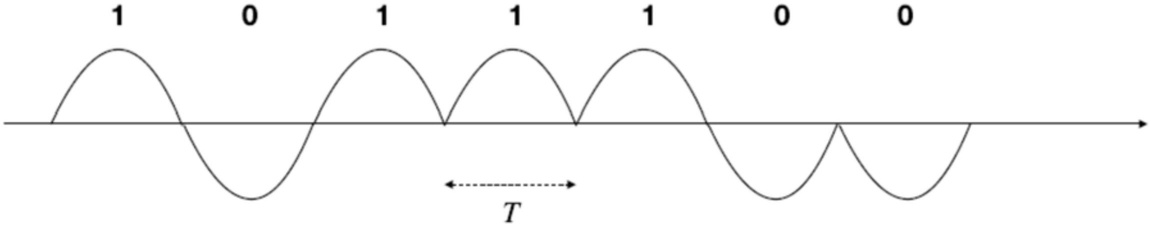
$$\mathcal{E}\big(X(t_1+nT)X(t_2+nT)\big) = \mathcal{E}\left(\sum_{i=-\infty}^{\infty} d_i g(t_1+nT-iT) \sum_{j=-\infty}^{\infty} d_j g(t_2+nT-jT)\right)$$

$$= \mathcal{E}\left(\sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} d_i d_j g(t_1 + nT - iT)g(t_2 + nT - jT)\right)$$

$$= \mathcal{E}\left(\sum_{i=-\infty}^{\infty} d_i^2 g(t_1 + nT - iT)g(t_2 + nT - iT)\right)$$

$$= \mathcal{E}\left(\sum_{i=-\infty}^{\infty} d_i^2 g(t_1 - iT)g(t_2 - iT)\right) = C(t_1, t_2)$$

Where the last equation was obtained by changing the index of summation.



Random Shift of a Cyclo-Stationary Process

In some applications such as finding the power spectral density (discussed later) we like to average the auto-correlation function for a cyclo-stationary process over one period. We do this formally by modified the random process model $X(t)$ by introducing a random shift in time θ which is a random variable with uniform probability on $[0, T]$, where T is the period of the auto-correlation function.

In many applications we compute the auto-correlation function of $X(t)$ with θ included as a parameter and then average over θ, or take the expected value over θ. An alternative point of view is as follows:

If $X(t)$ is a cyclo-stationary process, formally we introduce the process $Y(t) = X(t + \theta)$, where θ is a random variable with uniform distribution in $[0, T]$. The process $Y(t)$ has the following CDF

$$F_{Y(t_1),\cdots,Y(t_k)}(x_1, \cdots, x_k) = \frac{1}{T}\int_0^T F_{X(t_1+\theta),\cdots,X(t_k+\theta)}(x_1, \cdots, x_k)d\theta$$

Example

Consider the above digital communication system signal with $P(d_k = 1) = P(d_k = -1) = 1/2$.

$$X(t) = \sum_{k=-\infty}^{\infty} d_k g(t - kT)$$

$$Y(t) = \sum_{k=-\infty}^{\infty} d_k g(t - kT + \theta)$$

Let us determine the mean and covariance

First we condition on θ and compute $\mathcal{E}(\mathcal{E}(Y(t|\theta))$. The first expected value (the inner one) treats θ as a constant. Then the outer expected values is with respect to θ.

$$\mathcal{E}\big(Y(t)\big) = \sum_{k=-\infty}^{\infty} \mathcal{E}\big(d_k g(t - kT + \theta)\big) = \sum_{k=-\infty}^{\infty} \mathcal{E}(d_k)\mathcal{E}\big(g(t - kT + \theta)\big) = 0$$

$$R_Y(\tau) = \mathcal{E}\big(Y(t)Y(t+\tau)\big) = \mathcal{E}\big(X(t+\theta)X(t+\tau+\theta)\big)$$

$$\mathcal{E}\{\mathcal{E}(X(t+\theta)X(t+\tau+\theta)|\theta)\}$$

The inner expected value, i.e. with respect to the $d_k's$ is (i.e. treat θ as a constant)

$$\mathcal{E}\left(\sum_{k=-\infty}^{\infty} d_k g(t - kT + \theta) \sum_{n=-\infty}^{\infty} d_n g(t + \tau - nT + \theta)\right)$$

$$= \sum_{k=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} \mathcal{E}\big(d_k g(t - kT + \theta) d_n g(t + \tau - nT + \theta)\big)$$

$$= \sum_{k=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} \mathcal{E}(d_k d_n)\, g(t - kT + \theta) g(t + \tau - nT + \theta)$$

$$= \sum_{k=-\infty}^{\infty} g(t - kT + \theta) g(t + \tau - kT + \theta)$$

Where we have used $\mathcal{E}(d_k d_n) = \delta_{kn}$.

Now, we take the expected value with respect to the random variable θ

$$\mathcal{E}\left(\sum_{k=-\infty}^{\infty} g(t - kT + \theta) g(t + \tau - kT + \theta)\right) = \sum_{k=-\infty}^{\infty} \mathcal{E}(g(t - kT + \theta) g(t + \tau - kT + \theta))$$

$$= \sum_{k=-\infty}^{\infty} \frac{1}{T}\int_0^T g(t - kT + \theta) g(t + \tau - kT + \theta)\, d\theta$$

$$= \frac{1}{T} \sum_{k=-\infty}^{\infty} \int_{t-kT}^{t-kT+T} g(u)g(\tau + u)\, du$$

$$R_Y(\tau) = \frac{1}{T} \int_{-\infty}^{\infty} g(u)g(u + \tau)du$$

In the last step we recognize that the infinite sum is over a set of integrals over non-overlapping intervals whose union is the real line.

Note that the auto-covariance function depends only on τ and not t. Hence the process $Y(t)$ is WSS.

Continuity in a Random Process

We know about continuity for a real function $f(t)$ as follows: The function $f(t)$ is continuous at the point $t = t_0$ if for every $\epsilon > 0$, there exists a $\delta > 0$, such that for $|t - t_0| < \delta$, $|f(t) - f(t_0)| < \epsilon$. We would write $\lim_{t \to t_0} f(t) = f(t_0)$.

In the same manner we may consider a sample path of a random process, $X(t, \zeta)$ and determine continuity at each point in time, t. In this case ζ is treated as a parameter. We can then write

$$\lim_{t \to t_0} X(t, \zeta) = X(t_0, \zeta)$$

For some random processes the sample paths are continuous functions for all outcomes ζ and all t_0. Hence we may say that the random process is a continuous process.

However in many cases the process can be determined to be continuous only in a *probabilistic sense.*

Mean Square Continuity

A random process $X(t)$ is continuous at the point t_0 in the mean square sense if

$$\lim_{t \to t_0} \mathcal{E}\left\{ \left(X(t) - X(t_0)\right)^2 \right\} = 0$$

Note that in the case of sample paths $X(t, \zeta)$ (for fixed ζ) when we take a limit we are considering a sequence of numbers and whether or not that sequence converges to a specific value. The metric for convergence is the usual metric that we use to compare real numbers. For example, for two real numbers x, y, we use the metric $|x - y|$. In the case of mean square convergence we are considering a sequence of functions of ζ, where each function is indexed by t. The metric for convergence is the metric derived from the $\mathcal{L}_2$ norm, i.e. for the two functions of ζ, $X(t)$ and $X(t_0)$ the distance between them is $\left|\left|X(t) - X(t_0)\right|\right| = \left(\mathcal{E}\left\{\left(X(t) - X(t_0)\right)^2\right\}\right)^{1/2}$. Note that this may be clearer if we think of the random process as $X_t(\zeta)$ where t is now clearly shown as a parameter and the function is a function of ζ. Hence the above would be

$$\left|\left|X_t - X_{t_0}\right|\right| = \left(\mathcal{E}\left\{\left(X_t - X_{t_0}\right)^2\right\}\right)^{1/2}$$

Sometimes mean square continuity is denoted as

$$\text{l.i.m.}_{\,t\to t_0} X(t) = X(t_0)$$

Where l.i.m. denotes “limit in the mean”.

We say that the process $X(t)$ is mean square continuous if it is mean square continuous for all points in time t_0.

Note that previously we discussed the convergence of a sequence of random variable. We considered convergence according to various criteria. In this case we are considering a sequence of random variable $X(t_n, \zeta)$ and a target (limit) random variable $X(t, \zeta)$, where the sequence t_n approaches t, i.e. $t_n \to t$. Then continuity in the mean square sense means that the sequence of random variables $X(t_n, \zeta)$ converges to the random variable $X(t, \zeta)$ in the mean square sense.

Continuity of Sample Paths Implies Mean Square Continuity

We can show that if a random process is sample-path continuous then it is also continuous in the mean square sense, but the converse is not necessarily true. There are processes that are continuous in the mean square sense at all points in time. However, each sample function exhibits at least one point of discontinuity, for all sample functions except possibly for a set of sample functions in a set of measure 0.

Continuity of the Mean and Continuity of the Auto-Correlation of a Random Process

Let $X(t)$ be a random process which is continuous at t_0 in the mean square sense. Then

$$\lim_{t\to t_0} m_X(t) = m_X(t_0)$$

Proof: $0 \le \text{var}\big(X(t) - X(t_0)\big) = \mathcal{E}\left\{\big(X(t) - X(t_0)\big)^2\right\} - \left\{\mathcal{E}\big(X(t) - X(t_0)\big)\right\}^2$

Hence $\left\{\mathcal{E}\big(X(t) - X(t_0)\big)\right\}^2 \le \mathcal{E}\left\{\big(X(t) - X(t_0)\big)^2\right\}$

or

$$\big(m_X(t) - m_X(t_0)\big)^2 \le \mathcal{E}\left\{\big(X(t) - X(t_0)\big)^2\right\}$$

If the process is mean-square continuous at t_0 then the right side approaches 0 as $t \to t_0$. Hence the left side must also approach 0. Hence $m_X(t) \to m_X(t_0)$ as $t \to t_0$.

Now, consider the auto-correlation

$$\mathcal{E}\left\{\big(X(t) - X(t_0)\big)^2\right\} = \mathcal{E}\{X^2(t) - 2X(t)X(t_0) + X^2(t_0)\} = R_X(t,t) + R_X(t_0,t_0) - 2R_X(t,t_0)$$

If the auto-correlation function is continuous at $t = t_0$ then $\lim_{t \to t_0} R_X(t,t) = R_X(t_0,t_0)$, and $\lim_{t \to t_0} R_X(t,t_0) = R_X(t_0,t_0)$. Hence the right side in the above approaches 0. Hence the left side approaches 0 and the process is continuous at $t = t_0$ in the mean-square sense.

The Wiener and Poisson Processes

Both the Wiener and the Poisson processes are mean-square continuous processes. We can check this as follows:

For the Wiener process

$$R_X(t_0,t) = \alpha \min(t_0,t)$$

We can test the following as $t \to t_0$.

$$R_X(t,t) + R_X(t_0,t_0) - 2R_X(t,t_0) = \alpha \min(t,t) + \alpha \min(t_0,t_0) - 2\alpha \min(t,t_0)$$

If $t \to t_0$ from the left the above becomes $\alpha t + \alpha t_0 - 2\alpha t = \alpha(t_0 - t)$ which approaches 0 as $t \to t_0$.

Similarly if $t \to t_0$ from the right

$$\alpha \min(t,t) + \alpha \min(t_0,t_0) - 2\alpha \min(t,t_0) = \alpha t + \alpha t_0 - 2\alpha t_0 = \alpha(t - t_0) \to 0$$

For the Poisson process we had

$$R_N(t_1,t_2) = \lambda \min(t_1,t_2)$$

In a similar manner as for the Wiener process we argue that the process is mean-square continuous.

Note that for the Poisson process all sample functions except those in a set of measure 0 are not continuous.

On the other hand we can show that for the Wiener process all sample functions except for those in a set of measure 0 are continuous.

Continuity in WSS Processes

For WSS processes the condition on the auto-correlation function for the process to be mean-square continuous is simpler:

The process is mean-square continuous if the autocorrelation function is continuous at 0, $\lim_{\tau \to 0} R_X(\tau) = R_X(0)$, and if this holds then it is mean-square continuous at all points in time t_0.

We show this in a similar manner to the above

$$\mathcal{E}\left\{\big(X(t+\tau) - X(t)\big)^2\right\} = 2\big(R_X(0) - R_X(\tau)\big)$$

Hence if $\lim_{\tau \to 0} R_X(\tau) = R_X(0)$, then $\lim_{\tau \to 0} \mathcal{E}\left\{\big(X(t+\tau) - X(t)\big)^2\right\} = 0$ and the process is mean-square continuous. Conversely if the process is mean-square continuous, i.e. the left side approaches 0 then the right side approaches 0, and the auto-correlation function is continuous at $\tau = 0$.

Derivatives of Random Processes

Consider a random process $X(t)$. If all the sample functions of $X(t)$, i.e. each function $X(t,\zeta)$ for a fixed ζ is differentiable, then we can find a new process that is the derivative of the process $X(t)$. We may refer to this new process as $Y(t) = X'(t)$, or $Y(t) = \frac{dX(t)}{dt}$. Each sample function of $Y(t)$ is the derivative of the corresponding sample function of $X(t)$.

Note that the derivative for the sample function $X(t,\zeta)$ is defined as

$$X'(t,\zeta) = \lim_{\epsilon\to 0}\frac{X(t+\epsilon,\zeta) - X(t,\zeta)}{\epsilon}$$

However, for random processes where the sample functions are not differentiable we can define derivates using a different convergence criteria, in fact we can use convergence in the mean-square sense. We define the derivative as

$$X'(t) = \text{l.i.m.}_{\epsilon\to 0}\ \frac{X(t+\epsilon) - X(t)}{\epsilon}$$

This means that there exists a random process $X'(t)$ such that the following holds.

$$\lim_{\epsilon\to 0}\mathcal{E}\left\{\left(\frac{X(t+\epsilon) - X(t)}{\epsilon} - X'(t)\right)^2\right\} = 0$$

Note that this does not give us a procedure to determine $X'(t)$ from $X(t)$. It only gives us a test to determine if a particular process (lets call it $Y(t)$) is the derivative of the process $X(t)$ in the mean square sense. We may refer to the resulting derivative, obtained under the mean square convergence criterion, as the mean square derivative of $X(t)$ and we denote it by $X'(t)$.

If we don't have the function $X'(t)$ and we wish to test for convergence to determine the existence of the derivative in the mean square sense, then we can use the Cauchy-Criterion for the convergence of a sequence of random variables, which states that a Cauchy sequence converges to a function in the space (if the space is complete) – in this case the function is a random variable. In our case the Cauchy criterion is equivalent to the existence of the following limit:

$$\lim_{\substack{\epsilon_1\to 0\\ \epsilon_2\to 0}}\mathcal{E}\left\{\left(\frac{X(t+\epsilon_1) - X(t)}{\epsilon_1} - \frac{X(t+\epsilon_2) - X(t)}{\epsilon_2}\right)^2\right\} = 0$$

The left side of the above means that for every $\epsilon > 0$, there exists a $\delta > 0$, such that for all $\epsilon_1 < \delta$ and $\epsilon_2 < \delta$, the left side is less than ϵ.

Expanding the square inside the expected value we obtain

$$\left(\frac{X(t+\epsilon_1)-X(t)}{\epsilon_1}\right)^2 - 2\frac{[X(t+\epsilon_1)-X(t)]}{\epsilon_1}\frac{[X(t+\epsilon_2)-X(t)]}{\epsilon_2} + \left(\frac{X(t+\epsilon_2)-X(t)}{\epsilon_2}\right)^2$$

The limit of the expected value for the 1st and 3rd terms is

$$\lim_{\epsilon\to 0}\frac{R_X(t+\epsilon,t+\epsilon)-R_X(t+\epsilon,t)-R_X(t,t+\epsilon)+R_X(t,t)}{\epsilon^2}$$

The above limit, if it exists, is equal to $\frac{\partial^2 R_X(t,t)}{\partial t_1 \partial t_2}$, hence the first and 3rd terms yield $2\frac{\partial^2 R_X(t,t)}{\partial t_1 \partial t_2}$

The limit of the expected value for the middle term is

$$\lim_{\substack{\epsilon_1\to 0\\ \epsilon_2\to 0}} -2\frac{R_X(t+\epsilon_1,t+\epsilon_2)-R_X(t+\epsilon_1,t)-R_X(t,t+\epsilon_2)+R_X(t,t)}{\epsilon_1\epsilon_2}$$

This limit, if it exists, is equal to $-2\frac{\partial^2 R_X(t,t)}{\partial t_1 \partial t_2}$. Note that we take R_X as a function of two variables t_1 and t_2, then take the partial derivatives and then evaluate at $(t_1,t_2)=(t,t)$.

Hence $\lim_{\substack{\epsilon_1\to 0\\ \epsilon_2\to 0}} \mathcal{E}\left\{\left(\frac{X(t+\epsilon_1)-X(t)}{\epsilon_1}-\frac{X(t+\epsilon_2)-X(t)}{\epsilon_2}\right)^2\right\} = 2\frac{\partial^2 R_X(t,t)}{\partial t_1 \partial t_2} - 2\frac{\partial^2 R_X(t,t)}{\partial t_1 \partial t_2} = 0.$

We can see that a sufficient condition for the mean square derivative to exist at t is that $\frac{\partial^2 R_X(t,t)}{\partial t_1 \partial t_2}$ exists. Because if $\frac{\partial^2 R_X(t,t)}{\partial t_1 \partial t_2}$ exists then the limit $\lim_{\substack{\epsilon_1\to 0\\ \epsilon_2\to 0}} \mathcal{E}\left\{\left(\frac{X(t+\epsilon_1)-X(t)}{\epsilon_1}-\frac{X(t+\epsilon_2)-X(t)}{\epsilon_2}\right)^2\right\} = 0$, which is the requirement for the mean square derivative to exist.

Note that for a WSS process the above formulations become simpler since the autocorrelation function becomes a function of one variable. **A sufficient condition for the mean-square derivative to exist is that $R_X''(0)$ exists.**

Mean of the Derivative of a Random Process $X(t)$

For a random process $X(t)$, if the derivative $X'(t)$ exists (in the mean square sense), we have

$$\lim_{\epsilon\to 0}\mathcal{E}\left\{\left(\frac{X(t+\epsilon)-X(t)}{\epsilon}-X'(t)\right)^2\right\}=0$$

This means that $\lim_{\epsilon\to 0}\mathcal{E}\left\{\frac{X(t+\epsilon)-X(t)}{\epsilon}-X'(t)\right\}=0$ (*)

Why? Let $Y=\frac{X(t+\epsilon)-X(t)}{\epsilon}-X'(t)$, then

$0\le \mathrm{var}(Y)=\mathcal{E}(Y^2)-\mathcal{E}(Y)^2$. Hence $\mathcal{E}(Y^2)\ge\left(\mathcal{E}(Y)\right)^2$.

Thus if $\mathcal{E}(Y^2) \rightarrow 0$, then $\left(\mathcal{E}(Y)\right)^2 \rightarrow 0$, and we must have $\mathcal{E}(Y) \rightarrow 0$. Thus, continuing with (*), we have

$$\lim_{\epsilon\to 0} \mathcal{E}\left\{\frac{X(t+\epsilon)-X(t)}{\epsilon}\right\} - \mathcal{E}\big(X'(t)\big) = 0$$

or

$$\mathcal{E}\big(X'(t)\big) = \lim_{\epsilon\to 0} \mathcal{E}\left\{\frac{X(t+\epsilon)-X(t)}{\epsilon}\right\} = m_X'(t)$$

Thus the expected value of the mean square derivative is the derivative of the mean for the process. **In a sense we exchange the order of taking expectations and limits**. $\mathcal{E}(\cdot)$ and $\frac{d(\cdot)}{dt}$ are linear operators where the order may be exchanged.

Mean of the Derivative of a WSS Process

Note that for a WSS process then $m_X(t)$ is a constant. As a result the mean-square derivative has zero mean, i.e. $\mathcal{E}\big(X'(t)\big) = 0$.

Cross-Correlation between $\boldsymbol{X(t)}$ and $\boldsymbol{X'(t)}$

Let $Y(t) = X'(t)$

Then $R_{XY}(t_1, t_2) = \mathcal{E}\big(X(t_1)X'(t_2)\big) = \mathcal{E}\left\{X(t_1)\cdot \text{l.i.m.}_{\epsilon\to 0} \frac{(X(t_2+\epsilon)-X(t_2))}{\epsilon}\right\}$

$$= \lim_{\epsilon\to 0} \frac{\mathcal{E}\big(X(t_1)X(t_2+\epsilon)\big) - \mathcal{E}(X(t_1)X(t_2)}{\epsilon}$$

$$= \lim_{\epsilon\to 0} \frac{R_X(t_1, t_2+\epsilon) - R_X(t_1, t_2)}{\epsilon}$$

$$R_{XX'}(t_1, t_2) = \frac{\partial R_X(t_1, t_2)}{\partial t_2}$$

Auto-Correlation Functions of Derivative Process

$$R_{X'}(t_1, t_2) = \mathcal{E}\{X'(t_1)X'(t_2)\} = \mathcal{E}\left\{\text{l.i.m.}_{\epsilon\to 0} \frac{X(t_1+\epsilon)-X(t_1)}{\epsilon} X'(t_2)\right\}$$

$$= \lim_{\epsilon\to 0} \frac{\mathcal{E}\big(X(t_1+\epsilon)X'(t_2)\big) - \mathcal{E}\big(X(t_1)X'(t_2)\big)}{\epsilon}$$

$$= \lim_{\epsilon\to 0} \frac{R_{XX'}(t_1+\epsilon, t_2) - R_{XX'}(t_1, t_2)}{\epsilon}$$

$$= \frac{\partial R_{XX'}(t_1,t_2)}{\partial t_1}$$

$$= \frac{\partial}{\partial t_1}\left(\frac{\partial R_X(t_1,t_2)}{\partial t_2}\right)$$

$$R_{X'}(t_1,t_2) = \frac{\partial^2}{\partial t_1 \partial t_2} R_X(t_1,t_2)$$

Differentiating Wide-Sense Stationary Processes

Let $X(t)$ be a WSS process. Then the above conditions for the existence of the mean-square derivative simplify. The auto-correlation function becomes a function of the time difference. In other words $R_X(t_1,t_2) = \tilde{R}_X(t_2 - t_1) = \tilde{R}_X(\tau)$. Hence a sufficient condition for the existence of the mean square derivative is the following partial derivative evaluated at $(t_1,t_2) = (t,t)$.

$$\frac{\partial^2}{\partial t_1 \partial t_2} R_X(t_1,t_2) = \frac{\partial^2}{\partial t_1 \partial t_2} \tilde{R}_X(t_2 - t_1) = \frac{\partial}{\partial t_1} \tilde{R}'_X(t_2 - t_1) = -\tilde{R}''_X(t_2 - t_1)$$

where the last equality is the evaluation at the point (t,t). Note that the condition for the derivative to exist is that $\frac{\partial^2}{\partial t_1 \partial t_2} R_X(t_1,t_2)$ evaluated at $(t_1,t_2) = (t,t)$ exists.

Note that generally we abuse notation and represent both functions as R_X even though they are different. In fact R_X (in the above) is a function of two variables, and $\tilde{R}_X$ is a function of one variable. We usually use the symbol R_X for both and rely on the context to avoid confusion.

The autocorrelation functions in the case of WSS processes are as follows:

$$R_{XX'}(\tau) = \frac{\partial}{\partial t_2} R_X(t_1 - t_2) = -\frac{d}{d\tau} R_X(\tau)$$

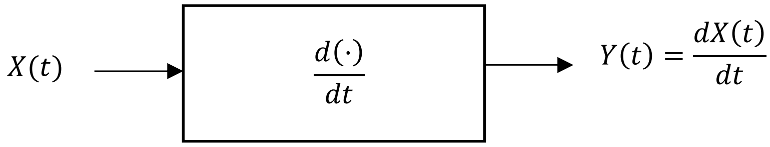
where $\tau = t_1 - t_2$.

$$R_{X'}(\tau) = \frac{\partial}{\partial t_1}\left\{\frac{\partial}{\partial t_2} R_X(t_1 - t_2)\right\} = \frac{\partial}{\partial t_1}\{-R'_X(t_1 - t_2)\} = -R''_X(t_1 - t_2) = -R''_X(\tau)$$

Differentiating a Gaussian Random Process

If $X(t)$ is a Gaussian random process for which the mean square derivative $X'(t)$ exists, then $X'(t)$ is also a Gaussian random process.

