EE507

LECTURE 3: ESTIMATION

We have now completed our study of random variables, and will shortly be beginning our study of random processes (random signals). Before delving into random process theory, I would like to spend a couple lectures to pursue an engineering application of the random-variable theory that we have developed. Specifically, we will focus on the use of random-variable methods for estimation (Section 1).

Section 1: Estimation

Estimation problems come up in a variety of application areas:

1. Given current atmospheric conditions, a meteorologist must predict the weather in Pullman in 3-days time.
2. A university must predict the fraction of accepted students who will actually join.
3. A team of autonomous vehicles must estimate the location of a target, based on their noisy/incomplete observations.

Estimation problems from these many application areas can often be distilled into the following canonical (basic) problems:

Canonical Problem 1: Consider a random variable Y. Given no observations of the random variable,

what is a good estimate $\hat{Y}$ for Y?

Canonical Problem 2

Consider a pair of random variables X and Y. Given an observation of one of the random variables, say X, what is a good estimate for the other random variable Y?

Let us address these two questions, assuming that the the relevant PDFs (the PDF of Y in Problem 1, and the the joint PDF of X and Y in Problem 2) are known to us and hence can be used in designing the estimator. Subsequently, we will briefly consider estimation when aspects of the distribution (joint distribution) are not known.

Solution to canonical problem 1

We are trying to generate a good estimate $\hat{y}$ for a random variable Y with known P.D.F. $f_Y(\alpha)$. In order to decide on a good estimate $\hat{y}$, we need a measure for the performance of the estimator.

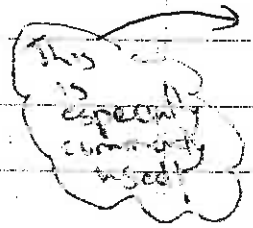
Typically, there are two types of measures considered. First, we may wish to choose $\hat{y}$, so that the P.D.F. $f_Y(\alpha)$ is maximized at $\alpha = \hat{y}$ (i.e., the measure is $f_Y(\alpha)$, and the goal is to maximize the measure):

$$\hat{y} = \max_{\alpha} f_Y(\alpha)$$

This α that ... can be found ... [illegible]

from the PDF.

A second sensible way to measure the performance of an estimator is based on the average distance of the estimate from the actual value of Y, i.e. an expectation of the form $E[g(|Y-\hat{y}|)]$, where $g(\cdot)$ is an increasing function. The most common estimators of this form are the following:



Minimum Mean Square Error (MMSE) Estimator:

$$\hat{y} = \min_{\alpha} E[(Y-\alpha)^2]$$

Minimum Mean Absolute Error (MMAE) Estimator:

$$\hat{y} = \min_{\alpha} E[|Y-\alpha|]$$

We can of course find the MMSE estimate (or the MMAE estimate) by explicitly computing $E[(Y-\alpha)^2]$. However, let's come up with a general expression for the MMSE estimator:

$$\hat{y} = \min_{\alpha} E[(Y-\alpha)^2]$$

Notice that the minimum can only be achieved for α s.t. $\frac{d}{d\alpha} E[(Y-\alpha)^2] = 0$, or as $\alpha \to \pm\infty$.

$$\frac{d}{d\alpha} E[(Y-\alpha)^2] = -2E[Y-\alpha] = 0 \Rightarrow E(Y) = \alpha$$

Also notice that $\frac{d^2}{d\alpha^2} E[(Y-\alpha)^2] = 2$, so this α is the minimum!

Thus, the MMSE estimator is $\boxed{\hat{y} = E[Y]}$. In words, the best estimate for Y is simply the expected value of Y. (Notice that you had actually proved this in a different way in an earlier homework set; do you remember where?)

Notice that $E[(Y-\hat{y})^2] = \text{var}(Y)$ in this case! This indicates the performance of the estimator

It is worth noting that we just need $E[Y]$ to find $\hat{y}$; can we get $E[Y]$ experimentally?

I'll let you find the study the MMAE estimator in a homework set.

CANONICAL PROBLEM 2

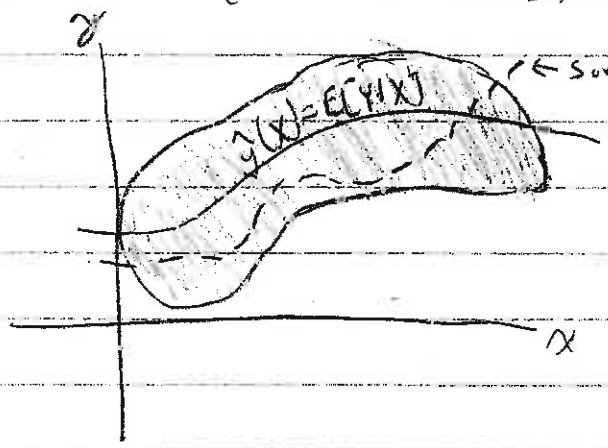
Now consider the case where two random variables X and Y are being considered, and we wish to estimate one, say Y, from the measurement of the other. That is, given that $X=x$, we wish to come up with an estimate $\hat{y}(x)$ for Y.

Let us consider design of the estimator $\hat{y}(x)$ so as to minimize a mean-square-error cost. In this case, notice that we are actually designing a family of estimators: given each possible value of X (i.e., $X=x$), we can choose an estimate $\hat{y}(x)$ (and we have the freedom to choose anything we want). Thus it is reasonable to choose the estimate $\hat{y}(x)$ so that

(*) $$\hat{y}(x) = \min_{\alpha} E\left[(Y-\alpha)^2 \mid X=x\right], \text{ for each } x.$$

To evaluate the estimate, simply notice the following: given $X=x$, we are trying to estimate Y so as to minimize a quadratic cost, where Y has the P.D.F. $f_{Y|X}(y|X=x)$. This is simply an instance of canonical problem 1 so we recover that the estimate is $\hat{y}(x)=E[Y|X=x]$. (Notice that we can also derive this result algebraically from Equation (*) also.)

We have shown that the estimator $\hat{y}(x)=E[Y|X=x]$ minimizes the expected squared estimation error given $X=x$, for all x. In fact, this estimator minimizes the overall mean square error $E[(Y-\hat{y}(X))^2]$, with respect to all possible estimators (functions of X), as diagrammed below:



Claim:

$$E[(Y-E[Y|X])^2] \le E[(Y-\hat{y}_2(X))^2]$$

Proof: Notice that

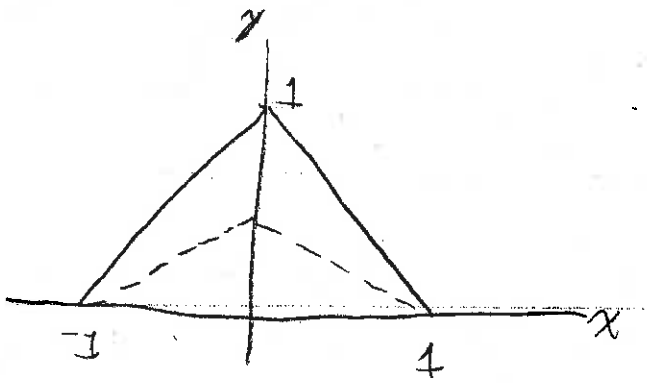
$$E[(Y-\hat{y}(X))^2] = \int_{-\infty}^{\infty} E[(Y-\hat{y}(x))^2|X=x] f_X(x)\,dx$$

This quantity is minimized for each x if the estimator $\hat{y}(x)=E[Y|X=x]$ is used, and so the integral over all x is smaller than for any other function $\hat{y}(x)$.