Note that differentiation is a linear operation and linear operations on Gaussian processes result in Gaussian processes. The proof would start with the expression $(X(t+\epsilon) - X(t))/\epsilon$ which is a linear combination of two Gaussian random variables and then take the limit as $\epsilon \rightarrow 0$ and show that the result is still a Gaussian random variable.



White Gaussian Noise (Derivative of Wiener Process)

We attempt to find the derivative of the Wiener process. The auto-correlation function is as follows:

$$R_X(t_1, t_2) = \alpha \min(t_1, t_2) = \begin{cases} \alpha t_2 & \text{for} \quad t_2 < t_1 \\ \alpha t_1 & \text{for} \;\; t_2 \geq t_1 \end{cases}$$

We can attempt to determine the auto-correlation function for the process $X'(t)$ using the expression derived above

$$R_{X'}(t_1, t_2) = \frac{\partial^2}{\partial t_1 \partial t_2} R_X(t_1, t_2)$$

Now we take partial derivatives with respect to t_1 and t_2.

$$\frac{\partial}{\partial t_2} R_X(t_1, t_2) = \begin{cases} \alpha & \text{for } \; t_2 < t_1 \\ 0 & \text{for } \; t_2 \geq t_1 \end{cases} = \alpha u(t_1 - t_2)$$

We note that if we fix t_1, the partial derivative is not defined at $t_2 = t_1$ because there is an abrupt change in slope.

Then taking the derivative with respect to t_1 we obtain $\alpha\delta(t_1 - t_2)$. Thus, we have

$$R_{X'}(t_1, t_2) = \alpha\delta(t_1 - t_2)$$

Strictly speaking the mean-square derivative of the Wiener process does not exist because the partial derivative condition is not satisfied, i.e. $\frac{\partial^2}{\partial t_1 \partial t_2} R_X(t, t)$ does not exist – it is infinite. However we may still work with the process in a formal way just like we work with delta functions in linear systems. The process $X'(t)$ has infinite power because $R_{X'}(0) = \infty$.

Note that the auto-correlation function is 0 for any $\tau \neq 0$ (i.e. $t_1 \neq t_2$). This means that the process varies very rapidly with time and with oscillations that approach infinity.

We will see later that it has a "flat" power spectral density. As a result we call it white noise, because it resembles white light, having power distributed over all frequencies.

Mean Square Integral (Integral of a Random Process in the Mean Square Sense)

We now define the integral for a random process $X(t)$ over the interval (t_0, t). We define it in the same manner as the Riemann Integral. We partition the interval (t_0, t) into a set of n equally spaced sub-intervals and form the Riemann sum

$$I_n = \sum_{k=1}^{n} X(t_k)\Delta$$

Where t_k is a point in the k^{th} sub-interval.

With deterministic functions we define the integral as the $\lim_{n\to\infty} I_n$, i.e. as $\Delta \to 0$. In more precise terms we form two approximations for the integral, one using a selection of points t_k that gives an upper-bond for the integral, i.e. it over-estimates the integral, and the other using another set of points t_k' which gives a lower both, i.e. it under-estimates the integral. Then we show that as $\Delta \to 0$, the two estimates converge to the same value, assuming appropriate conditions on the function that is being integrated.

In this case, where $X(t)$ is a random process, each of the sums I_n, when applied to a sample function $X(t,\zeta)$, is a random variable (i.e. a mapping from ζ to a real number. Hence we take the limit of the sequence of random variables I_n. As we have seen before there are different criteria to consider for convergence of sequences of random variables. In this case we use the mean square criterion.

We define the integral $\int_{t_0}^{t} X(t')dt' = \text{l.i.m.}_{\Delta\to 0} \sum_{k=1}^{n} X(t_k)\Delta$. Note that letting $\Delta \to 0$, is the same as letting $n \to \infty$. Note that the integral is defined in the same manner as the Riemann integral. For each sample function we compute a value for the integral in the same manner, i.e. Riemann sums. The difference lies in which random processes result in convergence (i.e. what are the conditions for convergence), and in this sense we have chosen the mean square convergence. Note that the integral is a random variable, i.e. for each sample function (i.e. each $\zeta \in \Omega$) we have a real number. If we consider the upper limit of the integral, i.e. t, as an index for the random variable then we have a random process which we may refer to as $Y(t)$, i.e.

$$Y(t) = \int_{t_0}^{t} X(t')dt'$$

Just as we have done for other random processes we may determine the mean and auto-correlation for the process $Y(t)$.

How do we test for convergence? For which processes does the integral exists?

As in the case of the use of a metric based on the $\mathcal{L}_2$ norm, we use the Cauchy criteria to determine convergence in the mean square sense. The condition for convergence is the following:

$$\mathcal{E}\left\{\left[\sum_{i=1}^{n_1} X(t_i)\Delta_1 - \sum_{k=1}^{n_2} X(t_k')\Delta_2\right]^2\right\} \to 0$$

as $\Delta_1, \Delta_2 \to 0$.

$$= \mathcal{E}\left\{\Delta_1^2 \sum_{i=1}^{n_1}\sum_{j=1}^{n_1} X(t_i)X(t_j) - 2\Delta_1\Delta_2 \sum_{i=1}^{n_1}\sum_{k=1}^{n_2} X(t_i)X(t'_k) + \Delta_2^2 \sum_{k=1}^{n_2}\sum_{m=1}^{n_2} X(t'_k)X(t'_m)\right\}$$

$$= \Delta_1^2 \sum_{i=1}^{n_1}\sum_{j=1}^{n_1} \mathcal{E}\left(X(t_i)X(t_j)\right) - 2\Delta_1\Delta_2 \sum_{i=1}^{n_1}\sum_{k=1}^{n_2} \mathcal{E}\left(X(t_i)X(t'_k)\right) + \Delta_2^2 \sum_{k=1}^{n_2}\sum_{m=1}^{n_2} \mathcal{E}(X(t'_k)X(t'_m))$$

$$= \Delta_1^2 \sum_{i,j} R_X(t_i, t_j) - 2\Delta_1\Delta_2 \sum_{i,k} R_X(t_i, t'_k) + \Delta_2^2 \sum_{k,m} R_X(t'_k, t'_m)$$

If the limit as $\boldsymbol{\Delta_1, \Delta_2 \to 0}$ of the above expression exists, then we can show that the expression has a limit of 0. Each of the three terms then converges to the following integral

$$\int_{u=t_0}^{t}\int_{v=t_0}^{t} R_X(u,v)dudv$$

Hence a sufficient condition for the existence of the integral of $X(t)$ in the mean square sense is that the above double integral of the auto-correlation function exists.

The Mean and Auto-Correlation of an Integral

In the above we have defined a new process, $Y(t)$ as the integral of the process $X(t)$. Just like for any process we can determine the mean and auto-correlation. In a sense the integral is a linear combination of the random variables $X(t)$ (indexed by t). We take the expectation of $Y(t)$ in the same way that we take the expectation of any linear combination, i.e. the expected value of a linear combination is the linear combination of the expected values. This means that we can exchange the order of the expectation operation and the integral operation

$$\mathcal{E}\big(Y(t)\big) = \mathcal{E}\left\{\int_{t_0}^{t} X(t')dt'\right\} = \int_{t_0}^{t} \mathcal{E}\big(X(t')\big)dt' = \int_{t_0}^{t} m_X(t')dt'$$

For the auto-correlation we have

$$R_Y(t_1, t_2) = \mathcal{E}\big(Y(t_1)Y(t_2)\big) = \mathcal{E}\left\{\int_{t_0}^{t_1} X(u)du \int_{t_0}^{t_2} X(v)dv\right\}$$

$$= \int_{t_0}^{t_1}\int_{t_0}^{t_2} \mathcal{E}\big(X(u)X(v)\big)dudv = \int_{t_0}^{t_1}\int_{t_0}^{t_2} R_X(u,v)dudv$$

Integral of a Gaussian Process

As a result of the integral of a random process being a linear operation on a set of random variables, if the process $X(t)$ is a Gaussian process then the integral process $Y(t)$, as discussed above, is also a Gaussian random process.

The Integral of White Gaussian Noise

Let $Z(t)$ be a white Gaussian noise process as discussed above. Define the integral process as the integral of $Z(t)$ over the interval $[0, t]$, i.e.

$$X(t) = \int_0^t Z(u)du$$

What is the auto-correlation function for $X(t)$?

Note that the process $Z(t)$ has auto-correlation function $R_Z(t_1, t_2) = \alpha\delta(t_1 - t_2)$. If we rename the function $R_Z(\cdot)$ as a function of one variable $\tau = t_1 - t_2$, we have $R_Z(\tau) = \alpha\delta(\tau)$.

The auto-correlation function of $X(t)$ is then

$$R_X(t_1, t_2) = \int_0^{t_1}\int_0^{t_2} R_Z(u, v)dudv = \alpha\int_0^{t_1}\int_0^{t_2} \delta(u - v)dudv = \alpha\int_0^{t_1}\int_{-v}^{t_2 - v} \delta(x)dxdv$$

$$= \alpha\int_0^{t_1} U(t_2 - v)dv = \begin{cases} \alpha t_1 & \text{if } t_1 < t_2 \\ \alpha t_2 & \text{if } t_1 \geq t_2 \end{cases}$$

Hence

$$R_X(t_1, t_2) = \alpha \min(t_1, t_2)$$

This is the auto-correlation function for the Wiener process. Hence, we see that if we differentiate the Wiener Process we obtain the white noise, and if we integrate white noise process we obtain the Wiener process. If the white noise is also Gaussian, as in the case of the derivative of the Wiener process then the integral of the white noise is also a Gaussian process.

Summary of Continuity, Derivative, and Integral

We now summarize the various concepts in the above starting with Continuity. In all the cases we needed a notion of convergence of some sequence of random variables. We used the notion of mean square convergence. In all the cases we consider the process $X(t)$

Mean Square Continuity at t_0

$$\lim_{t\to t_0} \mathcal{E}\left\{\left(X(t) - X(t_0)\right)^2\right\} = 0$$

Continuity of Sample Paths

If all the sample paths of a random process are continuous at t_0 then the process is mean square continuous at t_0. The converse is not necessarily true.

Continuity of the Mean of $X(t)$

We saw that $\left(m_X(t) - m_X(t_0)\right)^2 \leq \mathcal{E}\left\{\left(X(t) - X(t_0)\right)^2\right\}$

Hence if $X(t)$ is mean square continuous at $t = t_0$ then the mean $m_X(t)$ is continuous at $t = t_0$, i.e. **mean square continuity => the mean is continuous**

Continuity of the Autocorrelation function

We saw that $\mathcal{E}\left\{\left(X(t) - X(t_0)\right)^2\right\} = R_X(t,t) + R_X(t_0,t_0) - 2R_X(t,t_0).$ Hence

Continuity of the auto-correlation function at t_0 => $X(t)$ is mean square continuous at $t = t_0$.

Wiener Process and Poisson Process

Wiener Process has continuous sample paths for all t. Poisson Process does not have continuous sample paths. Both processes are continuous in the mean square sense.

Continuity for WSS processes

If $X(t)$ is WSS and $\lim_{\tau \to 0} R_X(\tau) = R_X(0)$ then $X(t)$ is mean square continuous at all points t.

Derivates of a Random Process

Sample path derivatives

$$X'(t,\zeta) = \lim_{\epsilon \to 0} \frac{X(t+\epsilon,\zeta) - X(t,\zeta)}{\epsilon}$$

Derivative in the Mean Square Sense

$$X'(t) = \text{l.i.m.}_{\epsilon \to 0} \frac{X(t+\epsilon) - X(t)}{\epsilon}$$

Condition for existence of mean square derivative at $t = t_0$, $\frac{\partial^2 R_X(t_1,t_2)}{\partial t_1 \partial t_2}$ exists at $(t_1,t_2) = (t,t)$.

Condition for existence of mean square derivative for a WSS process: $R_X''(\tau)$ exists at $\tau = 0$.

Mean of the Derivative of a Random Process

$$\mathcal{E}\left(X'(t)\right) = m_X'(t)$$

i.e. mean of the derivative of the process is the derivative of the mean function.

Mean of the Derivative of a WSS Process

$\mathcal{E}\left(X'(t)\right) = 0.$

Cross-Correlation between $\boldsymbol{X(t)}$ and $\boldsymbol{X'(t)}$

$$R_{XX'}(t_1,t_2)=\frac{\partial R_X(t_1,t_2)}{\partial t_2}$$

Autocorrelation of $\boldsymbol{X'(t)}$

$$R_{X'}(t_1,t_2)=\frac{\partial^2}{\partial t_1 \partial t_2}R_X(t_1,t_2)$$

Differentiating Wide-Sense Stationary Processes

The mean square derivative of $X(t)$ exists if $\tilde{R}_X''(0)$ exists.

Autocorrelation function for $X'(t)$ is $R_{X'}(\tau)=-R_X''(\tau)$.

Derivative of a Gaussian process is a Gaussian process

Derivative of the Wiener process is the white noise process.

Mean-Square Integral of $\boldsymbol{X(t)}$

$$\int_{t_0}^{t} X(t')dt' = \text{l.i.m.}_{\Delta\to 0}\sum_{k=1}^{n} X(t_k)\Delta$$

Condition for existence of the integral of $X(t)$: The following integral exists

$$\int_{u=t_0}^{t}\int_{v=t_0}^{t} R_X(u,v)dudv$$

Mean of the Integral equals integral of the mean

$$\mathcal{E}\left\{\int_{t_0}^{t} X(t')dt'\right\}=\int_{t_0}^{t} m_X(t')dt'$$

Autocorrelation Function for the integral $Y(t)=\int_{t_0}^{t} X(t')dt'$

$$R_Y(t_1,t_2)=\int_{t_0}^{t_1}\int_{t_0}^{t_2} R_X(u,v)dudv$$

Estimating the Mean of a Random Process

Suppose we wish to estimate the mean of a random process:

We may repeat the experiment N times with each time yielding the realization $X(t,\zeta_i)$. Then the mean may be estimated as

$$\widehat{m}_X(t) = \frac{1}{N}\sum_{i-1}^{N} X(t, \zeta_i)$$

In a sense what we are doing here is introducing a sequence of independent random processes $X_1(t,\zeta), X_2(t,\zeta), X_3(t,\zeta) \cdots$. If we fix the time $t = t_0$, then $X_1(t_0,\zeta), X_2(t_0,\zeta), X_3(t_0,\zeta) \cdots$ is a sequence of independent random variables, and under the appropriate conditions we can appeal to the law of large numbers and estimate the mean at each $t = t_0$.

However we may also attempt to estimate the mean for the process $\boldsymbol{X(t,\zeta)}$ by fixing $\zeta = \zeta_0$, and considering the sequence of random variables $X(t_{1,}, \zeta_0), X(t_2,\zeta_0), X(t_3,\zeta_0) \cdots$ and then taking the average

$$\frac{1}{N}\sum_{i=1}^{N} X(t_i, \zeta_0)$$

We assume that the points t_i are close in time, i.e. the spacing between them is small, and they are uniformly spaced by Δ starting at $t_1 = -T$ and ending at $t_N = T$. This means $\Delta \approx \frac{2T}{N}$, i.e. $\frac{1}{N} = \frac{\Delta}{2T}$. The above estimate is then

$$\frac{\Delta}{2T}\sum_{i=1}^{N} X(t_i, \zeta_0)$$

For large N we may want to estimate this sum as the integral

$$\frac{1}{2T}\int_{-T}^{T} X(t, \zeta_0)dt$$

The question now is the following: Do these two methods of estimating the mean yield the same result? Note that the first method, resulting from the introduction of a sequence of random processes, is conceptual, because it would not be practical in practice, since it requires many realizations of the process.

The second method is much more practical because it uses only one realization of the process.

It turns out that for some processes these two methods yield the same result, thus allowing us to use the second method to estimate the mean. The processes for which this approach to estimate the mean is valid are generally called *Ergodic Processes*. However we need to be more precise and also to generalize this idea of estimation for other parameters of a random process other than the mean.

Ergodic Processes

A process is ergodic if the statistical properties at a set of points in time can be inferred from the time properties of one sample function. This is not a precise definition. Generally we say that the process, $X(t,\zeta)$ is ergodic if time averages, i.e. averages over t, equal ensemble averages, i.e. averages over ζ. This is also not a precise definition of ergodicity, but it gives an idea.

Consider the following moments:

The mean:

The ensemble average at the time t_0, as $N \to \infty$ yields $\mathcal{E}\big(X(t_0)\big)$ according to the law of large numbers.

Time average, as $T \to \infty$, is: $\langle X(t)\rangle = \lim\limits_{T - \infty} \frac{1}{2T}\int_{-T}^{T} X(t)dt$

Autocorrelation:

Ensemble Average with times $\boldsymbol{t}$ and $\boldsymbol{t+\tau}$:

$R_X(\tau) = \mathcal{E}\{X(t+\tau)X(t)\}$

Note that for an ergodic process the ensemble average cannot depend on t, otherwise there is no possibility that they would equal the time averages from one sample function. There are many ensemble averages (one for each t) and only one time average for a single sample function. Hence the only way that ensemble averages are equal to time averages is if the ensemble averages are all equal to a constant.

Time average:

$$\langle X(t+\tau)X(t)\rangle = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} X(t+\tau)X(t)dt$$

Mean-Ergodic Processes

A process $X(t)$ is said to be mean ergodic if the time average $\langle X(t)\rangle$ computed from a **single** realization of the process $X(t)$ equals the ensemble average $\mathcal{E}\big(X(t)\big)$. This definition requires that the process have a constant mean in order to be mean-ergodic. This definition may be summarized as follows:

$$\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} X(t,\zeta)dt = \mathcal{E}(X(t))$$

Note that in the left we have a single realization of the process indexed by ζ. The variable t in the left is a dummy variable of integration corresponding to a single realization. On the right we have an expected value for a specific random variable at the fixed time t. This expected value must be the same for all t, because there is no dependency on t in the left side.

For discrete time processes X_n, the above definition (requirement) for a mean-ergodic process becomes

$$\lim_{N\to\infty}\frac{1}{2N+1}\sum_{k=-N}^{N} X_k(\zeta) = \mathcal{E}(X_n)$$

Mean-Ergodic Process: Example 1

Consider a discrete time process X_n, where the X_n for different n are independent random variables with the same distribution, i.e. i.i.d. random variables with equal mean m_X. Note that for a discrete process n corresponds to t for the continuous time case.

Then according to the strong law of large numbers we have

$$\frac{1}{2N+1}\sum_{k=-N}^{N} X_k(\zeta) - m \to 0$$

as $N \to \infty$ almost surely. In other words $P\left(\lim_{N\to\infty} \frac{1}{2N+1}\sum_{k=-N}^{N} X_k(\zeta) - m = 0\right) = 1$.

Hence the process X_n is mean ergodic.

We are interested in obtaining similar results for non i.i.d. random processes.

Example 2

Let $X(t) = A$ be a random process where A is a zero mean random variable with unit variance.

The ensemble average at time t is $\mathcal{E}\big(X(t)\big) = 0$.

However for a realization ζ, $X(t)$ is a constant function with value not necessarily equal to 0. Hence the time average is

$$\langle X(t)\rangle = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} X(t)dt = A$$

The time average is different for different samples functions! Hence the process $X(t)$ is not ergodic

Example 3

Consider two mean-ergodic processes $X_1(t)$ and $X_2(t)$. Form the process $Y(t)$, where

$$Y(t) = X_1(t) + \alpha X_2(t)$$

where α is a random variable independent of $X_2(t)$ taking values in $\{0,1\}$ with $P(\alpha = 0) = P(\alpha = 1) = \frac{1}{2}$. Is $Y(t)$ ergodic?

Ensemble average $\mathcal{E}\big(Y(t)\big) = \mathcal{E}\big(X_1(t)\big) + \mathcal{E}\big(\alpha X_2(t)\big) = m_1 + \mathcal{E}(\alpha)\mathcal{E}\big(X_2(t)\big) = m_1 + \frac{1}{2}m_2$

Time average is $\langle X_1(t) + \alpha X_2(t)\rangle = \langle X_1(t)\rangle + \langle \alpha X_2(t)\rangle = m_1 + \alpha m_2$

where as a result of the assumption on $X_1(t)$ and $X_2(t)$ being ergodic we have $\langle X_1(t)\rangle = m_1$, and $\langle X_2(t)\rangle = m_2$. As a result the process $Y(t)$ is not mean-ergodic because different realizations of $Y(t)$ will correspond to different values of the random variable α. For $\alpha = 0$ the time average is m_1, whereas for $\alpha = 1$, the time average is $m_1 + m_2$.

A Sufficient Condition for a Process to be Mean-Ergodic

A required condition for a process to be mean-ergodic is obviously that it has constant mean, i.e. $\mathcal{E}\big(X(t)\big) = m_X(t) = m$.

Theorem: A random process $X(t)$ is mean-ergodic if and only if the auto-covariance $C_X(t_1, t_2)$ satisfies the following:

$$\lim_{T\to\infty} \frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T} C_X(t,t')dtdt' = 0$$

Proof:

Let $m = \mathcal{E}\big(X(t)\big)$. We need to show that $\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} X(t)dt = m$, but we need a convergence criterion. The criterion that we use is mean-square convergence. Hence we need to show that

$$\lim_{T\to\infty} \mathcal{E}\left\{\left(\frac{1}{2T}\int_{-T}^{T} X(t)dt - m\right)^2\right\} = 0$$

$$\mathcal{E}\left\{\left(\frac{1}{2T}\int_{-T}^{T} X(t)dt - m\right)^2\right\} = \mathcal{E}\left\{\frac{1}{(2T)^2}\int_{-T}^{T}\int_{-T}^{T} X(t)X(t')dtdt' - \frac{2m}{2T}\int_{-T}^{T} X(t)dt + m^2\right\}$$

$$= \frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T} R_X(t,t')dtdt' - \frac{m}{T}\int_{-T}^{T} mdt + m^2$$

$$= \frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T} R_X(t,t')dtdt' - m^2$$

$$= \frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T} R_X(t,t')dtdt' - \frac{m^2}{4T^2}\int_{-T}^{T}\int_{-T}^{T} dtdt'$$

$$= \frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T} (R_X(t,t') - m^2)dtdt'$$

$$= \frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T} C_X(t,t')dtdt'$$

Taking the limit we have

$$\lim_{T\to\infty} \frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T} C_X(t,t')dtdt' = 0$$

Each of the above steps is basically an if and only if condition. Hence the theorem is proved.

Mean-Ergodic WSS Processes

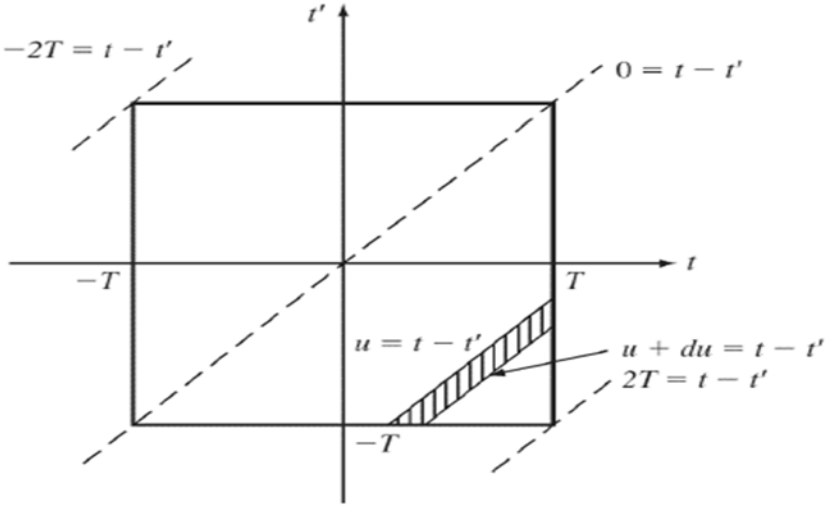
If the process $X(t)$ is WSS then the above requirement for ergodicity can be simplified by re-defining the function $C_X(\cdot)$ as follows:

$$\int_{-T}^{T}\int_{-T}^{T} C_X(t,t')dtdt' = \int_{-T}^{T}\int_{-T}^{T} C_X(t-t')dtdt'$$

We change variables in the double integral. Let $u = t - t'$, $v = t + t'$. Then integrate first over v and then over u to obtain

$$= \frac{1}{4T^2}\int_{-2T}^{2T}(2T-|u|)C_X(u)\,du = \frac{1}{2T}\int_{-2T}^{2T}\left(1-\frac{|u|}{2T}\right)C_X(u)\,du$$

See the Figure below



The requirement for mean-square ergodicity becomes

$$\lim_{T\to\infty}\frac{1}{2T}\int_{-2T}^{2T}\left(1-\frac{|u|}{2T}\right)C_X(u)\,du = 0$$

Example

Consider a process $X(t) = m + N(t)$, where $N(t)$ is a white noise process with zero mean and auto-correlation function $R_N(\tau) = N_0\delta(\tau)$. Is $X(t)$ mean ergodic?