Let us do a couple examples:

Example 1:

Let's say that X and Y are uniformly distributed in the triangular region shown below. Please find the best estimate for Y given $X=x$:

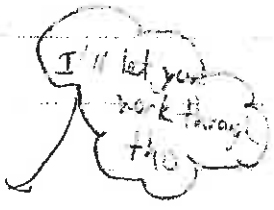


Notice that $$E[Y|X=x] = \begin{cases} \frac{1-x}{2}, & 0 \le x \le 1 \\ \frac{x+1}{2}, & -1 \le x \le 0 \end{cases}$$

This is the best estimator $f(x)$ (in a mean square sense) for Y given $X=x$.

- Notice that the expected error in the estimate given $X=x$ is $E[(Y-E[Y|X])^2 | X=x]$. This is simply the variance of Y given $X=x$, what is the variance of a random variable that is uniform on $[0, 1-x]$ for $0 \le x \le 1$, and uniform on $[0, 1+x]$ for $-1 \le x \le 0$.

Thus we obtain that the expected error is $$\frac{(1-x)^2}{12}, \quad 0 \le x \le 1 \quad \text{and} \quad \frac{(1+x)^2}{12}, \quad -1 \le x \le 0.$$



The overall average error can be found as $\int_{-\infty}^{\infty} E[(Y-E[Y|X])^2 | X=x] f_X(x)\, dx$.

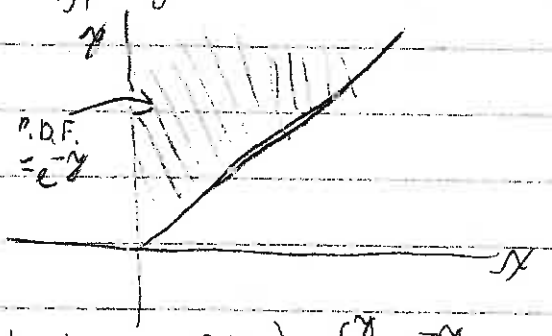
EXAMPLE 2

The position X of a target is an exponential random variable with mean 1. We cannot measure X directly, and instead measure $Y = X + Z$, where Z is exp. w/ mean 1. What is the MMSE estimate for X given $Y = y$?

The best estimate is $\hat{x}(y) = E[X|Y=y]$.

Note that $f_X(x) = e^{-x}$, $x \geq 0$, while $f_{Y|X}(y|X=x) = e^{-(y-x)}$, $y \geq x$.

Thus, $f_{X,Y}(x,y) = e^{-x} e^{-(y-x)} = e^{-y}$, $0 \leq x \leq \infty$, $x \leq y \leq \infty$



Thus, note that $f_Y(y) = \int_0^y e^{-y}\, dx = y e^{-y}$, $0 \leq y < \infty$.

Thus, $f_{X|Y}(x|Y=y) = \frac{e^{-y}}{y e^{-y}}$, $0 \leq x \leq y$

$f_{X|Y}(x|Y=y) = \frac{1}{y}$, $0 \leq x \leq y$, i.e. X is uniform between 0 and y.

Thus, $\hat{x}(y) = E(X|Y=y) = \frac{y}{2}$

This estimate makes sense, since we are in essence repeat measuring two times X.

Structure of the MMSE Estimator

We reiterate that the best estimate for Y given $X=x$ (in a mean square sense) is $\hat{y}(x)=E[Y|X=x]$. This estimator has a special structure. Let us first present the structural result, and then interpret the result:

Claim: $E[(Y-\hat{y}(X))g(X)]=0$, for any $g(X)$.

Proof:

$$E[(Y-\hat{y}(X))g(X)]$$
$$= E[E[(Y-\hat{y}(X))g(X)|X]]$$
$$= E[g(X)E[Y-\hat{y}(X)|X]]$$
$$= E[g(X)(E[Y|X]-\hat{y}(X))]$$
$$= E[g(X)(E[Y|X]-E[Y|X])] = 0.$$

Interpretation: $Y-\hat{y}(X)$, or in other words the error between Y and its estimate, is uncorrelated with any function of X. This is a sensible and appealing result, in that it suggests that any other function of X cannot give further information on the value of the error $Y-\hat{y}(X)$, and hence the error cannot be reduced by another estimator.

Linear Minimum Mean Square Error (LMMSE) Estimation

Although the MMSE estimator is ideal in terms of minimizing a reasonable cost metric, the MMSE estimator may be difficult to implement for a couple reasons:

1. The conditional expectation for one variable given the other must be found, which means that the joint distribution of the two random variables (or appropriate conditional distributions) must be known. An estimator that only required knowledge of the statistics of X and Y ($m_X, m_Y, \sigma_X^2, \sigma_Y^2, \text{cov}(X,Y)$) would potentially be easier to use.

2. The MMSE estimator is nonlinear. In situations where the estimator is being used as part of a larger system or must be implemented in hardware, nonlinear mappings may be hard to analyze/implement.

These motivations persuade us to try linear estimators, i.e. estimators that are constrained to be of the form $\hat{y}(x) = ax + b$. (It is clear that such estimators address motivation 2, and we'll see that they address the first motivation also.) If we constrain the estimator to be linear, we cannot hope to minimize the expected squared error given each value of x. Instead,

Let us design the estimator to minimize the (overall) mean square error. That is, let us find the estimator

$$\hat{y}(x) = a^* x + b^*, \quad \text{where}$$

$$a^*, b^* = \arg\min_{a,b} E[(Y - aX - b)^2]$$

At the minimum, $\frac{d}{da} E[(Y-aX-b)^2] = 0$ and $\frac{d}{db} E[(Y-aX-b)^2] = 0$.

(It's easy to rule out the extreme cases.)

From $\frac{d}{da} E[(Y-aX-b)^2] = 0$, we get that

$$E[X(Y-aX-b)] = 0 \Rightarrow E(XY) = aE(X^2) + bE(X) \quad (*)$$

From $\frac{d}{db} E[(Y-aX-b)^2] = 0$, we get $E(Y) = aE(X) + b$ $(**)$

Multiplying $(**)$ by $E(X)$ and subtracting from $(*)$, we get

$$E(XY) - E(X)E(Y) = a\left(E(X^2) - (E(X))^2\right)$$

$$\Rightarrow \boxed{a = \frac{cov(X,Y)}{var(X)}} = \frac{\rho_{X,Y}\sqrt{var(Y)}}{\sqrt{var(X)}}$$

$$b = E(Y) - aE(X) = \boxed{E(Y) - E(X) \cdot \frac{cov(X,Y)}{var(X)}}$$

Thus, the LMMSE estimator is

$$\boxed{\hat{y}(x) = \frac{cov(X,Y)}{var(X)} x + \left(E(Y) - E(X) \cdot \frac{cov(X,Y)}{var(X)}\right)}$$

Like the MMSE estimator, the LMMSE estimator has a special structure. Specifically, we find that $E[(Y-\hat{y}(X))X]=0$, i.e. the estimation error is uncorrelated with X. This uncorrelation can be viewed as a sort of orthogonality: the error is perpendicular to the random variable X, so our estimator has computed all the part of Y that can be obtained from X. I'll leave it to you to go through the details of the analysis.

It is also worth noting that estimation of one random variable from another one is sometimes based on a different performance measure. Another common goal is to maximize $f_{Y|X}(y|X=x)$ to estimate Y from a measurement of X.

I will give out another handout to address the case where the PDF of a "hidden variable" is unknown. This technique for "non-random estimation" complements the techniques described here.