Note that $C_X(\tau) = N_0\delta(\tau)$

$$\frac{1}{2T}\int_{-2T}^{2T}\left(1-\frac{|u|}{2T}\right)C_X(u)\,du = \frac{N_0}{2T}\int_{-2T}^{2T}\left(1-\frac{|u|}{2T}\right)\delta(u)du = \frac{N_0}{2T}$$

$$\lim_{T\to\infty}\frac{N_0}{2T}=0$$

Hence the process is mean ergodic.

Example

Consider the zero mean process $X(t)$ with auto-covariance function $C_X(\tau)=\beta e^{-2\alpha|\tau|}$. Is $X(t)$ ergodic?

Solution: Use the above theorem,

$$\frac{1}{2T}\int_{-2T}^{2T}\left(1-\frac{|u|}{2T}\right)C_X(u)du$$
$$\leq\frac{1}{2T}\int_{-2T}^{2T}\left|\left(1-\frac{|u|}{2T}\right)C_X(u)\right|du\leq\frac{1}{2T}\int_{-2T}^{2T}|C_X(u)|du=\frac{\beta}{2T}\int_{-2T}^{2T}e^{-2\alpha|\tau|}du$$

The above integral converges as $T\to\infty$, but then the factor $\frac{\beta}{2T}$ forces the result to zero as $T\to\infty$. Note that it is not necessary to evaluate the integral. We can just use any upper bound that we can show approaches a constant and then the whole result approaches zero.

As a result the process $X(t)$ is mean-ergodic.

Correlation Ergodic Processes

A WSS process $Y(t)$ is said to be correlation ergodic if the correlation function can be estimated from a single sample function. We can apply a test to determine correlation-ergodicity for $Y(t)$ as in the above, by setting $X(t)=Y(t)Y(t+\tau)$. We apply the test for each fixed value of τ.

The mean of $X(t)$ is $\mathcal{E}\big(Y(t)Y(t+\tau)\big)=R_Y(\tau)$. Hence for a fixed τ, the process $Y(t)$ is correlation ergodic if

$$\lim_{T\to\infty}\mathcal{E}\left\{\left(\frac{1}{2T}\int_{-T}^{T}Y(t)Y(t+\tau)dt-R_Y(\tau)\right)^2\right\}=0$$

Expanding the above we obtain

$$\mathcal{E}\left(\left(\frac{1}{2T}\int_{-T}^{T}Y(t)Y(t+\tau)dt\right)^2-2R_Y(\tau)\frac{1}{2T}\int_{-T}^{T}Y(t)Y(t+\tau)dt+R_Y^2(\tau)\right)$$
$$=\frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T}\mathcal{E}\big(Y(t)Y(t+\tau)Y(t')Y(t'+\tau)\big)dtdt'-\frac{R_Y(\tau)}{T}\int_{-T}^{T}R_Y(\tau)dt+R_Y^2(\tau)$$
$$=\frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T}\mathcal{E}\big(Y(t)Y(t+\tau)Y(t')Y(t'+\tau)\big)dtdt'-R_Y^2(\tau)$$

Hence the test for correlation-ergodicity is

$$\lim_{T\to\infty}\frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T}\left(\mathcal{E}\big(Y(t)Y(t+\tau)Y(t')Y(t'+\tau)\big)-R_Y^2(\tau)\right)dtdt'=0$$

This is a more difficult test than that for mean-ergodicity, but is easier to carry out for a Gaussian process.

Example (Mean-Ergodic)

Let $X(t)=A$ for all t be a random process, where A is a zero mean unit variance random variable. Use the above theorem to show that $X(t)$ is not an ergodic process.

$$\mathcal{E}\big(X(t)\big)=0$$

$$\mathcal{E}\big(X(t_1)X(t_2)\big)=\mathcal{E}(A^2)=1$$

Hence $C_X(t_1,t_2)=1$, and $X(t)$ is a WSS process. Now we apply the test

$$\frac{1}{2T}\int_{-2T}^{2T}\left(1-\frac{|u|}{2T}\right)C_X(u)du=\frac{1}{2T}\int_{-2T}^{2T}\left(1-\frac{|u|}{2T}\right)du=\frac{1}{T}\int_{0}^{2T}\left(1-\frac{|u|}{2T}\right)du=1$$

Now, take the limit as $T\to\infty$, and obviously the limit is 1 which is not 0. Hence the process is not mean-ergodic.

Discrete Time Ergodic Processes

For discrete time ergodic processes the time average is computed as

$$\langle X_n\rangle=\frac{1}{2N+1}\sum_{k=-N}^{N}X_k$$

and the correlation time average is (considering m constant and n as the time variable)

$$\langle X_{n+m}X_n\rangle=\frac{1}{2N+1}\sum_{k=-N}^{N}X_{k+m}X_k$$

The requirement for mean-ergodicity is that the mean is constant, i.e. $\mathcal{E}(X_n)=m$, as for the continuous time case, and then

$$\lim_{N\to\infty}\frac{1}{(2N+1)}\sum_{k=-2N}^{2N}\left(1-\frac{|k|}{2N+1}\right)C_X(k)=0$$

This can easily be shown with a similar procedure to the continuous case.

Example

Let X_n be a WSS discrete time process with mean m and covariance function

$C_X(k) = \sigma^2 \rho^{|k|}$ for $|\rho| < 1$, and $k = 0, \pm 1, \pm 2, \cdots$

(Note error in the textbook)

For the mean-ergodic test

$$\frac{1}{2N+1}\sum_{k=-2N}^{2N}\left(1-\frac{|k|}{2N+1}\right)\sigma^2\rho^{|k|} \le \frac{2}{2N+1}\sum_{k=0}^{2N}\sigma^2|\rho|^k \le \frac{2\sigma^2}{(2N+1)(1-|\rho|)}$$

Where for the last inequality we let $N \to \infty$, only in the sum to get a geometric series.

Taking the limit as $N \to \infty$, we obtain 0. Hence the process is mean-ergodic.

Example

Let $X_n = Z$, where Z is a Bernoulli random variable with $P(X_n = 1) = p$. Is X_n stationary? Is it ergodic?

Solution X_n is strict sense stationary since the probability distribution for a set of times does not depend on a time shift.

For the mean ergodicity we determine the covariance function

$$C_X(k) = \mathcal{E}\{(X_{n+k} - \mathcal{E}(X_{n+k}))(X_n - \mathcal{E}(X_n)) = \mathcal{E}(X_n^2) - (\mathcal{E}(X_n))^2 = p - p^2 = p(1-p)$$

Now determine

$$\frac{1}{2N+1}\sum_{k=-2N}^{2N}\left(1-\frac{|k|}{2N+1}\right)C_X(k)$$

$$= \frac{1}{2N+1}\sum_{k=-2N}^{2N}\left(1-\frac{|k|}{2N+1}\right)p(1-p)$$

$$= \frac{p(1-p)}{2N+1}\sum_{k=-2N}^{2N}\left(1-\frac{|k|}{2N+1}\right) \ge \frac{2p(1-p)}{2N+1}\sum_{k=1}^{2N}\left(1-\frac{k}{2N+1}\right)$$

$$= \frac{2p(1-p)}{2N+1}\left(2N - \frac{1}{2N+1}\sum_{k=1}^{2N} k\right) = \frac{2p(1-p)}{2N+1}\left(2N - \frac{1}{2N+1}\cdot\frac{2N(2N+1)}{2}\right)$$

$$= \frac{2p(1-p)}{2N+1}N = \frac{2p(1-p)}{2+1/N}$$

Taking the limit as $N \to \infty$, we obtain $p(1-p)$. Since the limit is not equal to zero, the process X_n is not mean ergodic.

Power Spectrum of a Random Process

Let $X(t)$ be a continuous-time WSS random process with mean m_X and auto-correlation function $R_X(\tau)$. Consider a truncated sample function $X_T(t) = X(t)$ for $|t| \leq T$, and zero elsewhere. We can take the Fourier Transform of $X_T(t)$ to obtain,

$$\tilde{X}_T(f) = \int_{-\infty}^{\infty} X_T(t)e^{-j2\pi f t}dt = \int_{-T}^{T} X(t)e^{-j2\pi f t}dt$$

For each sample function $X_T(t)$, $\left|\tilde{X}_T(f)\right|^2$ is the energy spectral density and we can define $\frac{1}{2T}\left|\tilde{X}_T(f)\right|^2$ as the power spectral density for this sample function over the interval $[-T, T]$.

Now, if we consider $X(t)$ as a process (i.e. a collection of sample functions), then $X_T(t)$ is also a process. It is a transformation of the process $X(t)$ (by truncation). Also we can consider the collection of functions $\tilde{X}_T(f)$ as a complex valued random process over the index set $f \in \mathcal{R}$. The second moment for this process is $\mathcal{E}\left\{\left|\tilde{X}_T(f)\right|^2\right\}$. The power spectral density of $X(t)$ is defined as the following

$$S_X(f) = \lim_{T\to\infty} \frac{1}{2T} \mathcal{E}\left\{\left|\tilde{X}_T(f)\right|^2\right\}$$

Wiener-Khinchin Theorem

The power spectral density of a WSS random process $X(t)$, $S_X(f)$, is equal to the Fourier Transform of the auto-correlation function of the process, $R_X(\tau)$, i.e.

$$S_X(f) = \int_{-\infty}^{\infty} R_X(\tau)e^{-j2\pi f\tau}d\tau$$

We do not prove this theorem here, but we can give an idea of why the theorem holds. First we consider the WSS processes that are mean-ergodic and correlation-ergodic. These processes require that the auto-covariance function be in a sense “narrow” (as we saw in the tests for ergodicity). This means that a Fourier Transform (FT) of $R_X(\tau)$ exists. Then we can use the ergodic property to state that

$$R_X(\tau) \approx \frac{1}{2T}\int_{-T}^{T} X_T(t+\tau)X_T(t)dt$$

For large T.

This is like a convolution (except for a positive sign, i.e. for a convolution we would have $X_T(-t + \tau)$ instead of $X_T(t+\tau)$ in the integral. We then recall that in Fourier analysis the FT of a convolution of two signals is the product of the FT's of the two signals. The FT of $R_X(\tau)$ would then be $\tilde{X}_T(f)\tilde{X}_T^*(f) = \left|\tilde{X}_T(f)\right|^2$. Note that for the regular convolution the FT would be $\tilde{X}_T^2(f)$. So we can see that the result of the Wiener-Khinchin theorem is not a surprise, although we have not stated the theorem in exact terms and have not given a rigorous mathematical proof.

Real Valued Random Processes

All the random processes that we have discussed so far are real-valued processes. However, it is possible to define complex valued processes, and these are very useful in signal processing applications.

For real-valued processes the auto-correlation function, as we have seen before, is an even function, i.e. $R_X(\tau) = R_X(-\tau)$. In this case we can see that the power spectral density will become

$$S_X(f) = \int_{-\infty}^{\infty} R_X(\tau)e^{-j2\pi f\tau}d\tau = \int_{-\infty}^{\infty} R_X(\tau)(\cos(2\pi f\tau) - j\sin(2\pi f\tau))d\tau$$
$$= \int_{-\infty}^{\infty} R_X(\tau)\cos(2\pi f\tau)d\tau$$

where we have used the fact that $R_X(\cdot)$ is an even function, and $\sin(\cdot)$ is an odd function. Then the product of an even function and an odd function is an odd function, and the integral of an odd function is 0.

Properties of Power Spectral Density

The power spectral density is a positive function, i.e. $S_X(f) \geq 0$ for all f. This is due to the fact that it is the expectation of $|X_T(f)|^2$.

The auto-correlation function can be recovered from the power spectral density by the inverse FT,

$$R_X(\tau) = \int_{-\infty}^{\infty} S_X(f)e^{j2\pi f\tau}d\tau$$

We refer to the second moment of the random process $X(t)$ as the average power of the process.

$$P = \mathcal{E}\big(X^2(t)\big) = R_X(0) = \int_{-\infty}^{\infty} S_X(f)df$$

It is for this reason that $S_X(f)$ is called the Power Spectral Density (PSD), i.e. its integral yields the average power of the process.

We have shown that the auto-correlation function and the auto-covariance functions are related as $R_X(\tau) = C_X(\tau) + m_X^2$, where m_X is the mean of the process.

Taking FT's we obtain $S_X(f) = \mathrm{F}\big(C_X(\tau)\big) + m_X^2\delta(f)$. In other words the PSD contains a delta function which results from the mean of the process. We also refer to this component as the d.c. component.

Cross-Power Spectral Density

Previously we also defined the cross-correlation function between two processes $X(t)$ and $Y(t)$, as $R_{XY}(\tau) = \mathcal{E}\big(X(t+\tau)Y(t)\big)$. Correspondingly we define the cross-Power Spectral Density function as the Fourier Transform of the cross-correlation function:

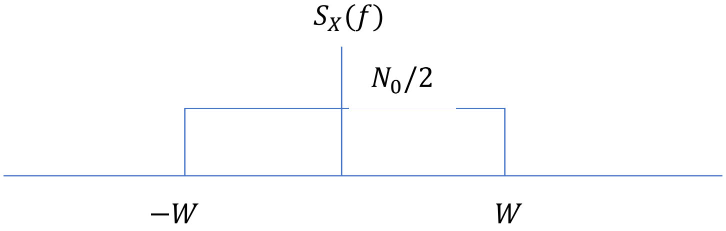
$$S_{XY}(f) = \mathrm{FT}\{R_{XY}(\tau)\}$$

Note that in general $S_{XY}(f)$ is a complex-valued function because in this case $R_{XY}(\cdot)$ is not an even function.

Example

A WSS random process has power spectral density given as follows:

$$S_X(f) = \begin{cases} \frac{N_0}{2} & \text{for } \ |f| \leq W \\ 0 & \text{elsewhere} \end{cases}$$



Find the average power and the auto-correlation function.

The auto-correlation function is the inverse FT

$$R_X(\tau) = \int_{-W}^{W} \frac{N_0}{2} e^{j2\pi f\tau} df = N_0 \frac{\sin 2\pi W\tau}{2\pi\tau} = N_0 W \ \mathrm{sinc}(2W\tau)$$

where the function $\mathrm{sinc}(\cdot)$ is defined as $\mathrm{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$.

Note that the average power can be determined in two ways, either the integral of $S_X(f)$, i.e. area under the curve, or the value of the auto-correlation function at 0, i.e. $R_X(0)$. The result is $N_0 W.$

As $W \to \infty$ in the above process, the average power approaches ∞, and the auto-correlation function approaches $\frac{N_0}{2}\delta(\tau)$. One may ask why we denote the power spectral density for this so-called white noise as $N_0/2$ and not simply N_0? The reason for this is that in applications engineers frequently work with what we call the one-sided power spectral density, i.e. the PSD over only positive frequencies. In such a case the one-sided PSD would be N_0 and the bandwidth of the noise is W. Hence the average power is $N_0 W$. In the case of using the two-sided PSD the (two-sided) bandwidth of the noise is $2W$. Hence the average power is $(N_0/2)(2W) = N_0 W$.

Power Spectral Density for Discrete Time Processes

For a discrete time process X_n, the auto-correlation function is $R_X(k)$, $k = 0, \pm 1, \cdots$ Discrete time processes often arise as sampled continuous time processes. In this sense the Fourier Transform of such sampled signals becomes a periodic signal of frequency. The sampled signal corresponds to the Fourier Series coefficients of a continuous domain signal. The power spectral density is then obtained from the auto-correlation function in a version of the Wiener-Khnichin theorem as follows:

$$S_X(f) = \text{FT}\big(R_X(k)\big) = \sum_{k=-\infty}^{\infty} R_X(k)\, e^{-j2\pi kf}$$

where $-\frac{1}{2} < f \leq \frac{1}{2}$

The inverse FT is $R_X(k) = \int_{-\frac{1}{2}}^{\frac{1}{2}} S_X(f) e^{j2\pi kf} df$

Example

Let X_n be a sequence of uncorrelated random variables with zero mean and variance σ_X^2.

Find $S_X(f)$.

The autocorrelation function is $R_X(k) = \sigma_X^2 \delta_k$, where $\delta_k = 1$ for $k = 0$, and 0 otherwise. This is **discrete time white noise**.

$$S_X(f) = \sum_{k=-\infty}^{\infty} R_X(k) e^{j2\pi kf} = R_X(0) = \sigma_X^2$$

Example

Let Y_n be a discrete time process where $Y_n = X_n + \alpha X_{n-1}$, and X_n is discrete white noise as in the previous example, and α is a constant. Find $\mathcal{E}(Y_n)$, $R_Y(k)$, and $S_Y(f)$.

$$\mathcal{E}(Y_n) = \mathcal{E}(X_n) + \alpha\mathcal{E}(X_{n-1}) = 0$$

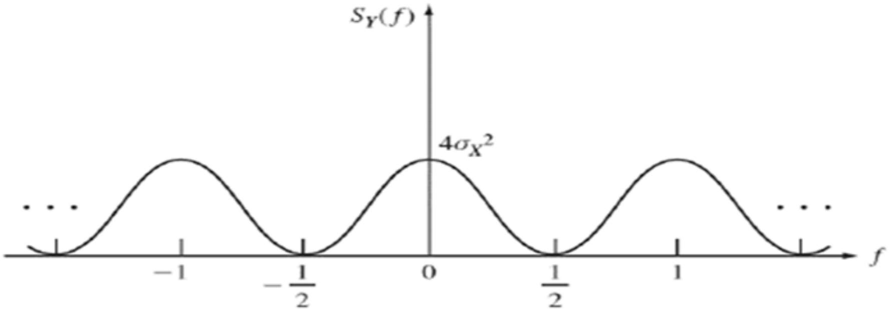
$$\begin{aligned} R_Y(k) = \mathcal{E}(Y_{n+k}Y_n) &= \mathcal{E}\big((X_{n+k} + \alpha X_{n+k-1})(X_n + \alpha X_{n-1})\big) \\ &= \sigma_X^2\delta_k + \alpha\sigma_X^2\delta_{k+1} + \alpha\sigma_X^2\delta_{k-1} + \alpha^2\sigma_X^2\delta_k \end{aligned}$$

$$R_Y(k) = \sigma_X^2(1 + \alpha^2)\delta_k + \alpha\sigma_X^2(\delta_{k-1} + \delta_{k+1})$$

The power spectral density is

$$\begin{aligned} S_Y(f) &= \sigma_X^2(1 + \alpha^2) + \alpha\sigma_X^2\big(e^{j2\pi f} + e^{-j2\pi f}\big) \\ &= \sigma_X^2[1 + \alpha^2 + 2\alpha\cos(2\pi f)] \end{aligned}$$

The Figure below shows the plot for $\alpha = 1$.



Linear Time Invariant Systems

Consider a system where an input signal, $x(t)$, is mapped to the output signal, $y(t)$, by the transformation

$$y(t) = T(x(t))$$

- The system is *linear* if the following property holds (superposition);

$$T\big(\alpha x_1(t) + \beta x_2(t)\big) = \alpha T\big(x_1(t)\big) + \beta T(x_2(t))$$

where $x_1(t)$ and $x_2(t)$ are arbitrary signals and α and β are arbitrary constants.

- The system is said to be *time-invariant* if $T\big(x(t-\tau)\big) = y(t-\tau)$, i.e.
$$\Big(T\big(x(t)\big) = y(t)\Big) \Rightarrow \Big(T\big(x(t-\tau)\big) = y(t-\tau)\Big)$$
for any input $x(t)$.
- The impulse response of a linear time invariant system is defined by

$$h(t) = T(\delta(t))$$

- The response (i.e. output) of the system to an arbitrary input $x(t)$ is given by the convolution

$$y(t) = (h * x)(t) = \int_{-\infty}^{\infty} h(s)x(t-s)ds = \int_{-\infty}^{\infty} h(t-s)x(s)ds$$

- We also define the Transfer Function for the system

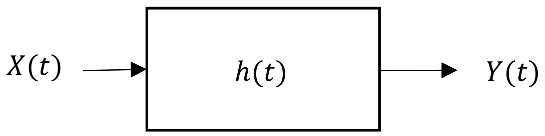
$$H(f) = \mathrm{FT}\big(h(t)\big) = \int_{-\infty}^{\infty} h(t)e^{-j2\pi ft}dt$$

- A system is said to be causal if the impulse response $h(t)$ satisfies $h(t) = 0$ for $t < 0$. This means that the output at time t_0 depends only on input values for time $t < t_0$. In this case the output (convolution integral) can be written as

$$y(t) = \int_{-\infty}^{t} h(t-s)x(s)ds$$

Random Process Inputs to Linear Time Invariant Systems (LTI)

Consider a random process $X(t)$ that is input to an LTI system with impulse response $h(t)$. The output is given by the convolution integral, as above, where the integral is defined in the mean-square sense as we have discussed before. We assume that the integral is defined as a Riemann integral and the mean-square criterion is a statement on which processes result in the convergence of the integral, i.e. for which processes is the integral defined. We represent the linear system as in the following Figure, where $X(t)$ and $Y(t)$ are random processes, and $h(t)$ is the impulse response of the system.



Mean and Auto-Correlation of $Y(t)$

Consider a WSS process $X(t)$, and the output is the process $Y(t)$. The mean of the output process is determined as

$$\mathcal{E}\big(Y(t)\big) = \mathcal{E}\left(\int_{-\infty}^{\infty} h(s)X(t-s)ds\right) = \int_{-\infty}^{\infty} h(s)\mathcal{E}\big(X(t-s)\big)ds = \int_{-\infty}^{\infty} h(s)m_X ds = H(0)m_X$$

Where $H(0)$ is the transfer function value at 0, and m_X is the mean of $X(t)$.

The auto-correlation is computed as follows:

$$\mathcal{E}(Y(t+\tau)Y(t)) = \mathcal{E}\left(\int_{-\infty}^{\infty} h(s)X(t+\tau-s)ds \int_{-\infty}^{\infty} h(r)X(t-r)dr\right)$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(s)h(r)\mathcal{E}\big(X(t+\tau-s)X(t-r)\big)dsdr$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(s)h(r)R_X(\tau-s+r)dsdr$$

Since the mean of $Y(t)$ is constant and the auto-correlation depends only on the time difference τ, the process $Y(t)$ is WSS.

The above can also be written as

$$R_Y(\tau) = \mathcal{E}\big(Y(t)Y(t+\tau)\big)$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(s)h(r)R_X(\tau-s+r)dsdr$$

$$= \int_{-\infty}^{\infty} h(r)\left(\int_{-\infty}^{\infty} h(s)R_X(t+r-s)ds\right)dr$$

$$= \int_{-\infty}^{\infty} h(r)(h * R_X)(t + r)dr$$

$$R_Y(\tau) = (h_m * h * R_X)(\tau)$$

where $h_m(r) = h(-r)$.