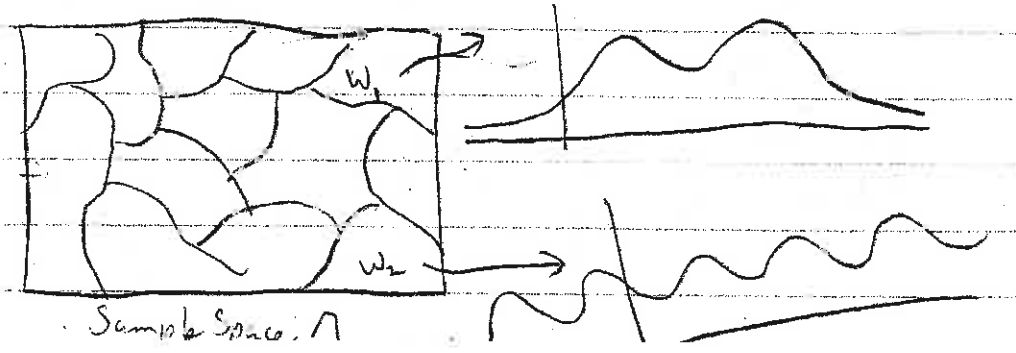
1

LECTURE 4: INTRO. TO RANDOM PROCESSES (PART 1)

Uncertainties play a critical role in a range of engineering applications. Very often, uncertainties (noise) impact a system over a duration of time. Thus, we are motivated to study random processes (random functions of time, random signals). In the next few weeks, we will delve into the theory of random processes. In this first lecture on random processes, we will define random processes, describe their basic analysis, and describe some common types of random processes.

Section 1: Definition of a Random Process

Consider an uncertain experiment with sample space Ω. A random process is an association of a time function with each outcome of the experiment, i.e. a mapping between the outcomes of an experiment and time functions. That is, a random process $x(t)$ is a mapping $x(t, \omega)$ from every outcome $\omega \in \Omega$ to a time signal. Here's an illustration:

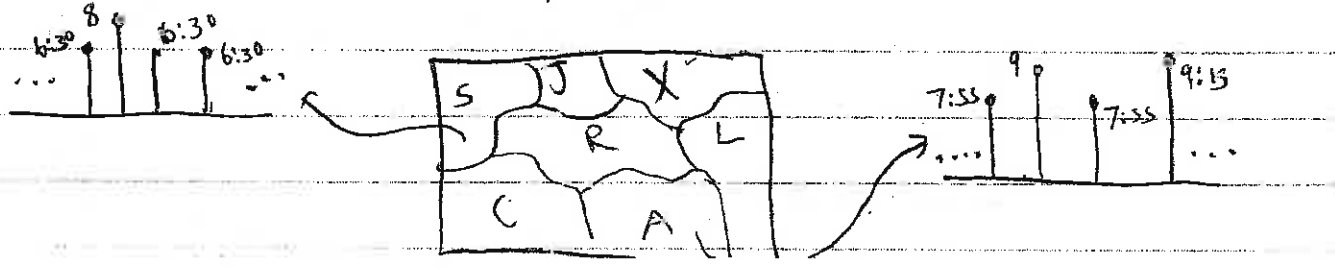


We can equivalently view a random process as a mapping from both outcomes and time to real numbers.

The domain of the time variable t for a random process is a subset of the real numbers. Commonly, we assume $t \in \mathbb{R}$ or $t \in \mathbb{R}^+$, i.e. t is any real number or any positive real number, in which case we call the random process a <u>continuous-time process</u>. Alternately, we consider $t \in \mathbb{Z}$ or $t \in \mathbb{Z}^+$, i.e. t is any integer or any positive integer, in which the random process is called a <u>discrete-time process</u>. A DT Random process is also known as a sequence of random variables

Let us give some examples of random processes:

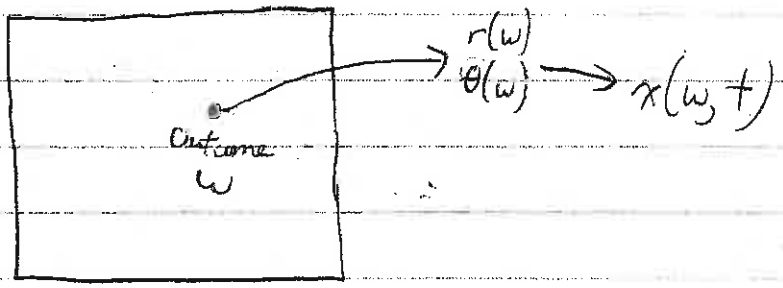
1. Consider the experiment where we pick a person at random in this class, and consider the sequence of times at which the person wakes up. (for simplicity, assume that the sequence is deterministic once the person is chosen). This sequence of time is a DT (discrete-time) random process: the experiment can be viewed as having each person in the class as an outcome, and each outcome maps to a sequence.



3

Example 2:

Consider the signal $x(t) = r\cos(\omega t + \theta)$, where r is Gaussian with mean 1 and variance 0.1, and θ is uniform on $[0, 4\pi)$, (and ω is fixed). Notice that this is a sinusoidal signal with random amplitude and phase. We claim that $x(t)$ is a continuous time random process. In particular, we note that the random variables r and θ can be generated from an uncertain experiment (for instance, the dartboard experiment). In turn, the signal $x(t)$ can be generated from these two random variables. Thus, each outcome of the experiment maps to a continuous-time function, and hence $x(t)$ is a continuous-time random processes. Here's an illustration:



Example 3

We toss a fair coin over and over. Every time the toss shows heads, we write down '1', and every time the toss shows tails, we write down '0'. (For instance, HHTHTTH... ⇒ 1,1,0,1,0,0,1...). We claim that this process is a discrete time random process. In particular, consider an experiment with

two outcomes, H and T. Then we can generate another experiment whose sample space is the Cartesian products of the sample spaces of the individual experiments, i.e. $\Omega_{whole} = \Omega_1 \times \Omega_1 \times \ldots$. Each outcome of this experiment maps to one possible sequence, and hence the sequence is a discrete time random process.

It is worth noting that multiple random processes can be defined for an experiment.

---

2. Outline

In the remainder of this lecture, we will present the basic analysis of random processes, and introduce several special types of random processes. In particular, we will pursue the following tasks:

1. We will develop useful techniques for analyzing the temporal characteristics of random processes (Part 1).
2. We will analyze, in particular, the asymptotics of random processes (Part 2).
3. We will introduce some interesting classes of random processes (Part 3).

5

3. Temporal Characteristics of Random Processes

When we first introduced random variables, we motivated the CDF (and in turn the PDF) as tools for characterizing the probabilities that random variables take on particular values. Similarly, we are interested in the value(s) that a random process takes at one or several times.

Notice that a random process evaluated at a particular time is simply a random variable, and hence we can study the value(s) taken on by a random process at one or more times by thinking about the (joint) distributions of random variables. These joint distributions of the random process evaluated at particular times, together with associated statistics, are the important descriptions of random process. Let us begin by introducing the joint distributions, and then delve into interesting statistics.

First-order Distribution/Density

Consider a random process $X(t) = X(\omega, t)$. At the most basic level, we may wish to characterize the values taken on by $X(t)$ at a single time $t = t_1$. With this goal in mind, we are interested in the first-order cumulative distribution function (or simply first-order density) $F(x; t_1) \triangleq P(X(t_1) \leq x)$. That is,

the first-order distribution is the function (of x and t) that captures the individual CDFs for $X(t)$ evaluated at each time $t=t_1$. Naturally, we define the <u>first-order PDF</u> or simply <u>first-order density</u> as the derivative of the first order CDF with respect to its argument x: $f(x,t) = \frac{\partial F(x,t)}{\partial x}$. Let us do an example to think about how the first-order distribution/density can be found/used.

<u>Example</u>

Consider the random process $X(t) = \cos(t+\theta)$, where θ is uniform on $[0, 2\pi)$. Let us find the first-order distribution. This is

$$F_X(x,t) = P(X(t) \le x) = P(\cos(t+\theta) \le x)$$
$$= P(\cos^{-1}(x) \le \underbrace{(t+\theta) \bmod (2\pi)}_{\text{this is uniform on } [0, 2\pi]} \le 2\pi - \cos^{-1}(x))$$

Thus, $F_X(x,t) = \frac{2\pi - 2\cos^{-1}(x)}{2\pi}$, $-1 \le x \le 1$

$$F_X(x,t) = \begin{cases} 1 - \frac{1}{\pi}\cos^{-1}(x), & -1 \le x < 1 \\ 1, & x \ge 1 \\ 0, & x < -1 \end{cases}$$

Thus, we find that the first-order density is

$$f_X(x,t) = \frac{d}{dx} F_X(x,t) = \begin{cases} \frac{1}{\pi} \frac{1}{\sqrt{1-x^2}}, & -1 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

7

Since you know that the marginal PDF of two R.V.s doesn't specify the joint PDF, you will not be surprised to hear that the first-order density/distribution is not a full probabilistic description of the process. For instance, consider the following:

Example

Consider the following two discrete-time random processes (specifically, random processes with $t = 1, 2, 3, \ldots$):

$X(t)$ is generated by repeatedly tossing a fair coin; $X(n)$ is set to 1 if the n^{th} coin toss shows heads, and 0 if it shows tails.	$Y(t)$ is generated by tossing a fair coin once; $Y(n)$ is is set to 1 for all n if the coin toss shows Heads, and is set to 0 for all n if the coin toss shows tails.

The two processes have the same first-order COF/PDF/PMF ($X(t) = 0$ w.p. $\frac{1}{2}$, $X(t) = 1$ w.p. $\frac{1}{2}$; similarly, $Y(t) = 0$ w.p. $\frac{1}{2}$, $Y(t) = 1$ w.p. $\frac{1}{2}$; for all t). However, $P(X(2) < X(4)) = \frac{1}{4}$, while $P(Y(2) < Y(4)) = 0$, hence the processes have not been fully characterized.

Second- and Higher Order Distributions/Densities

The second-order distribution for a random process $X(t)$ is the (family of) joint distributions $F(x_1, x_2; t_1, t_2) = P(X(t_1) \le x_1, X(t_2) \le x_2)$. The second order density (PDF) is $f(x_1, x_2; t_1, t_2) \triangleq \frac{\partial^2}{\partial x_1 \partial x_2} F(x_1, x_2; t_1, t_2)$.

Similarly, the n^{th}-order distribution for a random process $X(t)$ is $F(x_1, \ldots, x_n; t_1, \ldots, t_n) \triangleq P(X(t_1) \le x_1, \ldots, X(t_n) \le x_n)$. The n^{th}-order density can be defined from there. For most processes (and almost all useful processes), the statistical properties of the process are fully known if we know the n^{th} order distribution for arbitrary n.

Example

Consider a discrete-time process where each $x(t)$, $t = 1, 2, 3, \ldots$, is an independent exponential random variable with mean t_i. For this process, let us find the n^{th}-order joint PDF $f(x_1, \ldots, x_n; t_1, \ldots, t_n)$, where w.l.o.g. $t_1 \neq t_2 \neq \ldots \neq t_n$. Exploiting independence, we immediately obtain that

$$f(x_1, \ldots, x_n; t_1, \ldots, t_n) = \prod_{i=1}^{n} \frac{1}{t_i} e^{-x_i/t_i}$$

Although the n^{th} order joint distribution was straightforward to obtain in the example above, these joint distributions in general may be quite difficult to obtain; often, we require special structure in the description of the process to gain meaningful insight into the cross-time behaviors of random processes. This is one reason that we shall spend much

effort to understand some special classes of random processes.

In discussing n^{th} order joint distributions, it is worth noting that the notions of conditioning that we developed for models and in turn random variables are applicable here.

Statistics of Random Processes

Sometimes it is too ambitious to obtain the n^{th} order joint distributions or even the 1^{st} order distribution especially when we must characterize the process from some data. This motivates us to think about statistics that a random process will tend to at one or a small number of times. Of particular interest are first- and second-order statistics of the random process. Here, we define the statistics of interest; later, we shall show that much about a random process is captured by these few ideal statistics.

Mean The mean $m(t)$ of a random process $X(t)$ is a time function capturing the expected value vs. t:

$$m(t) = E[X(t)] = \int_{-\infty}^{\infty} x f(x,t)\, dx.$$

The mean is the first order statistic of the random process's name that it can be obtained from the 1^{st} order distribution of $X(t)$.

The autocorrelation captures the expected value of products of $X(t)$ at two times $t=t_1$ and $t=t_2$, and so tells us something about how values of the random process at a couple times relate to each other. Specifically, the autocorrelation of $X(t)$ is

$$R(t_1,t_2) = E[X(t_1)X(t_2)]$$

As always, we often wish to adjust the random variables to have zero mean in analyzing their relationship. This motivates us to define the autocovariance

$$\begin{aligned} C(t_1,t_2) &= E[(X(t_1)-m(t_1))(X(t_2)-m(t_2))] \\ &= R(t_1,t_2) - m(t_1)m(t_2). \end{aligned}$$

It's worth noting that $R(t,t) = E[X^2(t)]$, which is the average instantaneous power of the signal at time t, while $C(t,t) = \text{var}(X(t))$, i.e. the variance of $X(t)$ at time t.

Example (10-4 in text)

Consider the process

$x(t) = R\cos(\omega t + \theta)$, where ω is constant, R is a random variable with known mean and second moment, and θ is uniform on $(-\pi,\pi)$ and independent of R. Please find the mean, autocorrelation, and autocovariance of $x(t)$.

$$E[x(t)] = E\left(R\cos(\omega t+\theta)\right)$$

$$E[x(t)] = \int_r \int_{\theta=0}^{2\pi} r\cos(\omega t+\theta_1)\, d\theta_1\, dr$$

integral of a cosine wave over one period, equals 0.

$$\boxed{E[x(t)] = 0}$$

$$R_{xx}(t_1,t_2) = E[x(t_1)\,x(t_2)]$$

$$= E\left[R\cos(\omega t_1+\theta)\cdot R\cos(\omega t_2+\theta)\right]$$

$$= E\left[R^2\ \frac{\cos(\omega t_1-\omega t_2)+\cos(\omega(t_1+t_2)+\theta)}{2}\right]$$

$$= E\left[\frac{R^2\cos(\omega t_1-\omega t_2)}{2}\right] + 0 \leftarrow \text{since integrates to cosine over period}$$

$$\boxed{R_{xx}(t_1,t_2) = \tfrac{1}{2}E[R^2]\cos(\omega(t_1-t_2))}$$

$$C_{xx}(t_1,t_2) = R_{xx}(t_1,t_2) - E^2[x(t)]$$

$$\boxed{C_{xx}(t_1,t_2) = \tfrac{1}{2}E[R^2]\cos(\omega(t_1-t_2))}$$

The autocorrelation function has some special properties, which are useful in better understanding the temporal dynamics of random processes. Perhaps most importantly, we might expect that the diagonal values of $R_{xx}(t_1,t_2)$ (i.e. the values

when $t_1 = t_2$) are somehow large in comparison to the off-diagonal values (since e.g. $E[QR] \leq \sqrt{E(Q^2)E(R^2)}$ or similarly $\text{cov}(Q,R) \leq \sqrt{\text{var}(Q)\text{var}(R)}$). In fact, we can formalize this idea. In particular, consider the random process $X(t)$ evaluated at n times $t_1, \ldots, t_n$, and consider the matrix

$$R = \begin{bmatrix} R_{xx}(t_1,t_1) & \cdots & R_{xx}(t_1,t_n) \\ \vdots & & \vdots \\ R_{xx}(t_n,t_1) & \cdots & R_{xx}(t_n,t_n) \end{bmatrix}$$

The matrix R is <u>positive semidefinite</u> i.e. $\vec{a}^T R \vec{a} \geq 0$ for all $\vec{a} \neq 0$. In other words,

$\sum_i \sum_j a_i a_j R_{xx}(t_i,t_j) \geq 0$, which is a reflection of the fact that the diagonal of $R_{xx}(t,\tau)$ is dominant.