Power Spectral Density of $\boldsymbol{Y(t)}$

The power spectral density of the output $Y(t)$ is

$$S_Y(f) = \int_{-\infty}^{\infty} R_Y(\tau)e^{-j2\pi f\tau}d\tau = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(s)h(r)R_X(\tau - s + r)dsdr\right)e^{-j2\pi f\tau}d\tau$$

Letting $u = \tau - s + r$, i.e. $\tau = u + s - r$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(s)h(r)R_X(u)\, e^{-j2\pi f(u+s-r)}dudsdr$$

$$= \int_{-\infty}^{\infty} h(s)e^{-j2\pi fs}ds \int_{-\infty}^{\infty} h(r)e^{j2\pi fr}dr \int_{-\infty}^{\infty} R_X(u)e^{-j2\pi fu}du$$

$$= H(f)H^*(f)S_X(f)$$

$$S_Y(f) = |H(f)|^2 S_X(f)$$

Example

Find the output of an LTI system with impulse response $h(t)$ if the input is white noise.

Solution: $S_X(f) = \frac{N_0}{2}$

$$S_Y(f) = |H(f)|^2 \frac{N_0}{2}$$

If $H(f)$ is an ideal low-pass filter with bandwidth W, then $Y(t)$ is called bandlimited white noise.

Example – Ornstein-Uhlenbeck Process

Consider an LTI system with impulse response $h(t) = e^{-\alpha t}u(t)$, where $\alpha > 0,$ and $u(t)$ is the step function. Note that this is the impulse response of a first order low-pass filter. It is implemented as the input being connected to a resistor in series with a capacitor and the output taken across the capacitor. The transfer function is

$$H(f) = \frac{1}{\alpha + j2\pi f}$$

Assume that the input is white noise with PSD $S_X(f) = \frac{N_0}{2}$

The output is $S_Y(f) = \frac{N_0}{2}|H(f)|^2 = \frac{N_0}{2}\frac{1}{\alpha^2+(2\pi f)^2}$

The inverse FT can now be taken

$$\frac{N_0}{2}\frac{1}{\alpha^2+(2\pi f)^2} = \frac{N_0}{4\alpha}\left(\frac{1}{\alpha+j2\pi f}+\frac{1}{\alpha-j2\pi f}\right)$$

For the first term in the second factor the inverse FT is

$$e^{-\alpha t}u(t)$$

For the second term, the inverse FT is

$$e^{\alpha t}u(-t)$$

These two terms can be combined to yield the expression

$$e^{-\alpha|t|}$$

Hence we have

$$R_Y(\tau) = \frac{N_0}{4\alpha}e^{-\alpha|\tau|}$$

Discrete Time Systems

For discrete time Linear Time Invariant systems, the equivalent of the impulse response is the response of the system to the unit input signal

$$\delta_n = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 \end{cases}$$

The unit response that characterizes the system, i.e. the response to δ_n, is h_n.

Then the response of the system to an arbitrary input X_n is then

$$Y_n = X_n * h_n = \sum_{k=-\infty}^{\infty} h_k X_{n-k} = \sum_{k=-\infty}^{\infty} h_{n-k} X_k$$

The transfer function is

$$H(f) = \sum_{k=-\infty}^{\infty} h_k e^{-j2\pi fk}$$

If X_n is a discrete time WSS process then Y_n is also a discrete time WSS process.

$$\mathcal{E}(Y_n) = \mathcal{E}\left(\sum_{k=-\infty}^{\infty} h_k X_{n-k}\right) = \sum_{k=-\infty}^{\infty} h_k \mathcal{E}(X_{n-k}) = m_X \sum_{k=-\infty}^{\infty} h_k = m_X H(0)$$

The auto-correlation function is

$$R_Y(k) = \mathcal{E}\left(\sum_{i=-\infty}^{\infty} h_i X_{n+k-i} \sum_{j=-\infty}^{\infty} h_j X_{n-j}\right)$$

$$= \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} h_i h_j\, \mathcal{E}\big(X_{n+k-i}X_{n-j}\big)$$

$$= \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} h_i h_j R_X(k+j-i)$$

This can also be written as $R_Y(k) = (h * h_m * R_X)$

where * denotes discrete convolution and h_m is the time reversal of h, i.e. $h_m(i) = h(-i)$.

The power spectral density of the discrete time Fourier Transform.

We can show that it is given by

$$S_Y(f) = |H(f)|^2 S_X(f)$$

This is a similar result to the continuous time case.

Note that if X_n is a Gaussian process then Y_n is also a Gaussian process. This is due to the linear transformation of X_n.

Sampling of Bandlimited Random Processes

For a deterministic signal $x(t)$ with the Fourier Transform equal to 0 for $|f| \geq W$, we can sample the signal at the rate of $f_s = 1/T$ and obtain the sequence of samples, $\{\cdots, x(-2T), x(-T), x(0), x(T), x(2T), \cdots\}$. Then $x(t)$ can be recovered exactly from the samples if $f_s \geq 2W$. The minimum sampling rate is referred to as the Nyquist sampling rate.

The original signal can be reconstructed from the samples using the following interpolation formula:

$$x(t) = \sum_{k=-\infty}^{\infty} x(kT)p(t-kT)$$

Where $p(t) = \sin\frac{(\pi t/T)}{(\pi t/T)}$. Note that $p(t)$ is the impulse response for the ideal low-pass filter with bandwidth W, i.e.

$$P(f) = \mathrm{FT}\big(p(t)\big) = \begin{cases} 1 & |f| \leq W \\ 0 & |f| > W \end{cases}$$

The reconstruction process corresponds to inputting the signal

$$\sum_{k=-\infty}^{\infty} x(kT)\delta(t-kT)$$

to the above ideal low-pass filter, yielding the reconstructed signal at the output.

Sampling a WSS random process $\boldsymbol{X(t)}$ is done in a similar manner to that for a deterministic signal as the $x(t)$ above. Assume that the power spectral density $S_X(f)$ is 0 for $|f| > W$. We sample the process $X(t)$ with a sampling frequency $f_s > 2W$, i.e. $T < \frac{1}{2W}$.

The reconstructed process would then be

$$\hat{X}(t) = \sum_{k=-\infty}^{\infty} X(kT)p(t-kT)$$

We consider the mean square criterion for convergence, i.e. $\mathcal{E}\left\{\left(X(t)-\hat{X}(t)\right)^2\right\} = 0$.

Evaluating we obtain

$$\mathcal{E}\left\{X^2(t) - 2X(t)\sum_{k=-\infty}^{\infty} X(kT)p(t-kT) + \sum_{k=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} X(kT)X(jT)p(t-kT)p(t-jT)\right\}$$

$$= R_X(0) - 2\sum_{k=-\infty}^{\infty} R_X(t-kT)p(t-kT) + \sum_{k=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} R_X((k-j)T)p(t-kT)p(t-jT)$$

We can show that as a result of $R_X(\tau)$ being a band-limited signal, the above is equal to zero. This is clear for $t = 0$ since, in this case clearly the sum in the middle term is $R_X(0)$, because it is zero for all terms except $k = 0$, and the double sum is also $R_X(0)$ because it is zero for all terms except $k = 0$, and $j = 0$. But it also holds for all t, i.e. the reconstructed process $\hat{X}(t)$ is equal to the process $X(t)$ in the mean square sense.

Optimum Linear Prediction

We consider a scenario where a random variable X is correlated with a random variable Y. We observe Y and would like to estimate a value for X. Let the estimated value be $\tilde{X}$. We consider a linear estimator, i.e. the estimate of X is a linear function of Y. Let the estimate be $\tilde{X} = \alpha Y$, where α is a constant. The estimation error is $e = X - \tilde{X}$. We will choose the estimate so that the mean square vale of the error, e, is minimized.

$$\mathcal{E}(e^2) = \mathcal{E}\left(X-\tilde{X}\right)^2 = \mathcal{E}(X-\alpha Y)^2 = \mathcal{E}(X^2 - 2\alpha XY + \alpha^2 Y^2) = \sigma_X^2 - 2\alpha\sigma_{XY} + \alpha^2\sigma_Y^2$$

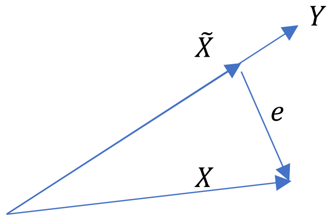
To minimize we differentiate with respect to α and set to 0

$$-2\sigma_{XY} + 2\alpha\sigma_Y^2 = 0$$

$$\alpha = \frac{\sigma_{XY}}{\sigma_Y^2}$$

Hence the estimate is $\tilde{X} = \frac{\sigma_{XY}}{\sigma_Y^2} Y$.

Now we may treat a random variable as a vector as we have discussed before. Then we may define a dot product between two random variables X, Y as $< X, Y >= \mathcal{E}(XY) = \int_\Omega XY dP$. The above then becomes $\tilde{X} = \frac{<X,Y>}{||Y||^2} Y$ where $||Y||^2 = \mathcal{E}(Y^2)$, may be thought of as the length of the "vector" Y squared. With the above vector interpretation we have the following Figure



In the Figure we see the vectors (random variables) corresponding to X, Y, the original two random variables, $\tilde{X}$, the estimate of X from the observed Y, and e the error vector. We will no show that e is orthogonal to Y.

$$< e, Y > = < X - \tilde{X}, Y > \ = \ < X, Y > - < \tilde{X}, Y > = < X, Y > - < \alpha Y, Y >$$

$$= < X, Y > -\alpha < Y, Y > = < X, Y > - \frac{< X, Y >}{||Y||^2} < Y, Y > = 0$$

where we have substituted for α. This means that using the mean square error criterion results in the error in the estimate being orthogonal to the observed random variable Y.

Next we modify the above so that the random variables X, Y are vectors. We observe Y and calculate an estimate for X, $\tilde{X}$. We assume a linear estimate $\tilde{X} = AY$, where X, Y are assumed to be n vectors and A is an $n \times n$ matrix. Following the same procedure as before we have

$$e = X - \tilde{X} = X - AY$$

The mean squared error is a function of the matrix A, $J(A) = \mathcal{E}\left(||e||^2\right) = \mathcal{E}(e^T e) = \mathcal{E}((X - AY)^T (X - AY))$, where we have assumed that the vectors are column vectors.

$$J(A) = \mathcal{E}(X^T X - X^T AY - (AY)^T X + (AY)^T AY) = \mathcal{E}(X^T X - X^T AY - Y^T A^T X + Y^T A^T AY)$$

Note that since $Y^T A^T X$ is a scalar then it is equal to its transpose which is $X^T AY$

Hence the above becomes $J(A) = \mathcal{E}(X^T X - 2X^T AY + Y^T A^T AY)$

Now we differentiate (partial derivatives) with each of the n^2 components of A and set each equation to 0. Formally we can differentiate $J(A)$ with respect to the matrix A and interchange the expectation and partial derivative operations as follows

$$\frac{\partial J(A)}{\partial A} = \mathcal{E}\left(\frac{\partial(X^T X)}{\partial A} - \frac{2\partial(X^T AY)}{\partial A} + \frac{\partial(Y^T A^T AY)}{\partial A}\right) = 0$$

$$\mathcal{E}(-2XY^T + 2AYY^T) = 0$$

$$-2R_{XY} + 2AR_Y = 0$$

$$A = R_{XY}R_Y^{-1}$$

Note that we have used the following $\frac{2\partial(X^T AY)}{\partial A} = XY^T$ and $\frac{\partial(Y^T A^T AY)}{\partial A} = AYY^T$. Nore that $X^T AY$ is a function $g(A)$ of n^2 variables which are the elements of the matrix A. We form n^2 partial derivatives, i.e. $\frac{\partial g(A)}{\partial a_{ij}}$ and write the resulting set of derivatives in a matrix form. This is what we mean by $\frac{\partial g(A)}{\partial A}$.

The equation to estimate X in terms of Y is then $\tilde{X} = R_{XY}R_Y^{-1}Y$. This is analogous to the above result for the case of scalar random variables, i.e. $\tilde{X} = \frac{\sigma_{XY}}{\sigma_Y^2}Y$.

We have considered estimation in the case of a scalar random variable and then the case of an n −vector. Now we can consider the case where the vector X has an infinite number of components, i.e. it is a discrete time random process.

To estimate X we will do it component by component. For linear estimation the equation is $\tilde{X} = AY$, where now the vectors are infinite dimensional and the matrix is also infinite dimensional.

The n^{th} component for $\tilde{X}$ is $\tilde{x}_n = \sum_{k=-\infty}^{\infty} a_{nk} y_k$. We refer to this operation as a generalized linear filter. Note that the filters that we are used to in signal processing basically are a special case of this where the rows of the matrix A are cyclic shifts of each other and the above sum is a discrete convolution $\tilde{x}_n = \sum_{k=-\infty}^{\infty} a_{n-k} y_k$. In this case $a_n, n = 0, \pm 1, \pm 2, \cdots$ is the impulse response. If the filter is causal then the sum reduces to

$$\tilde{x}_n = \sum_{k=-\infty}^{n} a_{n-k} y_k$$

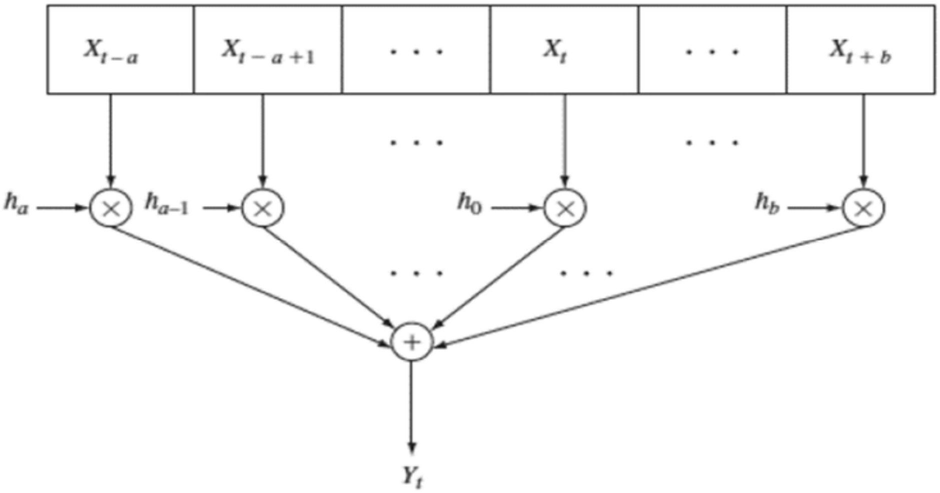
In the following we will consider various case of this type of filtering

Optimum Linear Filters

We consider a scenario where we observe a discrete-time zero-mean process X_n over a certain time interval $I = \{t - a, t - a + 1, \cdots, t + b\}$ and we are required to use the $a + b + 1$, resulting observations, i.e. $\{X_{t-a}, \cdots, X_t, \cdots X_{t+b}\}$ to obtain an estimate for some other (usually related) zero-mean process Z_t.

The estimator is a linear filter:

$$Y_t = \sum_{k=t-a}^{t+b} h_{t-k} X_k = \sum_{k=-b}^{a} h_k X_{t-k}$$

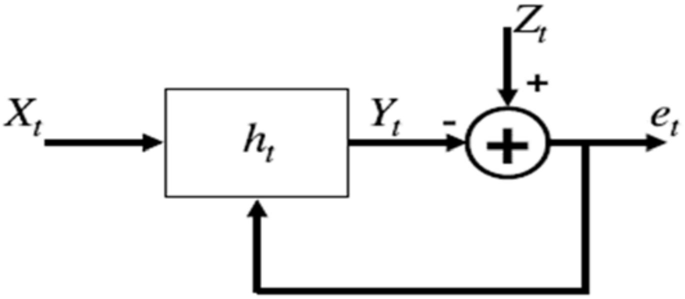


One example is the case where the observed signal is $X_n = Z_n + W_n$, where W_n is some noise signal. In this case the wanted signal Z_n is corrupted by noise W_n and the goals is to do some filtering in the observed signal X_n to obtain a good estimate for Z_n.

Figure of Merit

The figure of merit for the estimate is the mean-square error. Let the error for the signal at time t (note that t is an integer) is $e_t = Z_t - Y_t$. Hence we wish to compute a value Y_t so that $\mathcal{E}(e_t^2) =$

$\mathcal{E}\{(Z_t - Y_t)^2\}$ is minimized. Since $Y_t = h_t * X_t$. This means that we need to find the filter h_t that minimizes $\mathcal{E}(e_t^2)$.

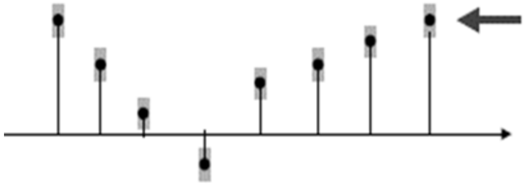


Filtering, Smoothing, and Prediction

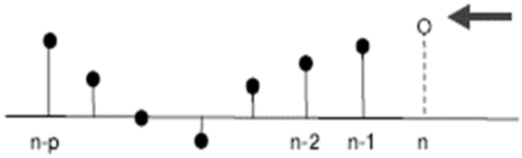
Three types of estimation problems: 1) filtering, 2) smoothing, 3) prediction

$$Y_t = \sum_{n=t-a}^{t+b} h_{t-n} X_n = \sum_{n=-b}^{a} h_n X_{t-n}$$

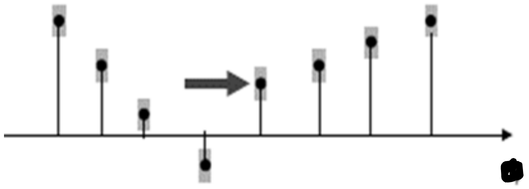
Filtering: $b = 0$. We have observations from the past and up to the current time t, denoted as X_t, and we are interested in estimating Z_t, at the current time. The idea is that there is a correlation between X_t and Z_t.



Prediction: $b < 0$. We have observations of the process Z_t in the past (not current time), i.e. $\{Z_r$, for $r < t\}$, and we are interested in estimating Z_s for $s \geq t$. There is a single process here and we are interested in predicting the future values in terms of past values.



Smoothing: $b > 0$. We have observations of the past and some future values, X_t, and are interested in estimating Z_t at the current time. This is similar to the Filtering case but also includes future observation values.



Orthogonality Principle

Consider the estimation value

$$Y_t = \sum_{n=-b}^{a} h_n X_{t-n}$$

and the error $e_t = Z_t - Y_t$.

We would like to minimize the mean square error by varying the filter coefficients h_n to determine the filter that yields the minimum mean-square error:

Consider the mean-square error as a function of $\boldsymbol{h} = (h_{-b}, \cdots, h_a)$. Differentiate with respect to the h_k and set to zero to find the vector $\boldsymbol{h}$ that yields the minimum mean-square error.

$$\frac{\partial}{\partial h_m}\mathcal{E}(e_t^2) = \mathcal{E}\left(\frac{\partial}{\partial h_m}e_t^2\right) = \mathcal{E}\left(2e_t\frac{\partial}{\partial h_m}e_t\right) = -2\mathcal{E}\left(e_t\left(\frac{\partial}{\partial h_m}Y_t\right)\right) = -2\mathcal{E}(e_tX_{t-m}) = 0$$

Hence the requirement to minimize the mean-square error is

$\mathcal{E}(e_tX_{t-m}) = 0$ for $-b \le m \le a$ ($a + b + 1$ equations)

The error is orthogonal to the observations. This is known as the orthogonality principle.

The optimum filter must satisfy these conditions. Note that the filter h_n impacts the error e_t.

To solve for the optimum filter we substitute for the error e_t:

$$\mathcal{E}(e_tX_{t-m}) = \mathcal{E}\big((Z_t - Y_t)X_{t-m}\big) = 0$$

$$\mathcal{E}(Z_tX_{t-m}) = \mathcal{E}(Y_tX_{t-m})$$

$$R_{ZX}(m) = \mathcal{E}\left(\left(\sum_{n=-b}^{a} h_nX_{t-n}\right)X_{t-m}\right) = \sum_{n=-b}^{a} h_n\mathcal{E}(X_{t-n}X_{t-m})$$

$$R_{ZX}(m) = \sum_{n=-b}^{a} h_nR_X(m-n) \qquad -b \le m \le a$$

Note that this is a set of $a + b + 1$ linear equations. The idea is that $R_{ZX}(m)$, and $R_X(k)$ are known and the above is a set of equations in the $a + b + 1$ unknowns h_k, for which we solve.

$$R_{ZX}(-b) = R_X(0)h_{-b} + R_X(-1)h_{-b+1} + \cdots + R_X(-b-a)h_a$$

$$R_{ZX}(-b+1) = R_X(1)h_{-b} + R_X(0)h_{-b+1} + \cdots + R_X(-b-a+1)h_a$$

$$\vdots$$

$$R_{ZX}(a) = R_X(a+b)h_{-b} + R_X(a+b-1)h_{-b+1} + \cdots + R_X(0)h_a$$

The above can be written in matrix form as follows

$$\begin{bmatrix} R_{ZX}(-b) \\ R_{ZX}(-b+1) \\ \vdots \\ R_{ZX}(a) \end{bmatrix} = \begin{bmatrix} R_X(0) & R_X(-1) & \cdots & R_X(-b-a) \\ R_X(1) & R_X(0) & \cdots & R_X(-b-a+1) \\ \vdots & \vdots & \ddots & \vdots \\ R_X(a+b) & R_X(a+b-1) & \cdots & R_X(0) \end{bmatrix} \begin{bmatrix} h_{-b} \\ h_{-b+1} \\ \vdots \\ h_a \end{bmatrix}$$

In matrix notation we have

$$\mathbf{R}_{ZX} = \mathbf{R}_X \mathbf{h}$$

Note that for real random processes the auto-correlation function is even, and the above matrix is symmetric.

A solution can be obtained by matrix inversion

$$\mathbf{h} = \mathbf{R}_X^{-1} \mathbf{R}_{ZX}$$

The mean square error corresponding to this choice for **h** can now be obtained. As in any optimization problem, first we find the argument of a function that minimizes the function and then we evaluate the minimum value achieved.

Mean Squared Error

Note first that as a result of the solution for the optimum **h**

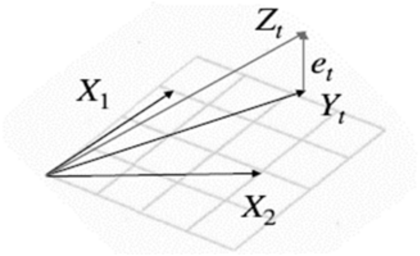
$$\mathcal{E}(e_t Y_t) = \mathcal{E}\left(e_t \sum_{n=-b}^{a} h_n X_{t-n}\right) = \sum_{n=-b}^{a} h_n \mathcal{E}(e_t X_{t-n}) = 0$$

Hence $\mathcal{E}(e_t^2) = \mathcal{E}\big(e_t(Z_t - Y_t)\big) = \mathcal{E}(e_t Z_t)$

$$\mathcal{E}(e_t^2) = \mathcal{E}\big(Z_t(Z_t - Y_t)\big) = \mathcal{E}(Z_t^2) - \mathcal{E}\left(Z_t \sum_{n=-b}^{a} h_n X_{t-n}\right)$$

$$= R_Z(0) - \sum_{n=-b}^{a} h_n \mathcal{E}(Z_t X_{t-n})$$

$$\boxed{\mathcal{E}(e_t^2) = R_Z(0) - \sum_{n=-b}^{a} h_n R_{ZX}(n)}$$



Wiener-Hopf Theorem (Discrete Time)

Let X_t and Z_t be discrete time zero mean jointly WSS random processes and let Y_t be an estimate of Z_t of the form

$$Y_t = \sum_{n=t-a}^{t+b} h_{t-n}X_n = \sum_{k=-b}^{a} h_k X_{t-k}$$

The filter, $\mathbf{h} = (h_{-b}, \cdots, h_a)$, that minimizes $\mathcal{E}\{(Z_t - Y_t)^2\}$ satisfies

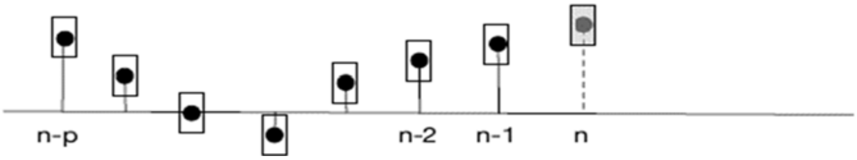
$$R_{ZX}(m) = \sum_{n=-b}^{a} h_n R_X(m-n) \qquad -b \le m \le a$$

and the mean square error satisfies

$$\mathcal{E}\{(Z_t - Y_t)^2\} = R_Z(0) - \sum_{n=-b}^{a} h_n R_{ZX}(n)$$

Example 1

Suppose that we are interested in estimating the signal Z_n from the most recent $p+1$ observations (including the present observation): $X_k = Z_k + N_k$, $k \in I = \{n-p, \cdots, n-1, n\}$. Find the set of linear equations for the optimum filter if Z_k and N_k are independent random processes.



Solution:

For this choice of observation interval we have $a = p$, and $b = 0$.

$$R_{ZX}(m) = \sum_{k=0}^{p} h_k R_X(m-k) \qquad m = 0,1,\cdots,p$$

The cross-correlation terms are given by

$$R_{ZX}(m) = \mathcal{E}\big(Z_{k+m}(Z_k + N_k)\big) = \mathcal{E}(Z_k Z_{k+m}) + \mathcal{E}(Z_{k+m}N_k) = R_Z(m) + \mathcal{E}(Z_{k+m})\mathcal{E}(N_k)$$

$$R_{ZX}(m) = R_Z(m)$$

The auto-correlation terms are given by

$$R_X(m-k) = \mathcal{E}(X_{n-k}X_{n-m}) = \mathcal{E}\big((Z_{n-k} + N_{n-k})(Z_{n-m} + N_{n-m})\big)$$

$$= \mathcal{E}(Z_{n-k}Z_{n-m} + Z_{n-k}N_{n-m} + N_{n-k}Z_{n-m} + N_{n-k}N_{n-m}) = \mathcal{E}(Z_{n-k}Z_{n-m}) + \mathcal{E}(N_{n-k}N_{n-m})$$

$$R_{ZX}(m-k) = R_Z(m-k) + R_N(m-k)$$

The optimum filter satisfies

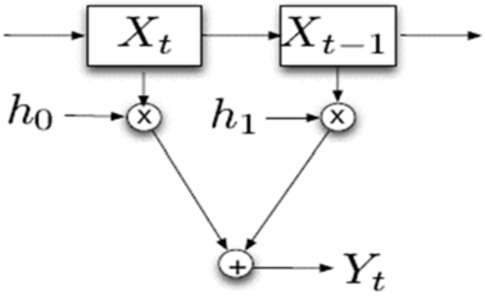
$$R_Z(m) = \sum_{k=0}^{p} h_k\,[R_Z(m-k) + R_N(m-k)] \qquad m = 0, \cdots, p$$

This set of linear equations is solved by matrix inversion (see above) to obtain the vector $\mathbf{h} = (h_0, \cdots, h_p)$. Note that we need to have the auto-correlation functions for the processes Z_n and N_n.

Example 2

Assume the scenario of Example 1, but now we are given the auto-correlation functions for Z_n, and N_n. $R_Z(m) = 4\left(\frac{3}{4}\right)^{|m|}$, and N_n is white noise with $\sigma_N^2 = 1$.

a) Find the optimum filter with $p = 1$.
b) Find the mean square error for the resulting filter.



Solution:

Since $p = 1$

$$Y_t = \sum_{k=0}^{1} h_k X_{t-k} = h_0 X_t + h_1 X_{t-1}$$

The equations for the filter coefficients are

$$R_Z(0) = h_0[R_Z(0) + R_N(0)] + h_1[R_Z(-1) + R_N(-1)]$$
$$R_Z(1) = h_0[R_Z(1) + R_N(1)] + h_1[R_Z(0) + R_N(0)]$$

Now we substitute for the auto-correlation values

$R_Z(0) = 4$, $R_Z(\pm 1) = 3$

$R_N(0) = 1$, $R_N(\pm 1) = 0$

The equations are therefore

$$5h_0 + 3h_1 = 4$$
$$3h_0 + 5h_1 = 3$$

The solution is

$$h_0 = \frac{11}{16}, \quad h_1 = \frac{3}{16}$$

The error of the estimate is

$$\mathcal{E}\{(Z_n - Y_n)^2\} = R_Z(0) - \sum_{k=0}^{1} h_k R_{ZX}(k)$$

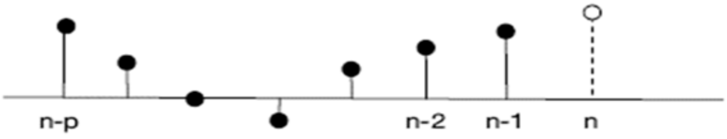
Now, $R_{ZX}(k) = \mathcal{E}(Z_{n+k}(Z_n + N_n) = \mathcal{E}(Z_{n+k}Z_n) + \mathcal{E}(Z_{n+k}N_n) = R_Z(k)$.
Hence the error is

$$\mathcal{E}\{(Z_n - Y_n)^2\} = R_z(0) - h_0 R_Z(0) - h_1 R_Z(1)$$
$$= (1 - h_0)R_z(0) - h_1 R_z(1)$$
$$\frac{5}{16}4 - \frac{3}{16}3 = \frac{11}{16}$$

Prediction

Consider a WSS random process Z_n. We wish to predict the random variable at a fixed n in terms of the random variables $Z_{n-1}, Z_{n-2}, \cdots, Z_{n-p}$. This falls into the general framework that we have discussed above, where $(X_{n-1}, \cdots, X_{n-p}) = (Z_{n-1}, \cdots, Z_{n-p})$, $a = p$, and $b = -1$. The prediction equation is

$$Y_n = \sum_{k=1}^{p} h_k X_{n-k}$$



The equations to be solved, according to the Wiener-Hopf theorem are

$$R_{ZX}(m) = \sum_{n=1}^{p} h_n R_X(m-n) \qquad 1 \le m \le p$$

$$R_{ZX}(m) = R_Z(m)$$

$$R_X(m-n) = R_Z(m-n)$$

The equations, listed in terms of matrices are

$$\begin{bmatrix} R_Z(1) \\ R_Z(2) \\ \vdots \\ R_Z(p) \end{bmatrix} = \begin{bmatrix} R_Z(0) & R_Z(-1) & \cdots & R_Z(p-1) \\ R_Z(1) & R_Z(0) & \cdots & R_Z(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_Z(p-1) & R_Z(p-2) & \cdots & R_Z(0) \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_p \end{bmatrix}$$

$$\boldsymbol{R_Z} = \mathbf{R}_Z\mathbf{h}$$

These are known as Yule-Walker equations. The solution can be obtained in a standard way by matrix inversion, although there are more efficient algorithms as a result of the special structure of $\mathbf{R}_Z$.

The **mean square error** is

$$\mathcal{E}(e_n^2) = R_Z(0) - \sum_{k=1}^{p} h_k R_Z(k)$$

Example

Consider the random process X_n. Estimate the value X_{n+m} in terms of X_n, i.e. $p = m,$ but to fit into the previous framework we find a solution where $h_1 = h_2 = \cdots = h_{m-1} = 0$

$$Y_n = \sum_{k=1}^{m} h_k\, X_{n-k} = h_m X_{n-m}$$

Assume $R_X(k) = 4\left(\frac{3}{4}\right)^{|k|}$

In this case there is only one filter coefficient, hence there is only one Yule-Walker equation.

$$R_Z(m) = h_m R_Z(0)$$

Note that $Z_n = X_n$

Hence $h_m = \frac{R_Z(m)}{R_Z(0)} = \frac{R_X(m)}{R_X(0)} = \frac{4\left(\frac{3}{4}\right)^m}{4} = \left(\frac{3}{4}\right)^m.$

The mean square error is $R_Z(0) - h_m R_Z(m) = R_X(0) - h_m R_X(m) = 4 - \left(\frac{3}{4}\right)^m 4\left(\frac{3}{4}\right)^m = 4\left(1 - \left(\frac{3}{4}\right)^{2m}\right) = 4\left(1 - \left(\frac{9}{16}\right)^m\right).$

Non-Causal Wiener Filter – Discrete Time Case

Suppose that Z_t is to be estimated by a linear function Y_t of the entire realization X_t, i.e. $a = b = \infty$.

$$Y_t = \sum_{i=-\infty}^{\infty} h_i X_{t-i}$$

The optimum filter will satisfy: $R_{ZX}(m) = \sum_{i=-\infty}^{\infty} h_i R_X(m-i)$, for all m

Note that $R_{ZX}(m) = (h * R_X)(m)$.

Taking the Discrete Fourier Transform we obtain

$$S_{ZX}(f) = H(f) S_X(f)$$

Hence the optimum filter satisfies

$$H(f) = \frac{S_{ZX}(f)}{S_X(f)}$$

This is called the Wiener Filter. Note that we call it a non-causal filter because at a time $t = t_0$ the output Y_t is a function of the present value, i.e. $t = t_0$, the past values, i.e. $t < t_0$, and the future values, i.e. $t > t_0$. The solution for the optimum filter is easy to compute if we do not put the restriction that the filter has to be causal, i.e. the output at $t = t_0$ depends only on the values $t \leq t_0$.

Estimation of Continuous-Time Processes

We seek a linear estimate $Y(t)$ of the continuous-time process $Z(t)$ in terms of the observation of a continuous time random process $X(t)$ in the time interval $[t-a, t+b]$, i.e. we utilize a linear filter with impulse response $h(t)$ to obtain an estimate of the process $Z(t)$ from the process $X(t)$. The estimate is as follows:

$$Y(t) = \int_{t-a}^{t+b} h(t-u) X(u) du = \int_{a}^{b} h(u) X(t-u) du$$

It can be shown that the optimum filter $h(t)$ that minimizes the mean squared error satisfies the following equation

$$R_{ZX}(\tau) = \int_{-b}^{a} h(u) R_X(\tau - u) du \qquad -a \leq \tau \leq a$$

Mean Square Error for the Optimum Filter

Note that $\mathcal{E}(e_t Y_t) = \mathcal{E}\left(e_t \int_{-b}^{a} h(u) X(t-u) du\right) = \int_{-b}^{a} h(u) \mathcal{E}\left(e_t X(t-u)\right) du = 0$

The last equality is due to the fact that the error signals is orthogonal to the observed signals. This is the orthogonality principle.

Then we evaluate

$\mathcal{E}(e^2(t)) = \mathcal{E}\big(e(t)(Z(t) - Y(t))\big) = \mathcal{E}(e(t)Z(t)) = \mathcal{E}\left(\big(Z(t) - Y(t)\big)Z(t)\right) = \mathcal{E}(Z^2(t)) - \mathcal{E}(Y(t)Z(t))$. Thus,

$$\mathcal{E}(e^2(t)) = R_Z(0) - \mathcal{E}\left(Z(t)\int_{-b}^{a} h(u)X(t-u)du\right)$$

$$\mathcal{E}(e^2(t)) = R_Z(0) - \int_{-b}^{a} h(u)R_{ZX}(u)du$$

Wiener-Hopf Theorem (Continuous Time)

Let $X(t)$ and $Z(t)$ be continuous-time zero-mean jointly wide-sense stationary processes and let $Y(t)$ be an estimate of $Z(t)$ of the form

$$Y(t) = \int_{t-a}^{t+b} h(t-u)X(u)du = \int_{-b}^{a} h(u)X(t-u)du$$

Then, the filter, $h(t)$, that minimizes the mean square error $\mathcal{E}\left\{\big(Z(t) - Y(t)\big)^2\right\}$ satisfies the equation

$$R_{ZX}(\tau) = \int_{-b}^{a} h(u)R_X(\tau - u)du \qquad -b \le \tau \le a$$

And the error is

$$\mathcal{E}\left\{\big(Z(t) - Y(t)\big)^2\right\} = R_Z(0) - \int_{-b}^{a} h(u)R_{ZX}(u)du$$

Continuous Time Non-Causal Wiener Filter

In the case of continuous time and with no restrictions on the observed signal, i.e. $a = b = \infty$, the estimate is the following:

$$Y(t) = \int_{-\infty}^{\infty} h(u)X(t-u)du$$

Then

$$R_{ZX}(\tau) = \int_{-\infty}^{\infty} h(u)R_X(\tau - u)du \qquad \text{for all } \tau$$

Which is the convolution

$$R_{ZX}(\tau) = (h * R_X)(\tau)$$

in the frequency domain it becomes

$$S_{ZX}(f) = H(f)S_X(f)$$

The solution for the filter in the frequency domain, i.e. the transfer function is then

$$H(f)=\frac{S_{ZX}(f)}{S_X(f)}$$

In general this filter is non-causal, i.e. $h(t)\neq 0$ for some values $t<0$.

Example

Let $X(t)=Z(t)+N(t)$, where $Z(t)$ and $N(t)$ are independent, $Z(t)$ has the power spectral density $S_Z(f)=\frac{4}{4+4\pi^2 f^2}$, and $N(t)$ is white noise with power spectral density $S_N(f)=\frac{N_0}{2}=\frac{1}{3}$. Find the optimum Wiener filter for the estimation of $Z(t)$, with no constraint of the filter being causal.

Solution:

Since $Z(t)$ and $N(t)$ are independent, we have $R_{ZX}(\tau)=R_Z(\tau)$, hence $S_{ZX}(f)=S_Z(f)$

Hence the transfer function for the optimum Wiener filter is

$$H(f)=\frac{S_Z(f)}{S_X(f)}=\frac{S_Z(f)}{S_Z(f)+S_N(f)}$$

$$\frac{\frac{4}{4+4\pi^2f^2}}{\frac{4}{4+4\pi^2f^2}+\frac{1}{3}}=\frac{12}{16+4\pi^2f^2}$$

Note that the transfer function can also be written as

$$H(f)=\frac{\frac{3}{2}}{4+j2\pi f}+\frac{\frac{3}{2}}{4-j2\pi f}$$

The impulse response is obtained as the inverse Fourier Transform

$$h(t)=\begin{cases}\frac{3}{2}e^{-4t} & t\geq 0\\ \frac{3}{2}e^{4t} & t<0\end{cases}$$

Note that the filter is non-causal, since we do not have $h(t)=0$ for $t<0$.

Continuous Time Causal Wiener Filter (infinite memory)

Now we assume that we wish to estimate the signal $Z(t)$ based only on past and present observations of $X(t)$, but not the future, i.e.

$$Y(t) = \int_0^\infty h(u)X(t-u)du$$

Then we must solve the equation

$$R_{ZX}(\tau) = \int_0^\infty h(u)R_X(\tau-u)du \qquad \text{for all } \tau$$

Similarly in the Discrete Case we must solve

$$R_{ZX}(m) = \sum_{i=0}^{\infty} h_i R_X(m-i) \qquad \text{for all } m$$

These equations are known as the Wiener-Hopf equations. They cannot be solved using the Fourier Transform techniques above because we are forcing the constraint that $h(t) = 0$ for $t < 0$, i.e. we are forcing the constraint that the filter must be causal.

Solution of the Wiener-Hopf Equations

Let us consider the discrete case of the Wiener-Hopf equations. The equations are

$$\begin{aligned}
&\vdots\\
R_{ZX}(-1) &= h_0R_X(-1) + h_1R_X(-2) + h_2R_X(-3) + \cdots\\
R_{ZX}(0) &= h_0R_X(0) + h_1R_X(-1) + h_2R_X(-2) + \cdots\\
R_{ZX}(1) &= h_0R_X(1) + h_1R_X(0) + h_2R_X(-1) + \cdots\\
R_{ZX}(2) &= h_0R_X(2) + h_1R_X(1) + h_2R_X(0) + \cdots\\
&\vdots
\end{aligned}$$

It is difficult to solve the above equations directly for a general set of correlation functions.

Special Case

Consider the case where the process X_t is white, i.e. $R_X(m) = \delta_m$. Then we can easily solve the above equations and obtain

$$h_m = \frac{R_{ZX}(m)}{R_X(0)} = R_{ZX}(m)$$

The corresponding filter transfer function is

$$H(f) = \sum_{m=0}^{\infty} h_m e^{-j2\pi fm} = \sum_{m=0}^{\infty} R_{ZX}(m)e^{-j2\pi fm}$$

Note that $H(f)$ is not equal to $S_{ZX}(f)$ because the sum does not include values of the cross-correlation $R_{ZX}(m)$ for negative m. But we can obtain the filter from the cross Power Spectral Density $S_{ZX}(f)$ by taking the inverse Fourier Transform and setting $h_m = 0$ for $m < 0$.

Solution for the General Case – Spectral Factorization

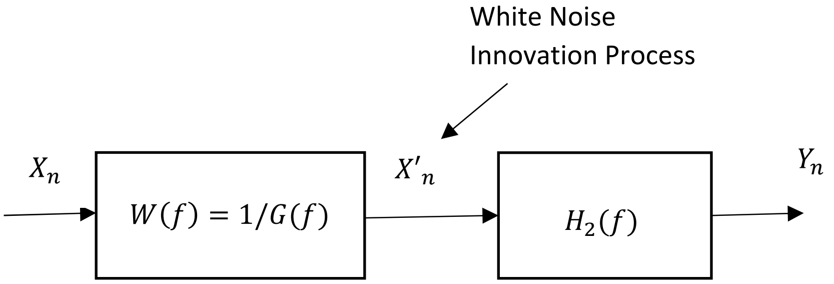
It can be shown that under very general conditions the power spectral density of a random process, $X(t)$, i.e. $S_X(f)$, can be factored as follows:

$$S_X(f) = |G(f)|^2 = G(f)G^*(f)$$

With $G(f)$ and $1/G(f)$ being causal filters.

Note that $S_X(f)$ is positive for all f. Hence we could take the square root $\sqrt{S_X(f)}$. Then we could set $G(f) = \sqrt{S_X(f)}$, and we see that $S_X(f) = G(f)G^*(f)$, because $\left(\sqrt{S_X(f)}\right)^* = \sqrt{S_X(f)}$. So we have achieved a factorization of $S_X(f)$. But this $G(f)$ is not necessarily the transfer function for a causal filter. Also any filter $Q(f)$ can be written as $Q(f) = A(f)e^{j\theta(f)}$, where $A(f)$ is the amplitude response and $\theta(f)$ is the phase response. Hence in factoring $S_X(f)$ into $G(f)G^*(f)$ we have one choice for the amplitude response but many choices for the phase response. The idea is that one of the choices for the phase response results in a causal filter. For example, suppose $S_X(f) = \frac{1}{1+4\pi^2 f^2}$. Then we can factor it as $\frac{1}{1+j2\pi f} \cdot \frac{1}{1-j2\pi f}$, i.e. with $G(f) = \frac{1}{1+j2\pi f}$. In this case the amplitude response for $G(f)$ is $A(f) = \frac{1}{\sqrt{1+4\pi^2 f^2}}$, and the phase response $\theta(f) = -\operatorname{atan}(2\pi f)$. And this results in a causal filter with impulse response $g(t) = e^{-t}$ for $t \geq 0$, and 0 elsewhere. But there are many other factorizations, for example using $\theta(f) = 2\pi f$, resulting in non-causal filters. This is because for any phase response $\theta(f)$, $G(f)G^*(f) = A(f)e^{j\theta(f)}A(f)e^{-j\theta(f)} = A^2(f)$.

Now if we can find a filter $G(f)$ such that $1/G(f)$ is causal then inputting X_n into the filter yields an output process X_n' with power spectral density $S_{X'}(f) = \frac{1}{|G(f)|^2} S_X(f) = \frac{|G(f)|^2}{|G(f)|^2} = 1$. This means that the filter output has a white power spectral density. It is called a white *noise innovation process*. The filter $W(f) = 1/G(f)$ is known as a whitening filter.



The innovation process contains all the information to reconstruct X_n, i.e. by using the filter $\frac{1}{W(f)} = G(f)$.

We now find the best estimate for Z_n using the whitened observation process X'_n instead of X_n. However the solution for the estimation filter, denoted in the figure as $H_2(f)$, is now easier because of the white input signal.

The solution for the optimum filter $H_2(f)$, is obtained as

$$H_2(f) = \sum_{m=0}^{\infty} R_{ZX'}(m) e^{-j2\pi fm}$$

But now we need the correlation $R_{ZX'}(m)$. Let the impulse response of the filter $W(z)$ be w_i, $(i = 0,1,2,\cdots)$

$$R_{ZX'}(k) = \mathcal{E}(Z_{n+k}X'_n)$$

$$= \mathcal{E}\left(Z_{n+k}\sum_{i=0}^{\infty} w_i X_{n-i}\right)$$

$$= \sum_{i=0}^{\infty} w_i \mathcal{E}(Z_{n+k}X_{n-i})$$

$$R_{ZX'}(k) = \sum_{i=0}^{\infty} w_i R_{ZX}(k+i)$$

We want to write the above as a convolution of two sequences, i.e. two discrete time signals. Let $v_i = w_{-i}$, by which we mean, if $\boldsymbol{w} = \cdots 0, w_0, w_1, w_2, \cdots$, then $\boldsymbol{v} = \cdots v_{-2}, v_{-1}, v_0, 0, \cdots$. $\boldsymbol{v}$ is the time reversed discrete signal of $\boldsymbol{w}$.

Let the Fourier Transform of $\boldsymbol{w}$ be $W(f) = \sum_{k=-\infty}^{\infty} w_k e^{-j2\pi fk} = \sum_{k=0}^{\infty} w_k e^{-j2\pi fk}$. Then the Fourier transform of v is

$$V(f) = \sum_{k=-\infty}^{\infty} v_k \, e^{-j2\pi fk} = \sum_{k=-\infty}^{0} v_k \, e^{-j2\pi fk} = \sum_{k=0}^{\infty} v_{-k} \, e^{j2\pi fk} = \sum_{k=0}^{\infty} w_k \, e^{j2\pi fk} = W^*(f)$$

The above correlation can be written as a convolution of the discrete signals $\boldsymbol{v} = (\cdots v_{-2}, v_{-1}, v_0, 0, \cdots)$ and $R_{ZX}(k)$. Change the index of summation from i to $-i$. The sum becomes

$$R_{ZX'}(k) = \sum_{i=0}^{-\infty} w_{-i} R_{ZX}(k-i) = \sum_{i=-\infty}^{0} w_{-i} R_{ZX}(k-i) = \sum_{i=-\infty}^{0} v_i R_{ZX}(k-i)$$

$$= (v * R_{ZX})(k)$$

Taking the Fourier Transform of both sides we obtain

$$S_{ZX'}(f) = V(f)R_{ZX}(f) = W^*(f)R_{ZX}(f)$$

Now invert use the inverse Fourier Transform to determine $R_{ZX'}(m)$ from $S_{ZX'}(f)$. Then keeping the values for $m = 0,1,2,\cdots$, we obtain the solution for the filter $H_2(f)$ in the time domain.

Summary of Solution for the Causal Wiener Filter

We wish to estimate the signal Z_n from observations of the signal X_n. We are given the power spectral densities $S_X(f)$ and $S_{ZX}(f)$.

1. Factor $S_X(f) = G(f)G^*(f)$ and obtain a causal whitening filter $W(f) = 1/G(f)$.
2. Determine $R_{ZX'}(m)$ from
$$R_{ZX'}(m) = \mathcal{F}^{-1}\left\{\frac{S_{ZX}(f)}{G^*(f)}\right\}$$
3. $H_2(f)$ is then given by $H_2(f) = \sum_{m=0}^{\infty} R_{ZX'}(m)e^{-j2\pi fm}$
4. The optimum filter is $H(f) = W(f)H_2(f)$

This procedure is also valid for the solution of the causal filter for continuous time processes, after changes are made, mostly using integrals instead of sums.