<u>Proof</u>

Note that $E\left[\left(\sum_{i=1}^{n} a_i X(t_i)\right)^2\right] \geq 0$, and thus

$$E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j X(t_i) X(t_j)\right] \geq 0$$

$$\Rightarrow \boxed{\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j R_{xx}(t_i,t_j) \geq 0}$$

It's worth noting that, for continuous time processes, the inequality $\int_t \int_\tau a(t)\, a(\tau)\, R_{xx}(t,\tau)\, d\tau\, dt \geq 0$ also holds.

It is worth noting that higher moments can also be defined for a random process, but are not in general useful.

13

We have now given a full probabilistic and statistical description of a pair of random processes. Before continuing on to the asymptotics of random processes, let us think briefly about pairs of random processes.

Notice that a pair of random processes $X(t)$ and $Y(t)$ are simply two random processes defined from the same uncertain experiment. A full probabilistic study of the pair of processes requires us to consider the joint CDFs/PDFs of both random processes evaluated at sets of time, i.e.

$$F(x_1, \ldots, x_n; y_1, \ldots, y_m; t_1, \ldots, t_n, t_1', \ldots, t_m')$$

$$\triangleq P(X(t_1) \leq x_1, \ldots, X(t_n) \leq x_n, Y(t_1') \leq y_1, \ldots, Y(t_m') \leq y_m)$$

for each n, m, and $t_1, \ldots, t_n, t_1', \ldots, t_m'$.

Such joint CDFs are usually a big pain to find/use, so instead we often focus on statistics, and in particular second-order ones. Of particular interest, we define the cross-correlation of $X(t)$ and $Y(t)$ as

$$R_{XY}(t_1, t_2) = E[x(t_1) y(t_2)], \text{ i.e. it}$$

captures the correlation of X at one time t_1 with Y at another time t_2. Similarly, the cross covariance is defined as

$$C_{XY}(t_1, t_2) = R_{XY}(t_1, t_2) - m_X(t_1) m_Y(t_2)$$

Example:

4. Asymptotics of Random Processes

In the previous section, we have given a general probabilistic and statistical description of a random process. When studying random processes, we sometimes do not need to concern ourselves with this full description, and only wish to understand the asymptotics of the process (the dynamics when the time variable becomes large).

For the sake of convenience, we will limit our study of asymptotics to processes with $t \in \mathbb{Z}^+$, i.e. sequences of random variables $X(0), X(1), X(2), \ldots$ (We'll sometimes also use the notation $X_0, X_1, X_2, \ldots$)

However, the definitions and results presented here can readily be adapted to the continuous-time case.

(Continued in Part 2)

4. Asymptotics of Random Processes

In the previous section, we have given a general probabilistic and statistical description of a random process. When studying random processes, we sometimes do not need to concern ourselves with this full description, and only wish to understand the asymptotics of the process (the dynamics when the time variable becomes large).

For the sake of convenience, we will limit our study of asymptotics to processes with $t \in \mathbb{Z}^+$, i.e. sequences of random variables $X(0), X(1), X(2), \ldots$ (we'll sometimes also use the notation $X_0, X_1, X_2, \ldots$). However, the definitions and results present here can readily be adapted to the continuous-time case.

(Continued in Part 2)

Let's begin our discussion of convergence by reviewing the notion of convergence for deterministic signals. Recall that convergence of sequences is important when we are trying to evaluate the steady-state value of a signal. Formally, we say that a sequence of numbers $x(1), x(2), \ldots$ converges to a number x, if for every ε, there exists n such that

$$|x(k) - x| \leq \varepsilon \text{ for all } k \geq n.$$

Let us now develop several notions of stochastic convergence. Our first idea is an obvious generalization of the deterministic notion of convergence. We say that the random sequence converges everywhere (meaning for all experiment outcomes) to x, if $X(w, k)$ converges to x for every w. That is, the sequence converges everywhere, if for every possible outcome of the experiment, the sequence converges to x.

Convergence everywhere seems like a sensible notion, but in fact it is usually way too stringent. Let us describe one example of a random process that common sense says is convergent, but is not convergent everywhere.

The example before suggests that we should permit the sequences for some initial conditions to not converge, as long as such sequences have probability 0 (notice this is possible because single outcomes have zero probability). This leads us to the following definition of convergence almost everywhere.

Let A be the event consisting of all outcomes ω such that $X(\omega,t)$ converges to x. If the probability of the event A is 1, then the random sequence is said to converge almost everywhere to x. In short, $X(\omega,t)$ converges almost everywhere to x if $P(\{X(t)\to x\}) = 1$.

Convergence almost everywhere is also known as convergence with probability 1.

Convergence almost everywhere turns out to be a quite sensible notion, but is sometimes can be difficult to test because it requires knowledge of the joint density of $X(t)$ over many times t. Also, we sometimes do not care whether entire sequences converge, but simply want to know whether the probability that the random process is near x gets bigger at larger times. We are thus motivated to pursue yet one more definition of convergence. A random sequence is said to converge in probability to x if $P(|X(n)-x|>\varepsilon)\to 0$ as $n\to\infty$ for all ε. That is, $X(n)$

converges in probability to x, if the probability that $X(n)$ differs from x by more than ε at time n approaches zero as n approaches infinity, for any ε. We notice that convergence in probability only requires knowledge of the first-order distribution of $X(n)$, and in no way implies that the entire sequence converges for any outcome in the sample space.

We can easily see that convergence almost everywhere implies convergence in probability. Do you see why?

Also, let us give an example which shows that convergence in probability does not imply convergence with probability 1.

As we discussed in Part A, we sometimes may only be able to, or only wish to, characterize moments of a random process. In fact, first- and second-moment information are sufficient to meaningfully define the convergence of a random sequence: we can think of a sequence as converging if the spreads of the sequence values around a point decrease with increasing t. In particular, we say that $X(t)$ converges to x in a mean square sense, if

$$E[(X(t)-x)^2] \to 0 \quad \text{as } t \to \infty.$$

Let us check for mean square convergence in an example:

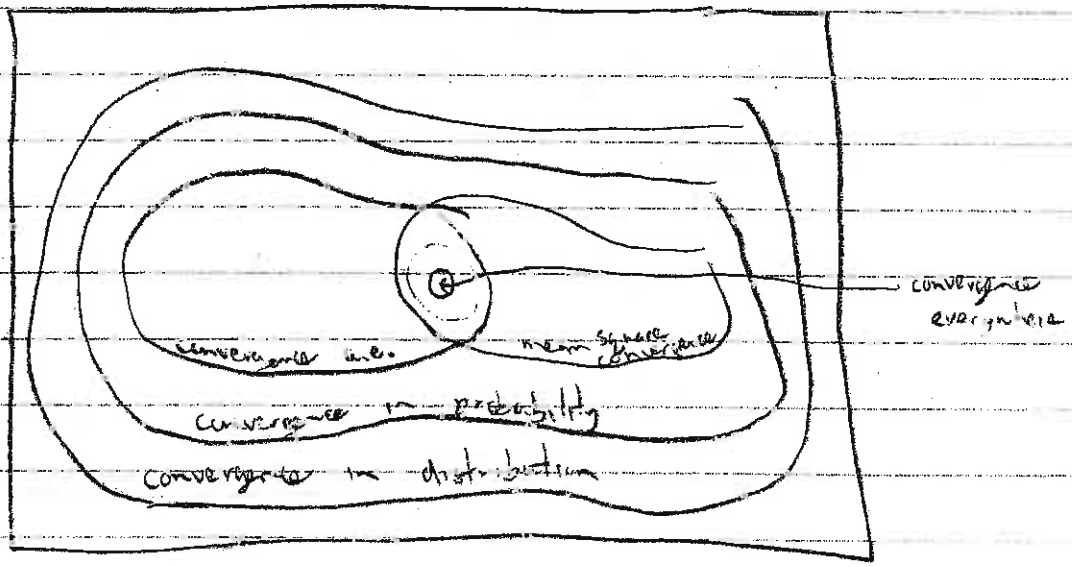
We note that mean square convergence is quite a strong notion of convergence, in that it implies convergence in probability. This can be proved using Chebyshev's inequality:

It's also worth noting that mean square convergence neither implies nor is implied by convergence almost everywhere. You will do some examples verifying this idea in your homework.

The condition for mean square convergence can sometimes be difficult to test, because we may not know the limit point to which the sequence converges. The Cauchy criterion (see your text) yields a simpler test for convergence. In particular, applying it to the mean square convergence test, we obtain that if $E[|x_{n+m} - x_n|^2] \to 0$ as $n \to \infty$ for all $m > 0$, then the sequence converges in a mean square sense. Similar tests for the other notions of convergence that do not depend on the limit point can be developed (though this idea is used most often in checking mean-square convergence).

Finally, there are many occasions where we a random process does not converge to a value, but the first-order CDF of the random process does converge to a fixed function, i.e. $F(x,n) \to F(x)$ for all x as $n \to \infty$. Convergence in distribution is a much weaker notion of convergence and is applicable to any problem where the CDF/PDF reach a steady-state with time. We note that all the other notions of convergence automatically imply convergence in distribution.

To summarize, let us present a Venn diagram relating the notions of convergence (Figure 8-3 in your text):



Convergence of Some Special Sequences

There are some special sequences that we encounter over and over, and so their convergence is of special interest. In these cases, analysts have gained great insight into the sequences' convergence; let us take a moment to study these special cases.

Special Case 1

We can pose the problem of finding the fraction of trials of an experiment that are successful in terms of random processes. In particular, let us consider a Bernoulli process X_i,

i.e. a process where each X_i is an independent random variable that equals 1 with probability p and equals 0 otherwise. We note that $X_i = 1$ can be interpreted as a success on i^{th} trial, and $X_i = 0$ a failure. We can thus interpret the random process

$$Y_i = \frac{\sum_{j=1}^{i} X_j}{i}$$

as representing the fraction of trials that are successes.

We claim that Y_i approaches p in probability. This result is known as the <u>law of large numbers</u>.

Proof:

Note that $E[Y_i] = \frac{E[\sum_{j=1}^{i} X_j]}{i} = \frac{\sum_{j=1}^{i} E(X_j)}{i} = \frac{pi}{i} = p.$

Thus, from Chebyshev's inequality, we know that $P(|Y_i - p| < \varepsilon) \geq 1 - \frac{\text{var}(Y_i)}{\varepsilon^2}$

$$\text{var}(Y_i) = \text{var}\left(\frac{\sum_{j=1}^{i} X_j}{i}\right) = \frac{1}{i^2} \sum_{j=1}^{i} \text{var}(X_j) = \frac{1}{i^2}(ip(1-p)) = \frac{p(1-p)}{i}$$

↑ Since independent

Thus, $P(|Y_i - p| < \varepsilon) \geq 1 - \frac{p(1-p)}{i\varepsilon^2} \to 1$ as $i \to \infty$, for each ε.

Thus, we have proved convergence in probability. □

In fact, it can be shown that Y_i approaches p not only in probability but almost everywhere. This

stronger result is known as the strong law of large numbers. We will not take the time to prove the strong law here.

Special Case 2

Consider a sequence of independent random variables $X_1, X_2, \ldots$, and consider the random variable $S = X_1 + X_2 + \ldots + X_n$ (i.e., the sum of the first n elements of this sequence). Notice that $E[S] = \sum_{i=1}^{n} E[X_i]$, while $\text{var}(S) = \sum_{i=1}^{n} \text{var}(X_i)$.

As long as the following conditions hold:

1. $\sigma_1^2 + \ldots + \sigma_n^2 \xrightarrow[n\to\infty]{} \infty$
2. $\int_{-\infty}^{\infty} x^{\alpha} f_i(x)\,dx < K < \infty$ for some $\alpha > 2$ and some K.

the CDF of the R.V. S approaches the CDF of a Gaussian random variable with mean $E[S]$ and variance $\text{var}(S)$.

Proof: the proof is rather messy and complicated, so I'll skip it for now.

LECTURE 4, PART 3

In this third part of our introduction to random processes, we will identify several interesting classes of random process.

Our motivation for spending so much time in exploring special classes of random processes is that, philosophically, our general definition of a random process is too broad. An arbitrary random process is quite difficult to characterize (CDFs/PDFs/statistics are hard to find, let alone use to compute quantities of interest) Luckily, most random processes are far more structured, and we can exploit this structure to achieve much sharper characterizations of these processes. Our goal here is to identify some useful yet tractable special classes of random processes.

Here are the special classes that we shall consider:

1. Stationary processes
2. Ergodic processes
3. Markov Processes
4. White processes (White noise)
5. Independent increments processes
6. Gaussian processes
7. α-dependent processes

For each of these types, we shall define the notion, think about how we can check whether a process is of the type, and think about the special analyses that are possible.

Stationary Processes

Many random processes of interest maintain a similar probabilistic description over time, or (more precisely) are probabilistically identical with respect to shifts of the origin. For instance, we might expect the intensity and probabilistic description of ambient noise affecting a circuit to remain relatively constant. Similarly, if we are keeping track of the times required for service of customers at a checkout counter at Dissmore's, the statistical description of the process should not change with time. Processes of this sort are known as stationary processes.

Formally, we define two types of stationarity, one based on CDFs/PDFs and one based on statistics.

1. We say that a process is strong sense stationary (SSS), if the n^{th}-order PDF is shift invariant for each n, i.e. if

$$f(x_1, \ldots, x_n; t_1, \ldots, t_n) = f(x_1, \ldots, x_n; t_1+c, \ldots, t_n+c),$$

for each $n, t_1, \ldots, t_n$, and c.

Example: For a strong sense stationary process we expect the joint distribution of $X(1)$, $X(3)$, and $X(8)$ to be the same as the joint distribution of $X(4)$, $X(6)$, and $X(11)$.

Example 2: Consider the process that is generated as follows. We flip a fair coin. If the coin shows head, $X(t)=1$ at all times, and if the coin shows tails $X(t)=0$ at all times.

We can show that this process is strong sense stationary, by finding the n^{th}-order PDF for $X(t)$. In particular, notice that

$$\begin{aligned} f(x_1,\dots,x_n;t_1,\dots,t_n) &= f(x_1,\dots,x_n;t_1,\dots t_n|H)P(H) \\ &\quad + f(x_1,\dots,x_n;t_1,\dots,t_n|T)P(T) \\ &= \delta(x_1-1)\dots\delta(x_n-1)\cdot\tfrac{1}{2} + \delta(x_1)\dots\delta(x_n)\cdot\tfrac{1}{2} \end{aligned}$$

Notice that $f(x_1,\dots,x_n;t_1+c,\dots,t_n+c)$ is exactly the same as $f(x_1,\dots,x_n;t_1,\dots t_n)$ for any c, since the n^{th}-order density does not depend on t. Thus, the random process is SSS.

Although strong-sense-stationarity makes analysis simpler (as we shall see shortly), it is in practice both restrictive and unwieldy: it is a bit much to expect that PDFs of all orders are stationary, and anyway it's unlikely that we can find the n^{th} order PDFs to check.