Example

Let $X(t) = Z(t) + N(t)$, where $Z(t)$ and $N(t)$ are independent, $Z(t)$ has the power spectral density

$$S_Z(f) = \frac{4}{4 + 4\pi^2 f^2}$$

and $N(t)$ is white noise with power spectral density $\frac{N_0}{2} = 1/3$. Find the optimum Wiener causal filter for the estimation of $Z(t)$

Solution:

First we find the whitening filter $W(f)$

$$S_X(f) = S_Z(f) + S_N(f)$$

$$= \frac{4}{4 + 4\pi^2 f^2} + \frac{1}{3} = \frac{16 + 4\pi^2 f^2}{3(4 + 4\pi^2 f^2)}$$

We factor $S_X(f)$ as:

$$S_X(f) = \frac{4 + j2\pi f}{\sqrt{3}(2 + j2\pi f)} \cdot \frac{4 - j2\pi f}{\sqrt{3}(2 - j2\pi f)}$$

Let

$$G(f) = \frac{4 + j2\pi f}{\sqrt{3}(2 + j2\pi f)}$$

Then

$$W(f) = \frac{\sqrt{3}(2 + j2\pi f)}{4 + j2\pi f}$$

Now,

$$S_{ZX'}(f) = W^*(f) S_{ZX}(f) = \frac{S_{ZX}(f)}{G^*(f)} = \frac{S_Z(f)}{G^*(f)}$$

Substituting for $S_Z(f)$ and $G^*(f)$

$$S_{ZX'}(f) = \frac{\frac{4}{4 + 4\pi^2 f^2}}{\frac{4 - j2\pi f}{\sqrt{3}(2 - j2\pi f)}} = \frac{4\sqrt{3}(2 - j2\pi f)}{(4 + 4\pi^2 f^2)(4 - j2\pi f)} = \frac{4\sqrt{3}}{(2 + j2\pi f)(4 - j2\pi f)}$$

This can be write as

$$S_{ZX'}(f) = \frac{\frac{2\sqrt{3}}{3}}{2 + j2\pi f} + \frac{\frac{2\sqrt{3}}{3}}{4 - j2\pi f}$$

The inverse Fourier Transform becomes

$$R_{ZX'}(\tau) = \begin{cases} \frac{2\sqrt{3}}{3} e^{-2\tau} & \tau \geq 0 \\ \frac{2\sqrt{3}}{3} e^{4\tau} & \tau < 0 \end{cases}$$

Then $H_2(f)$ becomes

$$H_2(f) = \mathcal{F}\left\{\frac{2\sqrt{3}}{3} e^{-2\tau} u(\tau)\right\} = \frac{\frac{2\sqrt{3}}{3}}{2 + j2\pi f}$$

Where $u(\tau)$ is the step function.

The transfer function for the optimum filter is then

$$H(f) = \frac{H_2(f)}{G(f)} = \frac{\frac{\frac{2\sqrt{3}}{3}}{2 + j2\pi f}}{\frac{4 + j2\pi f}{\sqrt{3}(2 + j2\pi f)}} = \frac{2}{4 + j2\pi f}$$

The impulse response is $h(t) = 2e^{-4t}$ for $t \geq 0$, and 0 elsewhere.

Markov Processes

First we consider a discrete time process X_n. The process X_n is a Markov process if the following property holds

$$P\left(X_{i_{n+1}} \leq x_{n+1} \middle| X_{i_n} = x_n, X_{i_{n-1}} = x_{n-1}, X_{i_{n-2}} = x_{n-2}, \cdots\right) = P(X_{i_{n+1}} \ \leq x_{n+1} | X_{i_n} = x_n)$$

where $i_{n+1} > i_n > i_{n-1} > i_{n-2} > \cdots$

In other words, the conditional CDF for the random variable $X_{i_{n+1}}$, conditioned on a number of past random variables at time indices i_n, $i_{n-1}, i_{n-2}, \cdots$, depends only on the most recent random variable, i.e. the one at time index i_n.

Note that to simplify notation we could just pick a set of points in time, not necessarily consecutive integers and index the random variables at these points with consecutive integers $1,2,\cdots$. Then we would write the above condition as

$$P(X_{n+1} \leq x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \cdots) = P(X_{n+1} = x_{n+1} | X_n = x_n)$$

For a continuous time process the Markov property becomes as follows:

Consider a fixed point in time t_1, where $t_1 > t_0$

$$P\big(X(t_1) \leq x_1 \big| X(t_0) = x_0, X(\tau) \text{ for } \tau \in (-\infty, t_0)\big) = P(X(t_1) \leq x_1 | X(t_0) = x_0)$$

Again, we fix $t_1 > t_0$ and consider the collection of random variables $X(t)$ for $t \in (-\infty, t_0]$. Then the conditional CDF for $X(t_1)$ given the collection of random variables indexed by the set $(-\infty, t_0]$ is equal to the conditional CDF for $X(t_1)$ given $X(t_0)$. In other words, having knowledge of the collection of random variables indexed by the set $(-\infty, t_0]$ is the same as having knowledge of the single random variable $X(t_0)$, in determining the CDF for the random variable $X(t_1)$.

Note that processes where the random variables at different points in time are independent are in a sense easier to analyze. For example, the joint CDF and PDF at the two points factors into a product of 1-dimensional CDF, or PDF. In a sense the Markov property represents a simplification in the same direction.

Property: For a Markov process, given that we know the present, the past is independent of the future, i.e. having knowledge of the past does not help in determining the future and vice-versa, i.e. having knowledge of the future does not help in determining the past. To show this consider three points in time $t_1 < t_2 < t_3$. Let us focus on t_2 which we denote as the present. The point t_1 is therefore in the past, and the point t_3 is in the future. Let the corresponding random variables be X_1, X_2, X_3.

Recall the definition of conditional probability. For events A and B, $P(A|B) = P(A \cap B)/P(B)$.

We can also write $P(A \cap B) = P(A|B)P(B)$. As we have shown previously this relation applies to random variables and their PDFs. For example, for two random variables X and Y, $f_{XY}(x, y) = f_{(X|Y)}(x|y)f_Y(y)$. This also extends to more than two random variables. For example

$$f_{XYZ}(x, y, z) = f_{(X|YZ)}(x|y, z)f_{YZ}(y, z)$$

Now, let us consider the above random variables X_1, X_2, X_3 and determine the joint conditional PDF for X_1, X_3 given X_2.

$$f_{X_1X_3|X_2}(x_1, x_3|x_2) = \frac{f_{X_1X_3X_2}(x_1, x_2, x_3)}{f_{X_2}(x_2)}$$

$$= \frac{f_{X_3|X_1X_2}(x_3|x_1, x_2)f_{X_1X_2}(x_1, x_2)}{f_{X_2}(x_2)}$$

$$= \frac{f_{X_3|X_1X_2}(x_3|x_1, x_2)f_{X_1|X_2}(x_1|x_2)f_{X_2}(x_2)}{f_{X_2}(x_2)}$$

$$f_{X_1X_3}(x_1, x_3|x_2) = f_{X_3|X_2}(x_3|x_2)f_{X_1|X_2}(x_1|x_2)$$

The first equality in the above follows from the definition of conditional probability. The second is also related to the definition of conditional probability where we split the three random variables into two groups, X_3 and X_1, X_2. The third equality (second factor) arises again from the definition of conditional probability. The fourth equality arises from the Markov property, i.e. conditioning on X_1 and X_2 is equal to conditioning only on X_2.

Since the joint conditional PDF factors we have proven independence of the random variables X_1 and X_3 given X_2.

Joint PDFs for a Markov Process

We consider a Markov process which may be continuous or discrete time. Fix a set of points in time $t_1, t_2, \cdots, t_n$. We will denote the corresponding random variables at these points as (assuming continuous time)

$$X_1 = X(t_1),\ X_2 = X(t_2), \cdots, X_n = X(t_n)$$

Now the joint PDF can be written in terms of conditional PDFs as follows:

$$f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = f_{X_1}(x_1)f_{X_2,\cdots,X_n|X_1}(x_2, \cdots, x_n|x_1)$$

$$= f_{X_1}(x_1)f_{X_2|X_1}(x_2|x_1)f_{X_3\cdots X_n|X_1X_2}(x_3, \cdots, x_n|x_1, x_2)$$

$$= f_{X_1}(x_1)f_{(X_2|X_1)}(x_2|x_1)f_{(X_3|X_1X_2)}(x_3|x_1, x_2)\cdots f_{(X_n|X_1,\cdots,X_{n-1})}(x_n|x_1, \cdots, x_{n-1})$$

Using the Markov property on the above expression we obtain

$$f_{X_1,\cdots,X_n}(x_1, x_2, \cdots, x_n) = f_{X_1}(x_1)f_{(X_2|X_1)}(x_2|x_1)f_{(X_3|X_2)}(x_3|x_2)\cdots f_{(X_n|X_{n-1})}(x_n|x_{n-1})$$

This is a lot simpler than the general case since it involves the factorization of the joint PDF into n conditional PDFs of one variable. In typical applications these probabilities are a lot easier to determine than probabilities involving more random variables.

Example 1

Coin Tossing

Suppose you have a game where a coin is tossed at times $i = 1,2,\cdots$ At the i^{th} toss you win one dollar if "Heads" occur and you loose a dollar if "Tails" occur. We are interested in the total money won/lost by the n toss, $X(n)$. Let I_i be the win/loss amount in the i^{th} toss.

Then $X(n) = \sum_{i=1}^{n} I_i$.

Now $X(n)$ depends on the values $X(1), X(2), \cdots, X(n-1)$. For example, if $X(n-1)$ is large then $X(n)$ is also very likely to be large, because $X(n) = X(n-1) \pm 1$. Also the probabilities for $X(n)$ conditioned on any $X(i)$ for $1 \leq i < n$ depend on $X(i)$, but it is clear that if we are given the value $X(n-1)$ then to calculate the probabilities for $X(n)$ we don't need the information on $X(1), X(2), \cdots, X(n-2)$, rather knowing the value $X(n-1)$ is all that is required, i.e.

$$P(X_n | X_{n-1}, X_{n-2}, \cdots, X_1) = P(X_n | X_{n-1})$$

Hence the random process X_n is a Markov process.

Example 2

Consider the Poisson counting process $N(t)$. $N(t)$ is equal to the number of arrivals in the time interval $[0, t]$. Now determine the conditional probability $P(N(t) = k | N(\tau) \text{ for } \tau \in [0, t_1])$, where $t_1 < t$. It is clear that $N(t) = N(t_1) + \text{number of arrivals in } (t_1, t)$. $N(t)$ does not depend on $N(\tau)$ for any $\tau < t_1$. There are mays ways that the process could have evolved in the time interval $[0, t_1]$ in order for there to be k arrivals in total, i.e. in order for $N(t_1) = k$.

But all that matters is that at time t_1 we have a total of k arrivals, and this information is contained in the value $N(t_1)$.

Clearly this is a Markov process. The Poisson process is an example of a continuous time Markov process.

Example 3

The Wiener process $W(t)$ is another example of a continuous time Markov Process. Let $\sigma^2 t$ be the variance of $W(t)$.

Consider the conditional PDF for the random variable $W(t)$ given $W(t_1), W(t_2), \cdots, W(t_n)$ for any set of points $t_1 < t_2 < \cdots < t_n < t$, i.e.

$$f_{(W | W_1, \cdots, W_n)}(w | w_1, \cdots, w_n)$$

where $W_i = W(t_i)$. Let $m = W_n$. Then it is clear that the random variable W is a Gaussian distributed random variable with mean m and variance $\sigma^2(t - t_n)$. Hence the conditional probability density for W given W_i $(i = 1, \cdots, n)$ is equal to the conditional probability for W given W_n.

Types of Markov Processes

The values that a Markov process takes are also referred to as ***States***. Just like for any process we can have four types of Markov Processes accordingly as the state space, $\mathcal{S}$, and the time set, $\mathcal{T}$, are discrete or continuous. Hence we have the following four types:

Discrete Time – Discrete State	(DT-DS)
Continuous Time – Discrete State	(CT-DS)
Discrete Time – Continuous State	(DT-CS)
Continuous Time – Continuous State	(CT-CS)

Now, if the **state is discrete** then it facilitates a special type of theory. The probabilities are described by the Probability Mass Function (PMF) and various probabilities can be computing using matrices as we will see below. We call these processes ***Markov Chains***.

Example of a Discrete Time Markov Chain

Take a random walk where steps are taken at the times $t_k = k\Delta$, $k = 1,2, \cdots$ Each step is an increment in the state by $\pm h$. The time set is $\mathcal{T} = \{\Delta, 2\Delta, \cdots\}$, and the state space is $\mathcal{S} = \{0, \pm h, \pm 2h, \cdots\}$. This is a discrete time Markov Chain.

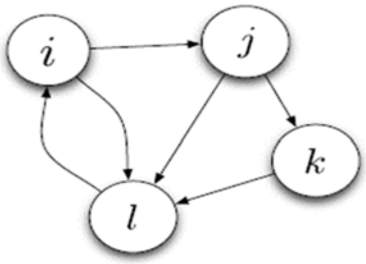
Example of a Continuous Time Markov Chain

Take the Poisson counting process with parameter λ. The time set is $\mathcal{T} = [0, \infty)$, and the state space is $\mathcal{S} = \{0,1,2, \cdots\}$. This is an example of a continuous time Markov Chain.

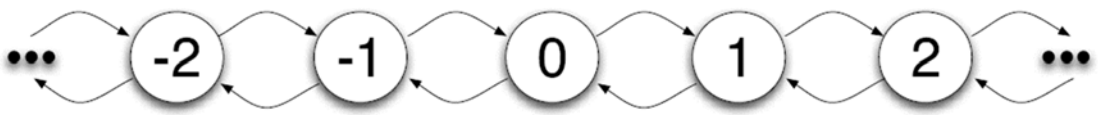
Representation of Markov Chains as Graphs

A Directed Graph is represented as $G = (V, E)$, where V is a set of vertices V and E a set of Directed Edges where a directed edge $e \in E$ is an ordered pair of vertices, i.e. $e = (v_1, v_2)$, $v_1, v_2 \in V$. Note that we can have the case where $v_1 = v_2$, i.e. an edge $e = (v, v)$.

A Markov Chain can be represented by a directed graph as follows: $V = \mathcal{S}$, i.e. the set of states. The set of edges is $E = \{(v_1, v_2): v_1, v_2 \in V$ with the condition that a transition from state v_1 to state v_2 occurs with **non-zero** probability$\}$. The following is an example of a Markov chain with 4 states and 6 possible transitions (i.e. transitions with non-zero probabilities as follows: $(i, j), (i, l), (j, k), (j, l), (k, l)$, and (l, i). In the case of a discrete time Markov Chain the transitions occur at discrete points in time n, if the time increments are 1. In the case of a continuous time Markov Chain the transitions may occur at any point in time t.



The following is the Graph for a Markov chain with an infinite number of states. This Markov Chain is associated with the Coin Tossing game In this example there are two possible transitions (non-zero probability) out of each state: a transition to the right, i.e. Heads outcome, or a Transition to the left (Tails outcome). The probability for the transition to the right is $P(\text{Heads})$, and the probability of transition to the left is $P(\text{Tails})$.



The resulting graph shown above are also referred to as state diagrams.

Discrete Time vs. Continuous Time Markov Chains

In the analysis of Markov Chains we consider two cases: **Discrete Time and Continuous Time**. Discrete Time Markov Chains are characterized by transition probabilities, i.e. at each point in Discrete Time, n, we define the one step transition probabilities $p_{ij}(n) = P(X_{n+1} = j|X_n = i)$. These probabilities must satisfy the condition $\sum_{j=1}^{N} p_{ij}(n) = 1$. If these probabilities are constant with respect to the time n then we refer to the Markov chain as a **Homogeneous Markov Chain.** In the following we will focus on Homogeneous Markov Chains.

In the case of Continuous time Markov processes the transitions may occur at any point in time. The transition probabilities are then defined as $p_{ij}(s,t) = P(X(t) = j|X(s) = i)$, with $0 \leq s < t$. The transition probabilities for a Markov Chain define the behaviour of the Markov Chain. In the case of discrete time the Markov Chain is described by an $n \times n$ matrix of transition probabilities, where n is the number of states. This includes an infinite dimensional matrix if the number of states is infinite. In the case of Continuous time the analysis may involve derivatives and we speak of transition rates between states. In the following we will focus on ***Discrete Time Markov Chains*** (DTMC).

Discrete Time Homogeneous Markov Chains

Note that the state space S can be finite or countably infinite. Note that in some mathematical formulations it is very convenient to think of the set of edges as $E = (v_1, v_2)$ with no restrictions on whether or not a transition is possible with non-zero probability. In this case if the number of states is n, then the number of edges is n^2. We would not draw these graphs because there would be no useful information in the graph, but in developing some equations it is convenient to think of the "full" set of edges. The Markov Chain can be represented by an $n \times n$ matrix where each element refers to an edge and the value of the matrix element is a transition probability between the two states associated with the particular element. If the matrix element is equal to 0 then the edge would not be present in the graph.

One Step Transition Probability Matrix

The one step Transition Probability Matrix for a Homogeneous Markov Chain is as follows:

$$\boldsymbol{P} = \begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0j} & \cdots \\ p_{10} & p_{11} & \cdots & p_{1j} & \cdots \\ \vdots & \vdots & \cdots & \vdots & \cdots \\ p_{i0} & p_{i1} & \cdots & p_{ij} & \cdots \\ \vdots & \vdots & \cdots & \vdots & \cdots \end{bmatrix}$$

For a finite state Markov Chain the Transition Probability Matrix is a square matrix with entries $0 \le p_{ij} \le 1$.

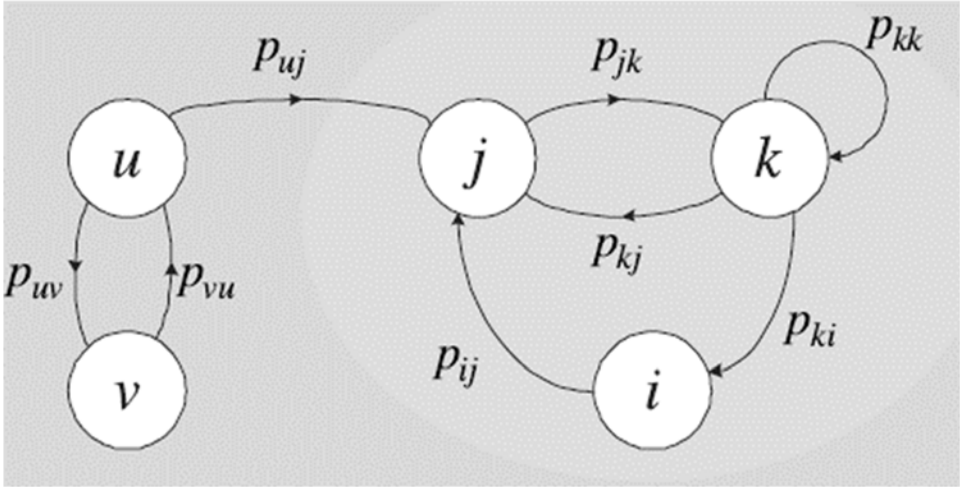
Each row of $\boldsymbol{P}$ represents the transition probabilities out of the state, include the probability that it remains in the state, i.e. p_{ii} for the i^{th} row. As a result we must have $\sum_{j=0}^{N-1} p_{ij} = 1$, where N is the number of states ($N = \infty$ for an infinite number of states).

Each column of $\boldsymbol{P}$ represents the probabilities into the corresponding state, i.e. the state indexed by the given column.

Graphical Representation of Homogeneous DTMC

The one step Transition Probability Matrix is used to represent the Markov Chain as a Graph. The weight of each edge of the graph is a corresponding element of the matrix. In the Figure below we shown a graph corresponding to the matrix $\boldsymbol{P}$. In the matrix we index the rows and columns by the integers 0,1,2,3,4, i.e. $(j, k, i, u, v) = (0,1,2,3,4)$

$$P = \begin{bmatrix} 0 & p_{jk} & 0 & 0 & 0 \\ p_{kj} & p_{kk} & p_{ki} & 0 & 0 \\ p_{ij} & 0 & 0 & 0 & 0 \\ p_{uj} & 0 & 0 & 0 & p_{uv} \\ 0 & 0 & 0 & p_{vu} & 0 \end{bmatrix}$$



.

A matrix such as the above P, where the rows sum to 1 is referred to as a Stochastic Matrix. Such a matrix can represent a Markov Chain.

The $\boldsymbol{n}$-step Transition Probability

The n-step transition probability for states i and j is the probability that given that the process is in state i at some time discrete t, then it will be in state j, at time $t + n$, i.e. in n-steps in the future. We will denote this probability as $p_{ij}(n)$. Note that we used this notation previously for a the case where the transition probabilities depend on time, i.e. non-homogeneous. Here we are using the notation for a different purpose.

The n-step Transition Probability Matrix for a Markov Chain with $N + 1$ states then becomes

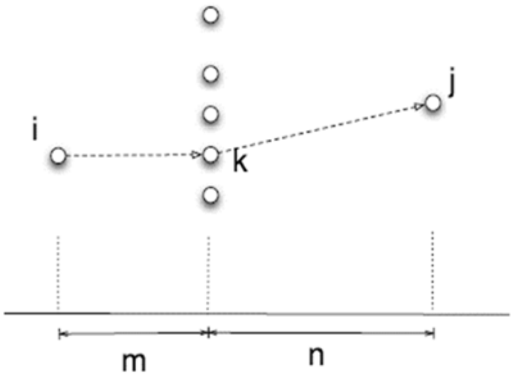
$$\boldsymbol{P_n} = \begin{bmatrix} p_{00}(n) & \cdots & p_{0N}(n) \\ \vdots & \ddots & \vdots \\ p_{N0}(n) & \cdots & p_{NN}(n) \end{bmatrix}$$

Note that since $\boldsymbol{P}$ is the one-step Transition Probability Matrix we have

$$\boldsymbol{P_1} = \boldsymbol{P}$$

Discrete-Time Chapman-Kolmogorov Equations

For a Homogeneous Discrete-Time Markov Chain the probability that the process reaches state j at time $m + n$ given that it starts at state i can be determined using the law of total probability after we condition that the process starts at state i at time 0. We condition on $X(0) = i$. Then we consider the set of events E_k where $E_k = \{\text{all sample functions with } X(0) = i, X(m) = k, \text{ and } X(m + n) = j\}$ – see the Figure below.



Let $E = \{\text{all sample functions with } X(0) = i \text{ and } X_{m+n} = j\}$. We can see that the sets E_k are disjoint and $\cup\, E_k = E$, that is the sets E_k form a partition of E.

By the law of total probability we have $P(E) = \sum_k P(E_k)$. But by the Markov property $P(E_k) = p_{ik}(m)p_{kj}(n)$. Now, $P(E) = p_{ij}(m+n)$. Hence we have

$$p_{ij}(m+m) = \sum_k p_{ik}(m)p_{kj}(n)$$

This can be written in terms of the n-step Transition Probability matrices as

$$\boldsymbol{P_{m+n} = P_m P_n}$$

Another way to write the above in more explicit terms is as follows:

$$p_{ij}(m+n) = P(X_{m+n} = j | X_0 = i)$$

$$= \sum_k P(X_{m+n} = j | X_0 = i, X_m = k) P(X_m = k | X_0 = i)$$

$$= \sum_k P(X_{m+n} = j | X_m = k) P(X_m = k | X_0 = i)$$

The equations are known as the Chapman-Kolmogorov equations.

Iterating the CK Equations

Using the above equations we can start with $m = 1, n = 1,$ and write

$$\boldsymbol{P_2 = P_{1+1} = P_1 P_1 = PP = P^2}$$

$$\boldsymbol{P_3 = P_{2+1} = P^2 P = P^3}$$

$$\dots$$

$$\boldsymbol{P_n = P_{(n-1)+1} = P^{n-1} P = P^n}$$

Hence all the n-step transition probabilities can be obtained from the single 1-step transition probability matrix $\boldsymbol{P}$.

State Probabilities

We can ask the question, what are the probabilities for the states at a given point in time n? We denote these as $p_i(n)$, where i is an index over all states, and n is the time index. In other words $p_i(n) = P(X_n = i)$. These are also known as marginal probabilities, just as for any discrete-time, discrete-space process. Using the law of total probability we can write these probabilities as follows:

$$p_i(n) = \sum_j p_j(0)p_{ji}(n)$$

where, as above, $p_{ji}(n)$ is the n-step transition probability between state j and state i.

Note that we can also compute the state probabilities recursively as follows:

$$p_i(n) = \sum_j p_j(n-1)p_{ji}$$

To do this we need the initial probabilities at time $n = 0$, i.e. $p_i(0)$.

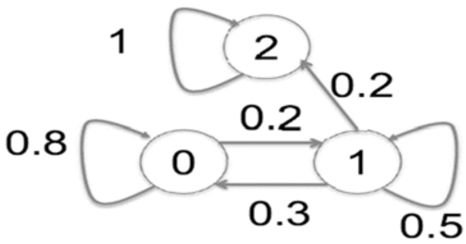
Define the state probability vector at time n as $\boldsymbol{p_n} = \left(p_0(n), p_1(n), \cdots, p_N(n)\right)$. Then for the times $n = 0,1,2,\cdots$

$$\boldsymbol{p_n} = \boldsymbol{p_0 P_n}$$

Example – Three state Markov Chain

Consider the Markov Chain in the Figure below. Assume that $\boldsymbol{p_0} = [1, 0, 0]$. Find $\boldsymbol{p_1}$ and $\boldsymbol{p_2}$.

Find $\lim_{n\to\infty} \boldsymbol{p_n} = \boldsymbol{p}_\infty$.



First we determine $\boldsymbol{P} = \begin{bmatrix} 0.8 & 0.2 & 0 \\ 0.3 & 0.5 & 0.2 \\ 0 & 0 & 1 \end{bmatrix}$

Then we compute $\boldsymbol{p_1} = \boldsymbol{p_0 P} = [1,0,0] \begin{bmatrix} 0.8 & 0.2 & 0 \\ 0.3 & 0.5 & 0.2 \\ 0 & 0 & 1 \end{bmatrix} = [0.8, 0.2, 0]$

$$p_2 = p_1P = [0.8, 0.2, 0]\begin{bmatrix} 0.8 & 0.2 & 0 \\ 0.3 & 0.5 & 0.2 \\ 0 & 0 & 1 \end{bmatrix} = [0.7, 0.26, 0.04]$$

Note that since there are no transitions out of state 2, once the process enters state 2 it will never exit. Hence $p_\infty = [0,0,1]$.

We can also test this numerically (e.g. use Matlab)

$$\boldsymbol{p_{10}} = [.398, .182, .42]$$

$$\boldsymbol{p_{100}} = [.00116, .00053, .998]$$

But there is another way to find the limit, if the limit exists which is the case in this example. Now, we can write $\boldsymbol{p_{n+1}} = \boldsymbol{p_n P},$ and if the sequence $\boldsymbol{p_n}$ converges to a limit then for large n, $\boldsymbol{p_{n+1}} = \boldsymbol{p_n}.$ Hence we can write $\boldsymbol{p_\infty} = \boldsymbol{p_\infty P},$ and solve for $\boldsymbol{p_\infty}.$ Hence

$$\boldsymbol{p_\infty I} = \boldsymbol{p_\infty P}$$

By taking the transpose of both sides, we can also write this as follows

$$\boldsymbol{P^T p_\infty} = \boldsymbol{p_\infty}$$

where we now consider $\boldsymbol{p_\infty}$ to be a column vector instead of the row vector above. Hence $\boldsymbol{p_\infty}$ is an eigenvector of $\boldsymbol{P^T}$ corresponding to the eigenvalue $\lambda = 1$.

$$(\boldsymbol{P^T} - \boldsymbol{I})\boldsymbol{p_\infty} = \boldsymbol{0}$$

In general this equation will have multiple solutions, i.e. multiple eigenvectors corresponding to the eigenvalue $\lambda = 1$. So lets see what we get in this case.

$$\boldsymbol{P^T} - \boldsymbol{I} = \begin{bmatrix} .8 & .2 & 0 \\ .3 & .5 & .2 \\ 0 & 0 & 1 \end{bmatrix}^T - \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} .8 & .3 & 0 \\ .2 & .5 & 0 \\ 0 & .2 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} -.2 & .3 & 0 \\ .2 & -.5 & 0 \\ 0 & .2 & 0 \end{bmatrix}$$

$$\begin{bmatrix} -.2 & .3 & 0 \\ .2 & -.5 & 0 \\ 0 & .2 & 0 \end{bmatrix}\boldsymbol{p_\infty} = \boldsymbol{0}$$

We solve this system of equations with the constraint that the sum of the probability of p_∞ equals 1, i.e. $[1,1,1]\boldsymbol{p_\infty} = 1$.

Now, lets solve the above system of equations using Gaussian elimination.

Solving the above using row-reduction we obtain

$$\begin{bmatrix} -.2 & .3 & 0 \\ .2 & -.5 & 0 \\ 0 & .2 & 0 \end{bmatrix} \sim \begin{bmatrix} -.2 & 0 & 0 \\ .2 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

So, let $\boldsymbol{p_\infty} = [p_0, p_1, p_2]^T$. Then $p_2 = \mu$, $p_1 = 0$, and $p_0 = 0$.

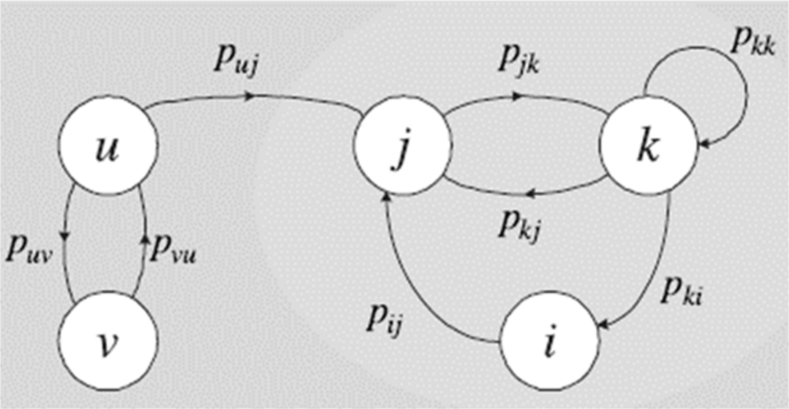
Hence $\boldsymbol{p_\infty} = \mu(0, 0, 1)^T$. Applying the constraint gives $\mu = 1$.

The solution is $(p_0, p_1, p_2) = (0, 0, 1)$ which is what we expected from above.

Examples

Consider the following Markov Chain that from above with state indexing

$$(j, k, i, u, v) = (0,1,2,3,4)$$



Assume the Transition Probability Matrix

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/4 & 1/2 & 1/4 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 2/3 & 0 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

The 10-step Transition Probability Matrix is

$$P_{10} = \begin{pmatrix} 0.285 & 0.572 & 0.143 & 0 & 0 \\ 0.286 & 0.571 & 0.143 & 0 & 0 \\ 0.287 & 0.57 & 0.143 & 0 & 0 \\ 0.284 & 0.571 & 0.141 & 4.115 \times 10^{-3} & 0 \\ 0.289 & 0.564 & 0.143 & 0 & 4.115 \times 10^{-3} \end{pmatrix}$$

The 100-step Transition probability Matrix is

$$P_{100} = \begin{pmatrix} 0.286 & 0.571 & 0.143 & 0 & 0 \\ 0.286 & 0.571 & 0.143 & 0 & 0 \\ 0.286 & 0.571 & 0.143 & 0 & 0 \\ 0.286 & 0.571 & 0.143 & 0 & 0 \\ 0.286 & 0.571 & 0.143 & 0 & 0 \end{pmatrix}$$

We observe that the transition probabilities into states $\{4,5\} = \{u, v\}$ are tending to 0, and the transition probabilities into states $\{j, k, i\}$ are all equal regardless of the starting state. This means that the state probabilities for large n converge to a single vector which in this case is $\boldsymbol{p}_\infty = [.286, .571, .143, 0, 0]$.

This can be seen from the state diagram, once the process leaves states $\{u, v\}$, it will never return to those states. We can of course solve for the limiting state probabilities as we have done for the 3-state example.

Stationary Markov Chains

Definition:

A Markov chain is stationary if $\boldsymbol{p_0} = \boldsymbol{p_0 P}$**.** This means that $\boldsymbol{p_n} = \boldsymbol{p_0}$

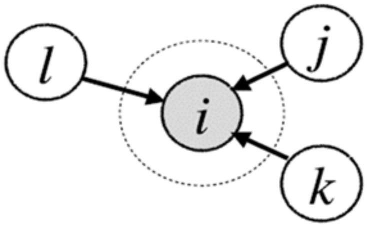
Example:

The above Markov Chain with initial state probabilities $\boldsymbol{p_0} = [0.285,\ 0.571,\ 0.143, 0, 0]$ is a stationary Markov Chain.

Note that the relation $\boldsymbol{p_0} = \boldsymbol{p_0 P}$ means that $\boldsymbol{p_0}$ is a left eigenvector of the matrix $\boldsymbol{P}$. Since $p_n = p_0$, the state probabilities become independent of time and we may write $\boldsymbol{p} = \boldsymbol{pP}$.

Global Balance Equations

In many cases when we solve the equation $\boldsymbol{p} = \boldsymbol{pP}$ for $\boldsymbol{p}$, especially if $\boldsymbol{P}$ is a sparse matrix, it is convenient to set up an equation for each state as follows: Consider state i and assume that the only states with transitions into state i, are the states j, k, l. Then we can write the equation $p_i = p_j p_{ji} + p_k p_{ki} + p_l p_{li}$. Performing this for each state gives a set of equations which can be represented in matrix form as $\boldsymbol{p} = \boldsymbol{pP}$. This states that the "flow probability into state i is equal to p_i. The resulting set of equations is also called the ***Global Balance Equations***.



Accessibility

A state j in a Markov Chain is ***accessible*** from state i if for some integer $n \geq 0$, $p_{ij}(n) > 0$. This means that in the Markov Chain graph there exists a path from state i to state j. In the above 5-state Markov Chain example state k is accessible from state v, but state v is not accessible from state k.

Communicating States

Definition: If state i in a Markov Chain is accessible from state j and state j is accessible from state i then we say that states i and j ***communicate***.

In the above 5-state example, states i and k communicate, but states v and k do not communicate.

Note that paths between two communicating states do not need to have the same number of edges.

Communicating Classes of States

It is easy to see that if state i communicates with state j and state j communicates with state k, then state i communicates with state k. We say that the relation "Communicates" is an ***Equivalence Relation***, and an equivalence relation on a set partitions the set into a set of classes which we call communicating classes. Any two classes are disjoint and the union of all classes is equal to the whole state space. The 5-state example above contains two communicating classes: $\{u, v\}$ and $\{i, j, k\}$.

Irreducible Markov Chain

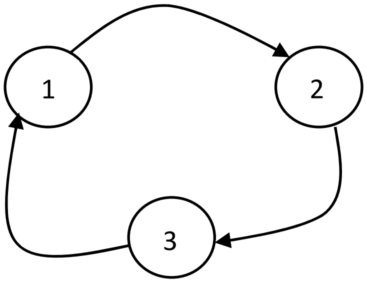
A Markov Chain is ***irreducible*** if it contains one communicating class. The 5-state example above is not an irreducible Markov Chain. In fact for large n we can reduce the Markov Chain to a Chain containing the states $\{i, j, k\}$. After this reduction we end up with a 3-state irreducible Markov Chain.

Periodic Markov Chains

Definition: The period of state i of a Markov Chain, written as d_i is the greatest common divisor (gcd) of all integers $n \geq 1$ for which $p_{ii}(n) > 0$.

Note that $p_{ii}(n) > 0$, means that starting at state i it is possible to return to state i in n-steps.

For the Markov Chain below: Starting at state 1 at time 0, returns are possible at steps 3, 6, 9, …

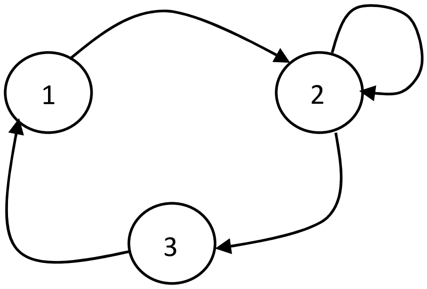


Now, $\gcd(3,6,9,\cdots) = 3$. Hence $d_1 = 3$.

In the same manner we can show that $d_2 = d_3 = 3$.

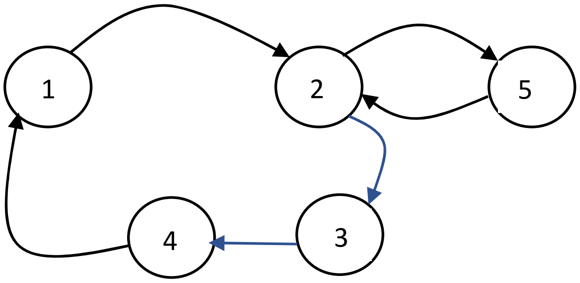
However if we modify the Markov Chain as below, then starting at state 1, returns are possible at times $n = 3,4,5,6,\cdots$. Now $\gcd(3,4,5,6,\cdots\} = 1$. Hence $d_1 = 1$.

Also, for state 2, returns are possible at $n = 1,2,3,\cdots$. $\gcd(1,2,3,\cdots) = 1$. Hence $d_2 = 1$. In a similar manner we show that $d_3 = 1$.



Now, consider the Markov Chain below. Starting at state 1 at $n = 0$, returns are possible at times $n = 4,6,8,10,\cdots$. $\gcd(4,6,8,10,\cdots) = 2$. Hence $d_2 = 2$.

Similarly starting at state 2 at $n = 0$, returns are possible at times $n = 2,4,6,8,10,\cdots$. $\gcd(2,4,6,8,10,\cdots) = 2$. Hence $d_2 = 2$. Similarly we can show that $d_3 = d_4 = d_5 = 2$.



Note that if a Markov Chain has a state with period $d_i > 1$, then the state probabilities $\boldsymbol{p_n}$ will ***not converge*** as $n \to \infty$.

Consider the first of the above Markov Chains (3 nodes, 3 edges). The Transition Probability Matrix is $\boldsymbol{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$

For the state probability vector to converge we must have P^n converge to a matrix with all rows equal to a constant vector. But consider

$$\boldsymbol{P^2} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$\boldsymbol{P}^3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

We can see that the sequence of matrices $\boldsymbol{P}, \boldsymbol{P}^2, \boldsymbol{P}^3, \boldsymbol{P}^4 = \boldsymbol{P}$ does not converge to a matrix with equal rows. In fact it is a periodic sequence of matrices with period equal to 3.

Hence the state probability vector does not converge to a constant vector as $n \to \infty$.

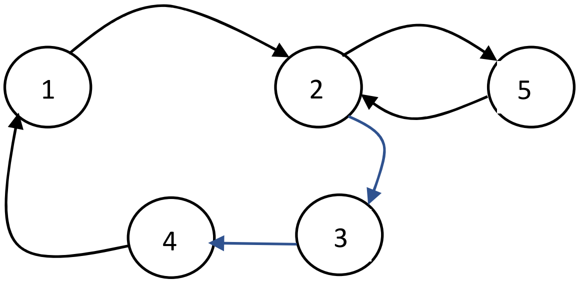
Period in an Irreducible Markov Chain

Theorem: In an irreducible Markov Chain for which states i and j communicate, we have $d_i = d_j$.

This theorem indicates that periodicity is a class property. All states in a communicating class have the same period. The textbook discusses the proof of this theorem in page 665.

Insight into Periodicity

To get more insight into periodicity we reconsider one of the above examples as follows



We will assume that $p_{25} = 1/2$ and $p_{23} = \frac{1}{2}$. All other transition probabilities are obviously 1.

The Transition probability matrix is as follows:

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & .5 & 0 & .5 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Let us determine the n-step probabilities, i.e. $\boldsymbol{P_n} = \boldsymbol{P^n}$. We have computed these for $n = 1, \cdots 10$, $n = 19$, $n = 20$, $n = 99$, and $n = 100$. In the first column in the page below we show all the values for n odd and in the second column we show the values for n even. We can see the convergence of the odd values and the even values. If the Markov chain was not periodic then there would be convergence to a single value.

$$P := \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & .5 & 0 & .5 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} \qquad P^2 = \begin{pmatrix} 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \end{pmatrix}$$

$$P^3 = \begin{pmatrix} 0 & 0.5 & 0 & 0.5 & 0 \\ 0.5 & 0 & 0.25 & 0 & 0.25 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0 & 0.5 & 0 \end{pmatrix} \qquad P^4 = \begin{pmatrix} 0.5 & 0 & 0.25 & 0 & 0.25 \\ 0 & 0.75 & 0 & 0.25 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0.5 & 0 & 0.25 & 0 & 0.25 \end{pmatrix}$$

$$P^5 = \begin{pmatrix} 0 & 0.75 & 0 & 0.25 & 0 \\ 0.25 & 0 & 0.375 & 0 & 0.375 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0.5 & 0 & 0.25 & 0 & 0.25 \\ 0 & 0.75 & 0 & 0.25 & 0 \end{pmatrix} \qquad P^6 = \begin{pmatrix} 0.25 & 0 & 0.375 & 0 & 0.375 \\ 0 & 0.625 & 0 & 0.375 & 0 \\ 0.5 & 0 & 0.25 & 0 & 0.25 \\ 0 & 0.75 & 0 & 0.25 & 0 \\ 0.25 & 0 & 0.375 & 0 & 0.375 \end{pmatrix}$$

$$P^7 = \begin{pmatrix} 0 & 0.625 & 0 & 0.375 & 0 \\ 0.375 & 0 & 0.313 & 0 & 0.313 \\ 0 & 0.75 & 0 & 0.25 & 0 \\ 0.25 & 0 & 0.375 & 0 & 0.375 \\ 0 & 0.625 & 0 & 0.375 & 0 \end{pmatrix} \qquad P^8 = \begin{pmatrix} 0.375 & 0 & 0.313 & 0 & 0.313 \\ 0 & 0.688 & 0 & 0.313 & 0 \\ 0.25 & 0 & 0.375 & 0 & 0.375 \\ 0 & 0.625 & 0 & 0.375 & 0 \\ 0.375 & 0 & 0.313 & 0 & 0.313 \end{pmatrix}$$

$$P^9 = \begin{pmatrix} 0 & 0.688 & 0 & 0.313 & 0 \\ 0.313 & 0 & 0.344 & 0 & 0.344 \\ 0 & 0.625 & 0 & 0.375 & 0 \\ 0.375 & 0 & 0.313 & 0 & 0.313 \\ 0 & 0.688 & 0 & 0.313 & 0 \end{pmatrix} \qquad P^{10} = \begin{pmatrix} 0.313 & 0 & 0.344 & 0 & 0.344 \\ 0 & 0.656 & 0 & 0.344 & 0 \\ 0.375 & 0 & 0.313 & 0 & 0.313 \\ 0 & 0.688 & 0 & 0.313 & 0 \\ 0.313 & 0 & 0.344 & 0 & 0.344 \end{pmatrix}$$

$$P^{19} = \begin{pmatrix} 0 & 0.666 & 0 & 0.334 & 0 \\ 0.334 & 0 & 0.333 & 0 & 0.333 \\ 0 & 0.668 & 0 & 0.332 & 0 \\ 0.332 & 0 & 0.334 & 0 & 0.334 \\ 0 & 0.666 & 0 & 0.334 & 0 \end{pmatrix} \qquad P^{20} = \begin{pmatrix} 0.334 & 0 & 0.333 & 0 & 0.333 \\ 0 & 0.667 & 0 & 0.333 & 0 \\ 0.332 & 0 & 0.334 & 0 & 0.334 \\ 0 & 0.666 & 0 & 0.334 & 0 \\ 0.334 & 0 & 0.333 & 0 & 0.333 \end{pmatrix}$$

$$P^{99} = \begin{pmatrix} 0 & 0.667 & 0 & 0.333 & 0 \\ 0.333 & 0 & 0.333 & 0 & 0.333 \\ 0 & 0.667 & 0 & 0.333 & 0 \\ 0.333 & 0 & 0.333 & 0 & 0.333 \\ 0 & 0.667 & 0 & 0.333 & 0 \end{pmatrix} \qquad P^{100} = \begin{pmatrix} 0.333 & 0 & 0.333 & 0 & 0.333 \\ 0 & 0.667 & 0 & 0.333 & 0 \\ 0.333 & 0 & 0.333 & 0 & 0.333 \\ 0 & 0.667 & 0 & 0.333 & 0 \\ 0.333 & 0 & 0.333 & 0 & 0.333 \end{pmatrix}$$

From the above we see convergence for n odd with the following limit

$$\boldsymbol{P}_{\infty}^{\text{odd}} = \begin{bmatrix} 0 & 2/3 & 0 & 1/3 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 \\ 0 & 2/3 & 0 & 1/3 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 \\ 0 & 2/3 & 0 & 1/3 & 0 \end{bmatrix}$$

And for n even we see convergence with the following limit

$$\boldsymbol{P}_{\infty}^{\text{even}} = \begin{bmatrix} 1/3 & 0 & 1/3 & 0 & 1/3 \\ 0 & 2/3 & 0 & 1/3 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 \\ 0 & 2/3 & 0 & 1/3 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 \end{bmatrix}$$

Aperiodic Markov Chains

Now let us modify the above Markov Chain by introducing a transition (a loop) from state 1 back to state 1 with probability $p_{11} = .5$. This means that we must change p_{12} to $p_{12} = .5$. The Transition Probability Matrix is now

$$\boldsymbol{P} = \begin{bmatrix} .5 & .5 & 0 & 0 & 0 \\ 0 & 0 & .5 & 0 & .5 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

We can easily see that this Markov Process has periodicity for state 1 as $d_1 = 1$. By the above theorem all states have periodicity equal to 1. Hence the Markov Chain is not periodic. We call it ***Aperiodic***. Let us compute the limit of P^n.

The limit exists and is equal to the following:

$$\boldsymbol{p}_{\infty} = \begin{bmatrix} 2/7 & 2/7 & 1/7 & 1/7 & 1/7 \\ 2/7 & 2/7 & 1/7 & 1/7 & 1/7 \\ 2/7 & 2/7 & 1/7 & 1/7 & 1/7 \\ 2/7 & 2/7 & 1/7 & 1/7 & 1/7 \\ 2/7 & 2/7 & 1/7 & 1/7 & 1/7 \end{bmatrix}$$

Theorem

If state i has period d_i then there exists an integer N depending on i such that for all integers $n \geq N$, we have $p_{ii}(nd_i) > 0$.

We will not prove this theorem, but we can see that in the above example it holds. The period for all states is 2. In the above example $N = 1$ for states 2 and 5, and $N = 2$ for states 1,3,4. And we see that in the first case for $n \geq 1$, nd_i takes values $2,4,6,\cdots$ and for all such values $p_{ii}(nd_i) > 0$

$(i = 2,5)$. In the second case $(N = 2)$ for $n \geq 2, nd_i$ takes values $4,6,8,\cdots$ and for all such values $p_{ii}(nd_i) > 0$ $(i = 1,3,4)$.

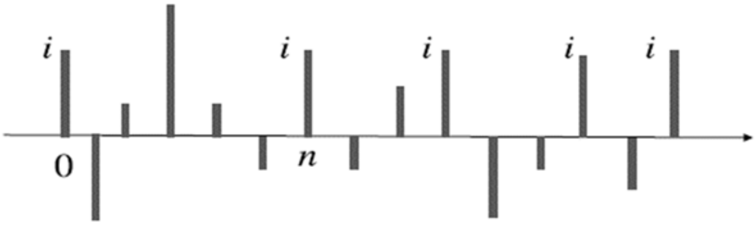
First Return

Let $F_{ii}(n)$ be the event corresponding to returning to state i in n steps for the first time, given that the process starts in state i. This means the set of sample functions that start in state i transition through various other states and are back at state i at the n^{th} step. It can be written as

$$F_{ii}(n) = \{X_n = i, X_v \neq i, v = 1,2,\cdots, n-1 | X_0 = i\}$$

For the different values of n these events are disjoint, i.e,

$F_{ii}(n) \neq F_{ii}(m), \quad n \neq m$



Probability of First Return

Let us define $f_{ii}(n)$, the probability of first return to state i at the n^{th} step, given that we start at state i at time $n = 0$

$$f_{ii}(n) = P(F_{ii}(n))$$

We set $f_{ii}(0) = 1$.

Now, if the process ever returns to state i, it will occur at either $n = 1, 2, \cdots$ The event that the process returns (ever) to state i is $E_i = \cup_{n=1}^{\infty} F_{ii}(n)$. And since the events $F_{ii}(n)$ (with i fixed and n varying) are mutually exclusive then we have

$$P(\text{return to state } i \,|\text{start at state } i) = P\left(\cup_{n=1}^{\infty} F_{ii}(n)\right) = \sum_{n=1}^{\infty} P\left(F_{ii}(n)\right) = \sum_{n=1}^{\infty} f_{ii}(n)$$

If $\sum_{n=1}^{\infty} f_{ii}(n) = 1$ then we are sure that the process will return to state i at some time in the future.