2. We are thus motivated to define stationary based only on second-order statistics of a random process. We say that a random process is wide sense stationary (WSS), if $E[X(t)]$ is a constant for all time, (i.e. $E[X(t)]=m$), and $R_{XX}(t_1,t_2)$ only depends on t_2-t_1 (i.e., $R_{XX}(t_1+c,t_2+c)=R_{XX}(t_1,t_2)$

In this case, notice that $R_{XX}(t+\tau, t) = E[X(t+\tau)X(t)]$ depends only on $t+\tau-t$ or τ, so $R_{XX}(t+\tau, t) = R(\tau)$ ← (some function of τ). This function $R(\tau) = E[X(t+\tau)X(t)] = E[X(t+\frac{\tau}{2})X(t-\frac{\tau}{2})]$ specifies the autocorrelation of $X(t)$, and so we equivalently refer to $R(\tau)$ as the autocorrelation.

Example:

Consider a WSS process $X(t)$. Let us find $E[X(t_1)X(t_2)]$ from the autocorrelation $R(\tau)$. To do so, notice that $E[X(t_1)X(t_2)] = E[X(t_1)X(t_1+t_2-t_1)]$ $= R(t_2-t_1)$. We thus also automatically recover that $E[X(t_1)^2] = R(0)$.

Notice that checking wide sense stationarity (WSS) is a lot simpler than checking SSS: all we have to do is verify the first- and second-moment conditions, which are often easy to obtain from a description of the random process. Nevertheless, WSS processes are especially amenable to analysis (as we shall argue in a moment) and hence have been widely studied. Also, they are a broader class than SSS (specifically, SSS implies WSS), and turn out to be widely applicable.

There are several reasons why WSS processes are specially tractable:

1. Since the autocorrelation only depends on the delay between the two time points, it can be obtained much more easily from data than the autocorrelation for a nonstationary process.

2. As we shall see during the next few weeks, certain operations on WSS processes are especially amenible to analysis. For example, the outputs of WSS processes driven by noise can be characterize easily.

3. The special structure of a WSS process permits us to meaningfully develop the notion of a spectrum for random processes. That is, we can think of a WSS random process as having components at different frequencies.

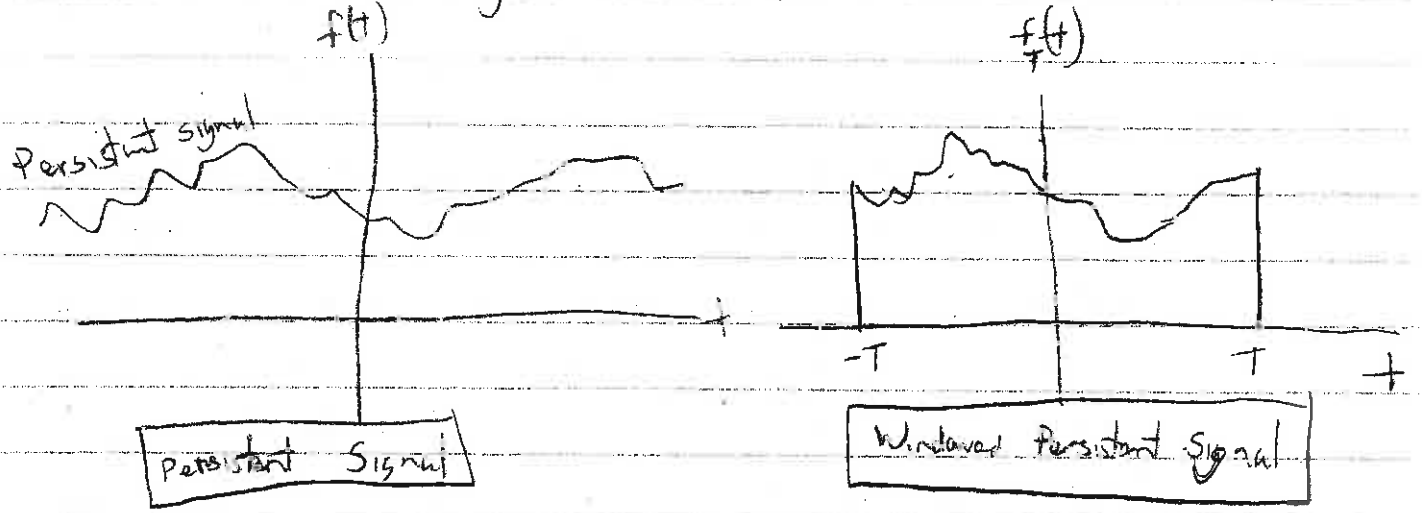
Let us pursue the third point above in some more detail, by developing the notion of a power spectrum for random processes, and finding a convenient expression for the power spectrum for WSS processes. We should begin by reviewing the idea of a spectrum (Fourier transform), energy, and power. In particular, recall that many signals $f(t)$ can be written as a sum of sinusoids:

$f(t) = \int_{-\infty}^{\infty} F(w) e^{jwt} dt$, where $F(w)$ is known as the Fourier transform of $f(t)$ and can be found as $F(w) = \int_{-\infty}^{\infty} f(t) e^{-jwt} dt$.

The energy contained in the signal $f(t)$ is given by $\int_{-\infty}^{\infty} f^2(t)\,dt$. We can straightforwardly show that this energy is equal to $\int_{-\infty}^{\infty} \frac{|F(w)|^2}{2\pi}\,dw$ (this is called Parseval's theorem).

Thus, we can think of the energy contained in a signal $f(t)$ as the sum (integral) of the energies contained in all the frequency bands. With this in mind we refer to $\Phi(w) = \int_{-\infty}^{\infty} \frac{|F(w)|^2}{2\pi}$ as the energy spectrum of a signal.

Unfortunately, most noise signals are persistent signals (they "go on" for ever, see below). The energy in a persistent signal is infinite, and further the signal doesn't have a Fourier transform.



For such persistent signals, it makes sense to compute the energy of a windowed version of the signal (i.e., the signal over an interval of time). Specifically, let us define $f_T(t) = \begin{cases} f(t), & -T \le t \le T \\ 0, & \text{otherwise} \end{cases}$

The energy in this signal is

$$\int_{-\infty}^{\infty} f_T^2(t)\,dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} |F_T(\omega)|^2\,d\omega,$$

where $F_T(\omega)$ is the Fourier transform of $f_T(t)$.

We can consider the energy in the signal at each frequency, i.e. we can consider the energy spectrum $\Phi_T(\omega) = \frac{|F_T(\omega)|^2}{2\pi}$.

The energy spectrum for windowed versions of a persistent signal tells us a lot about its frequency content. However, it is inconvenient that the energy spectrum grows as the windowing interval T grows. Thus, it is sensible for us to consider the energy per unit time or power, i.e.

$$\left(\frac{\text{J}\cdot\text{rad}}{\text{s}}\right) \rightarrow \frac{2\pi}{2T}\int_{-\infty}^{\infty} f_T^2(t)\,dt = \frac{1}{2T}\int_{-\infty}^{\infty} |F_T(\omega)|^2\,d\omega.$$

Based on this expression for the power, we can sensibly think about the power spectrum of the windowed signal, i.e. $P_T(\omega) = \frac{1}{2T}|F_T(\omega)|^2$. Unlike the energy spectrum, we expect the power spectrum to stay bounded — and in fact, in many cases, reach a limit — as T becomes larger and larger.

Finally, we are ready to consider the power spectrum for a random process. For a random process, we note that the power spectrum $P_T(\omega)$ is different for each outcome in the sample space (which maps to a different signal). In this case, it is very reasonable for us

to consider the expected value of the power at each frequency, i.e. $E[P_T(w)]$. We are especially interested in this expected power in the limit of large T, i.e. the expected power at each frequency w over the duration of the signal. We call this spectrum $S(w) = \lim_{T\to\infty} E[P_T(w)]$, the power spectrum.