Recurrent and Transient States

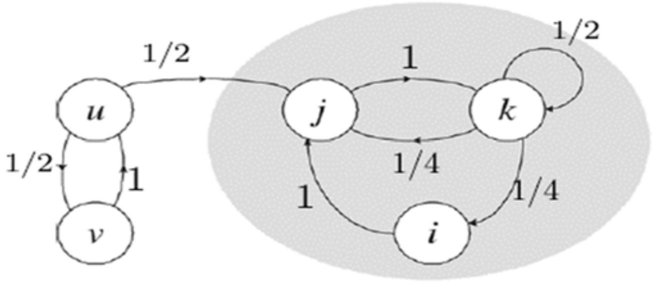
Definition: A state is referred to as ***recurrent*** if and only if $\sum_{n=1}^{\infty} f_{ii}(n) = 1$.

Definition: A Markov Chain for which all the states are recurrent is called a ***recurrent Markov Chain.***

Definition: A state i in a Markov Chain is referred to as ***transient*** if it is not recurrent, i.e. $\sum_{n=1}^{\infty} f_{ii}(n) < 1$.

Example: Transient and Recurrent States

Consider the following Markov Chain



$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/4 & 1/2 & 1/4 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

a) Show that state v is transient.
b) Show that state k is recurrent.

Solution:

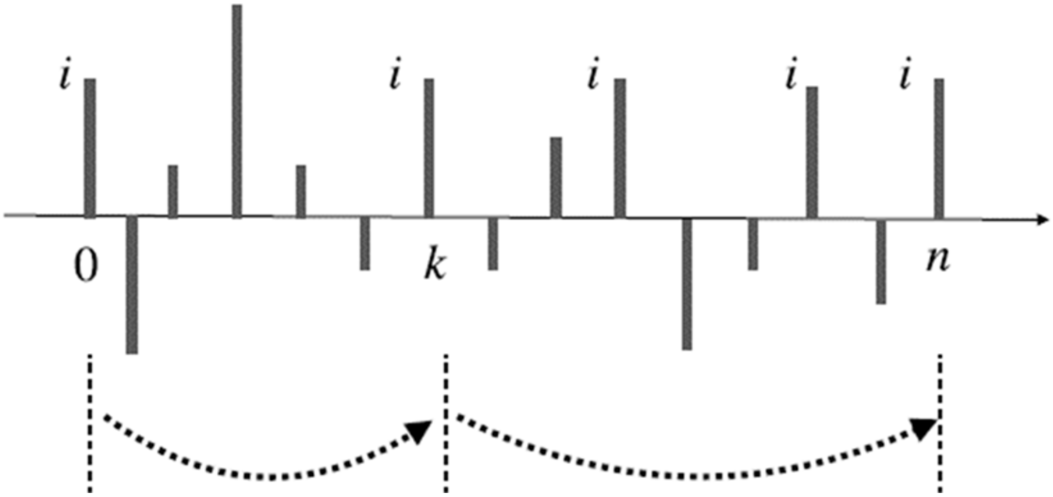
a) We can see that for state v the first return can only occur at $n = 2$, if the process does not return to v at $n = 1$, then it will never return because at $n = 2$ it must be in state j. Hence $f_{vv}(1) = 0, f_{vv}(2) = \frac{1}{2}, f_{vv}(n) = 0, n > 2$. Hence $\sum_{n=1}^{\infty} f_{vv}(n) = \frac{1}{2} < n$. Hence state v is transient.

b) For state k
If the first return occurs at $n = 1$, then it must be through a transition back to state k, hence $f_{kk}(1) = 1/2$
If the first return occurs at $n = 2$, then it must be through transitions $k \to j \to k$, hence $f_{kk}(2) = \frac{1}{4} \cdot 1 = \frac{1}{4}$.
If the first return is at $n = 3$ then it must be through transitions $k \to i \to j \to k$, hence $f_{kk}(3) = \frac{1}{4} \cdot 1 \cdot 1 = \frac{1}{4}$.
It is not possible that the first return occurs for $n > 3$, ie $f_{kk}(n) = 0$ for $n > 3$.

Hence $\sum_{n=1}^{\infty} f_{kk}(n) = \frac{1}{2} + \frac{1}{4} + \frac{1}{4} = 1$. The state k is therefore recurrent.

Relationship Between p_{ii} and f_{ii}

Using the law of total probability we can relate p_{ii} and f_{ii}. If we are going to return to state i at time n, then there is a path from state i to state i which may or may not include visits to state i during intermediate transitions. We compute $p_{ii}(n)$ by conditioning on the time that the first re-visit to state i occurs. The first visit will occur at one of the times $k = 1,2,\cdots,n$.



$$p_{ii}(n) = \sum_{k=1}^{n} P(\text{first visit occurs at time } k)P(\textit{start at state i and end at state i in n-k steps})$$

$$p_{ii}(n) = \sum_{k=1}^{n} f_{ii}(k)p_{ii}(n-k) \qquad n \geq 1$$

Note that if we use $p_{ii}(0) = 1$, then the above can be stated to be valid for $n \geq 0$.
The above expression looks like a convolution of two sequences as usually defined in discrete time systems, i.e. sequences starting at $n = 0$. But for a convolution the above sum needs to start at $k = 0$. Hence we modify the above equation to start the summation at $k = 0$, as follows:

$$p_{ii}(n) = \sum_{k=0}^{n} f_{ii}(k)p_{ii}(n-k) - f_{ii}(0)p_{ii}(n) \qquad n \geq 1$$

Since $f_{ii}(0) = 1$, this can be written as

$$p_{ii}(n) = \frac{1}{2}\sum_{k=0}^{n} f_{ii}(k)p_{ii}(n-k) \qquad n \geq 1$$

To include the case $n = 0$, i.e. make the above valid also for $n = 0$, we add the term $\frac{1}{2}\delta(n)$

$$p_{ii}(n) = \frac{1}{2}\sum_{k=0}^{n} f_{ii}(k)p_{ii}(n-k) + \frac{1}{2}\delta(n) \qquad n \geq 0$$

Hence we use transform techniques which we refer to here as the generating function. This is similar to the use of the Z transforms in discrete time linear systems. These generating functions are also used frequently in combinatorics.

Fix the state i and define the generating function for $p_{ii}(n)$ as

$$P_{ii}(s) = \sum_{n=0}^{\infty} p_{ii}(n)s^n \qquad s < 1$$

In the same manner define the generating function for $f_{ii}(n)$ as

$$F_{ii}(s) = \sum_{n=0}^{\infty} f_{ii}(n)s^n \qquad s \leq 1$$

Taking the generating function of both sides of the above equation involving a convolution, and noting that the generating function for $\delta(n)$ is equal to 1, we obtain the following:

$$P_{ii}(s) = \frac{1}{2}P_{ii}(s)F_{ii}(s) + \frac{1}{2} \qquad s < 1$$

Hence

$$P_{ii}(s) = \frac{1/2}{1 - \frac{1}{2}F_{ii}(s)} = \frac{1}{2 - F_{ii}s)} \qquad s < 1.$$

Proposition

A state i is recurrent if and only if

$$\sum_{n=1}^{\infty} p_{ii}(n) = \infty$$

Suppose that state i is recurrent. Then $\sum_{n=1}^{\infty} f_{ii}(n) = 1$, and since $f_{ii}(0) = 1$, this means $F_{ii}(1) = \sum_{n=0}^{\infty} f_{ii}(n) = 1 + \sum_{n=1}^{\infty} f_{ii}(n) = 1 + 1 = 2$.

But now we can see that it is not possible that $\sum_{n=1}^{\infty} p_{ii}(n) < \infty$, i.e. $P_{ii}(s)$ converges for $s = 1$. Because if it converges then we have $P_{ii}(1) = \frac{1}{2}P_{ii}F_{ii}(1) + \frac{1}{2}$, $P_{ii}(1) = P_{ii}(1) + \frac{1}{2}$, or $0 = 1$. This is a contradiction. So it must be the case that $\sum_{n=1}^{\infty} p_{ii}(n) = \infty$.

As for the converse, suppose that $\sum_{n=1}^{\infty} p_{ii}(n) = \infty$. This means that $P_{ii}(1) = \sum_{n=1}^{\infty} p_{ii}(n) + 1 = \infty$. Now we need to show that $F_{ii}(1) = 2$. But

$$F_{ii}(s) = \frac{2\left(P_{ii}(s) - \frac{1}{2}\right)}{P_{ii}(s)} = 2 - \frac{1}{P_{ii}(s)}$$

and $F_{ii}(1) = 2 - \frac{1}{P_{ii}(1)} = 2$.

What is the Expected Number of Returns to State *i* over all time?

Let $r_n = 1$ if there is a return at step n $(n \geq 1)$, and 0 otherwise. Then the number of returns in the interval $[1, N]$, R_n, is

$$R_n = \sum_{n=1}^{N} r_n$$

Then $\mathcal{E}(R_N) = \sum_{n=1}^{N} \mathcal{E}(r_n) = \sum_{n=1}^{N} p_{ii}(n)$. Now if the Markov Chain is recurrent and we let $N \to \infty$, then this sum is infinite. Hence if the Markov Chain is ***recurrent then the expected number of returns is infinite***.

Limit Theorems for Recurrent Markov Chains

Proposition:

For recurrent irreducible aperiodic Markov Chains we have

$$\lim_{n\to\infty} p_{ji}(n) = \lim_{n\to\infty} p_{ii}(n)$$

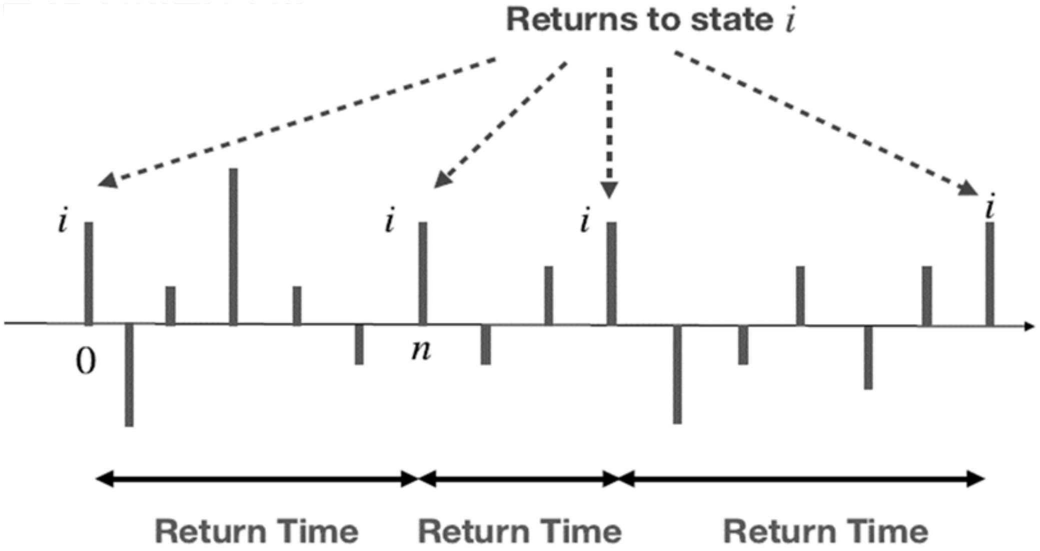
Earlier we saw in an example that for large n the $p_{ji}(n)$ approach a value that is independent of j, i.e. the starting state. This result formalizes that example as applicable to all recurrent irreducible, aperiodic Markov Chains. The state distribution of such chains converges to a constant distribution as $n \to \infty$.

Expected Value of First Return

The expected value of the first return is given by

$$\mu_i = \sum_{m=0}^{\infty} m f_{ii}(m)$$

μ_i small ⇒ visit state i frequencty
μ_i large ⇒ visit state i less frequently.



Proposition:

For recurrent irreducible aperiodic Markov Chains, we have

$$\lim_{n\to\infty} p_{ii}(n) = \frac{1}{\mu_i} = \frac{1}{\sum_{m=0}^{\infty} m f_{ii}(m)}$$

Idea of proof: Consider the times of a sequence of k first returns to state i, $T_i(1), T_i(2), \cdots, T_i(k)$, as shown in the above Figure. Due to the Markov Property these are independent random variables. The random variables $T_i(j)$ are independent and identically distributed with mean μ_i. Since the state is recurrent the Chain returns to the state i an infinite number of times. By the law of large numbers $\lim_{k\to\infty} \frac{1}{k}\sum_{j=1}^{k} T_i(j) = \mu_i$. The proportion of time spent in state i in the first k returns is

$$\frac{k}{T_i(1) + T_i(2) + \cdots + T_i(k)} = \frac{1}{\frac{T_i(1) + T_i(2) + \cdots + T_i(k)}{k}}$$

In the limit as $k \to \infty$ this becomes $\frac{1}{\mu_i}$. In the limit this proportion of time spent in state i is the steady state probability for state i and is equal to $\lim_{n\to\infty} p_{ii}(n)$. Hence we have $\lim_{n\to\infty} p_{ii}(n) = \frac{1}{\mu_i}$.

Positive Recurrent States

If for state i in a recurrent class $\lim_{n\to\infty} p_{ii}(n) = p_i > 0$. Then for any other j in the same class as i, $p_j > 0$. In this case we call the recurrent class, ***positive recurrent.*** **Note that this is the same as** $\mathcal{E}(T_i(j)) < \infty$.

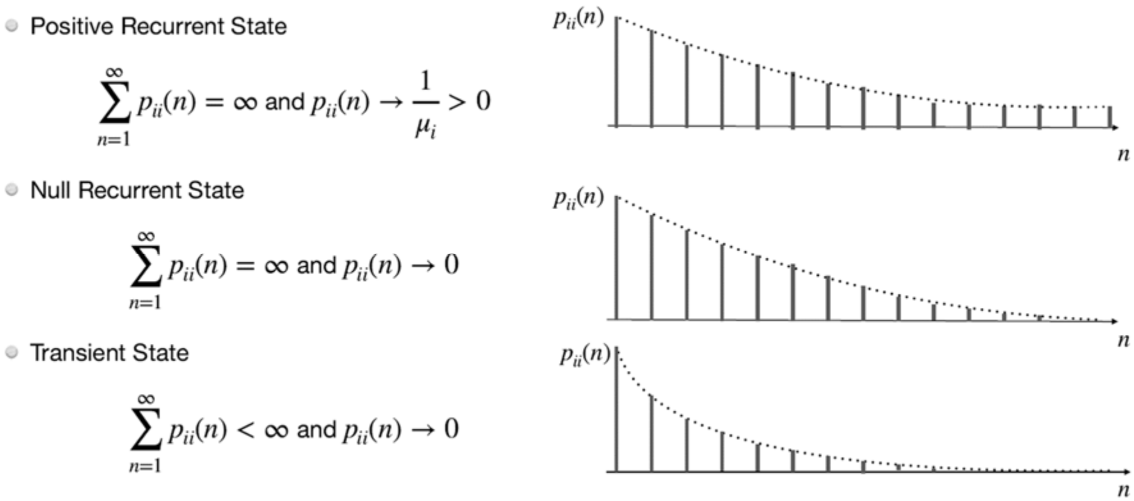
Null Recurrent States

If for an irreducible recurrent Markov Chain class $\mathcal{E}(T_i(j)) = \infty$, then the Markov Chain class is said to be ***null recurrent***.

Finite State Space: Recurrent States

In an irreducible aperiodic Markov Chain with a finite state space, all states are positive recurrent. A unique stationary distribution exists.

Null Recurrent vs. Transient States



- Positive Recurrent State

$$\sum_{n=1}^{\infty} p_{ii}(n) = \infty \text{ and } p_{ii}(n) \to \frac{1}{\mu_i} > 0$$

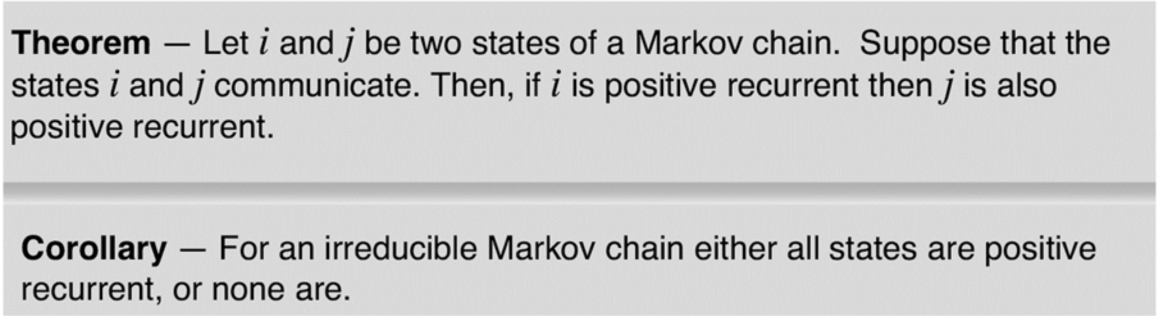
- Null Recurrent State

$$\sum_{n=1}^{\infty} p_{ii}(n) = \infty \text{ and } p_{ii}(n) \to 0$$

- Transient State

$$\sum_{n=1}^{\infty} p_{ii}(n) < \infty \text{ and } p_{ii}(n) \to 0$$

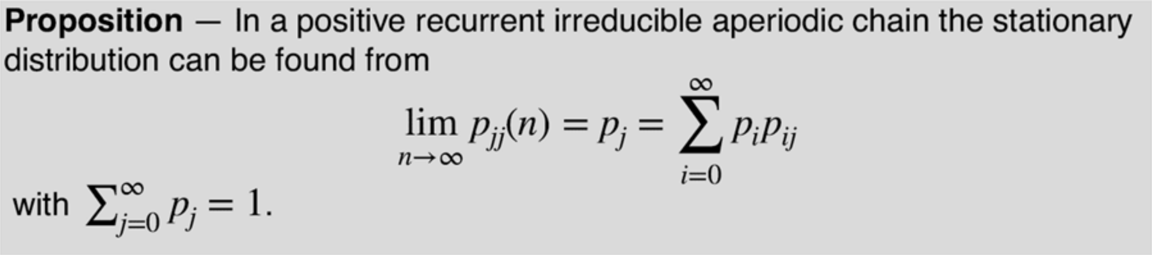
Recurrence of Communicating States



Theorem — Let i and j be two states of a Markov chain. Suppose that the states i and j communicate. Then, if i is positive recurrent then j is also positive recurrent.

Corollary — For an irreducible Markov chain either all states are positive recurrent, or none are.

Positive Recurrent Aperiodic Class

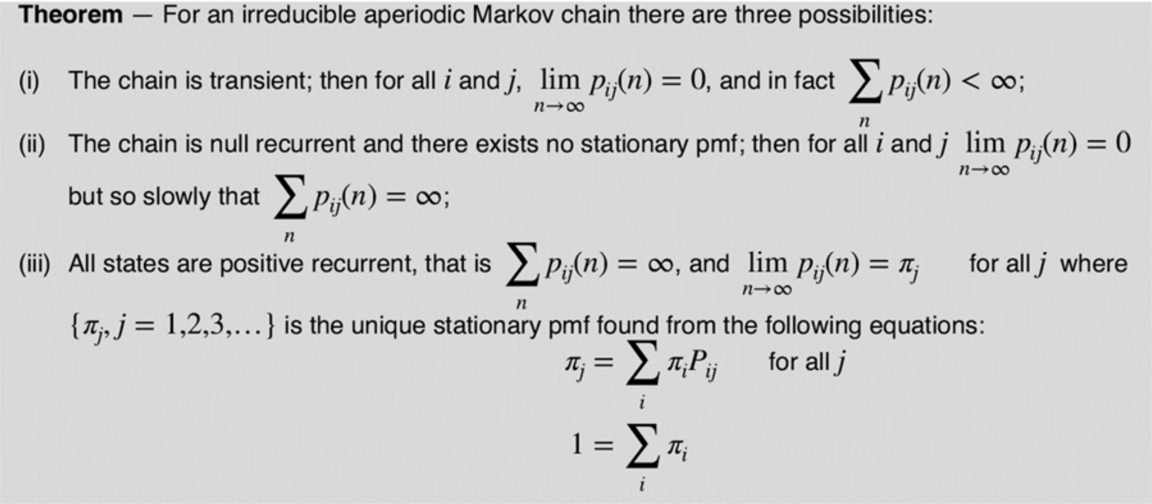


Proposition — In a positive recurrent irreducible aperiodic chain the stationary distribution can be found from

$$\lim_{n\to\infty} p_{jj}(n) = p_j = \sum_{i=0}^{\infty} p_i p_{ij}$$

with $\sum_{j=0}^{\infty} p_j = 1$.

Stationary State Probabilities



Theorem — For an irreducible aperiodic Markov chain there are three possibilities:

(i) The chain is transient; then for all i and j, $\lim_{n\to\infty} p_{ij}(n) = 0$, and in fact $\sum_n p_{ij}(n) < \infty$;

(ii) The chain is null recurrent and there exists no stationary pmf; then for all i and j $\lim_{n\to\infty} p_{ij}(n) = 0$ but so slowly that $\sum_n p_{ij}(n) = \infty$;

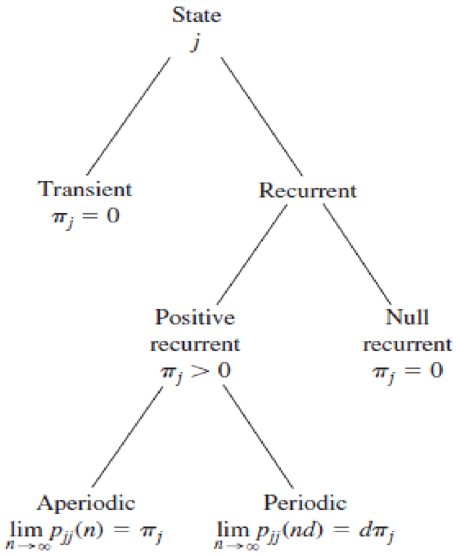
(iii) All states are positive recurrent, that is $\sum_n p_{ij}(n) = \infty$, and $\lim_{n\to\infty} p_{ij}(n) = \pi_j$ for all j where $\{\pi_j, j = 1,2,3,\ldots\}$ is the unique stationary pmf found from the following equations:

$$\pi_j = \sum_i \pi_i P_{ij} \qquad \text{for all } j$$

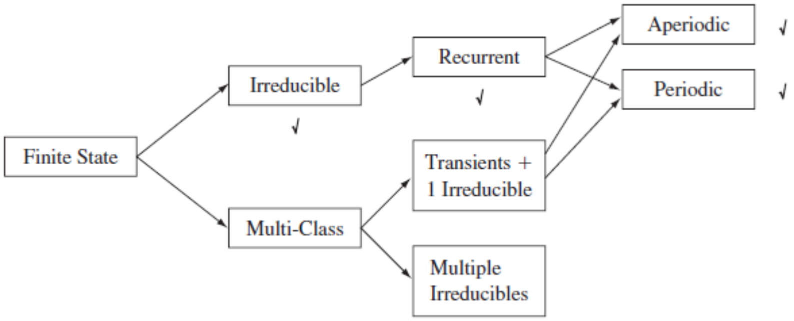
$$1 = \sum_i \pi_i$$

Classification of States and Associated Long Term Behaviour

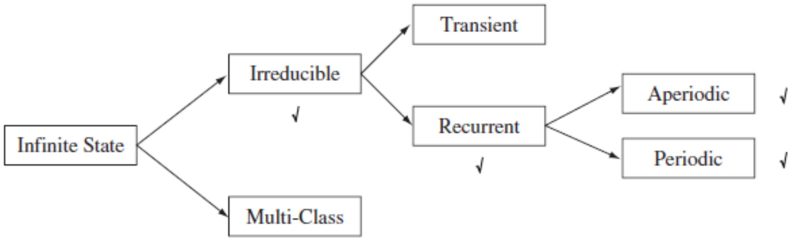
The proportion of time spent in state j is denoted π_j.



Possible Structures for Markov Chains



(a)



(b)