Definition: Power Spectrum: $S(w) = \lim_{T\to\infty} E[P_T(w)]$ (Power Spectral Density)

Assuming the limit exists

Let us come up with a useful expression for the power spectrum of a WSS random process $f(t)$.

To do so, note that

$$S(w) = \lim_{T\to\infty} E[P_T(w)] = \lim_{T\to\infty} \frac{1}{2T} E\left[|F_T(w)|^2\right]$$

$$S(w) = \lim_{T\to\infty} \frac{1}{2T} E\left[F_T(w) F_T^*(w)\right]$$

$$S(w) = \lim_{T\to\infty} \frac{1}{2T} E\left[\int_{-\infty}^{\infty} f_T(t) e^{-jwt}\, dt \int_{-\infty}^{\infty} f_T(\tau) e^{+jw\tau}\, d\tau\right]$$

$$S(w) = \lim_{T\to\infty} \frac{1}{2T} E\left[\int_{-T}^{T}\int_{-T}^{T} f(t) f(\tau) e^{-jw(t-\tau)}\, d\tau\, dt\right]$$

$$S(w) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} E[f(t) f(\tau)] e^{-jw(t-\tau)}\, d\tau\, dt$$

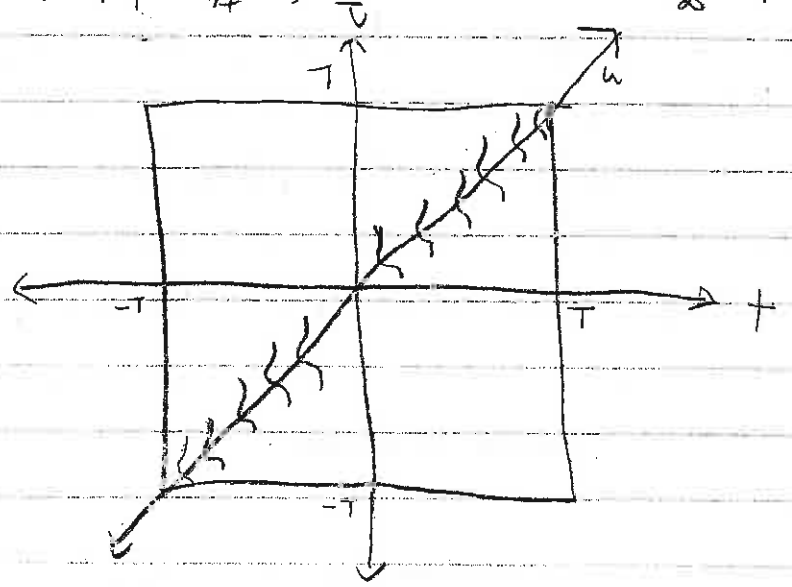
$$S(w) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T}\int_{-T}^{T} R_{ff}(t-\tau) e^{-jw(t-\tau)}\, d\tau\, dt.$$

$$= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T}\int_{t-T}^{t+T} R_{ff}(u) e^{-jwu}\, du\, dt$$

Let us assume that $R_{ff}(\alpha)$ becomes small as α becomes large. This is sensible (correlation dies out with time), and anyway the integral does not exist otherwise.

Then the following picture shows that, for large T,

$$\int_{-T}^{T}\int_{t-T}^{t+T} R_{ff}(u)e^{-jwu}\,du\,dt \approx 2T\int_{-\infty}^{\infty} R_{ff}(u)e^{-jwu}\,du$$



Thus we recover that

$$S(w) = \lim_{T\to\infty} \frac{1}{2T} \cdot 2T\int_{-\infty}^{\infty} R_{ff}(u)e^{-jwu}\,du$$

$$\boxed{S(w) = \int_{-\infty}^{\infty} R_{ff}(u)e^{-jwu}\,du}$$

In words, the power spectral density is the Fourier transform of the the autocorrelation.

This property is a key feature of stationary processes.

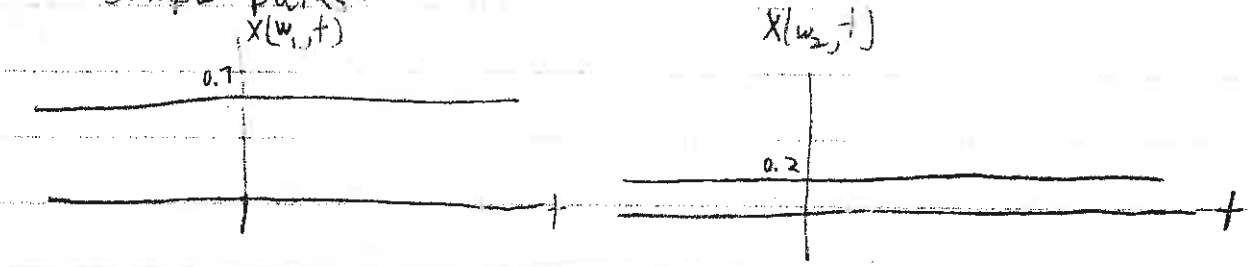
Ergodic Processes

Often, we need to infer statistical properties of a random process from a single instance of the process (i.e., the signal associated with one outcome of the underlying experiment). In general, there is no hope of doing this, because the statistical properties of the process change with time. However, for many stationary processes, statistical properties of the process can be found from a single instance, because the instance displays the average value over time. Random processes where averages can be found from single instances are called ergodic processes.

Specifically, a mean-ergodic process is a stationary process in which the time-average of the signal approaches the mean of the process. That is, consider the random variable $\eta_T = \frac{1}{2T}\int_{-T}^{T} X(t)\,dt$, where $X(t)$ is a random process. Note that $E[\eta_T] = \eta = E[X(t)]$. If $E[(\eta_T - \eta)^2] = \mathrm{var}(\eta_T)$ approaches 0 as $T \to \infty$, the random process is said to mean-ergodic. In other words, a mean-ergodic process is one in which the time average of the signal approaches the ensemble average (the average over sample paths or equivalently outcomes). Processes that are mean-ergodic are particularly easy to characterize from samples, because only one sample of the process (of sufficient duration) is needed to compute the mean.

Example 1

Consider a random process $X(t)$ that equals a constant C at all times, where C is a random variable that is uniform on $[0,1]$. Here are some sample paths:



The sample paths above suggest that the process is not mean-ergodic. To prove this formally, note that $\eta_T = \frac{1}{2T}\int_{-T}^{T} X(t)\,dt = \frac{1}{2T}\int_{-T}^{T} C\,dt = \frac{2CT}{2T} = C$

Thus, $\text{var}(\eta_T) = \frac{1}{12}$ for all T, which does not approach 0.

Example 2

Consider a DT random process $X(t)$ that is defined as follows:

$$X(t) = \begin{cases} 0, & \text{w.p. } 0.4 \\ 1, & \text{w.p. } 0.6 \end{cases}, \quad t = -\infty, \ldots, -1, 0, 1, 2, \ldots \infty$$

where $X(t)$ at each time is independent.

In our discussion of convergence, we proved that $\frac{1}{2T}\sum_{t=-(T-1)}^{T} X(t)$ approaches 0.6 in a mean square sense. Hence, this signal is mean-ergodic.

- The shorter the time constant, the larger we need the amplitude of the signal to be to get the same integral over a particular period. Thus, in the limiting case where the time constant is arbitrarily small, we need arbitrarily large amplitude to get meaningful interpretation for integrals of the signal.

The time-constant interpretation of white noise also makes clear that white noise models are appropriate when the speed of the noise process is fast compared to other signals and system responses.

Discrete-time: $C_{xx}(t,\tau) = \begin{cases} q(t), & t-t=\tau \\ 0, & \text{otherwise} \end{cases}$

Computing var(η_T) to check mean-ergodicity

In general, let $Y(t) = \frac{1}{2T}\int_{-T}^{T} X(\alpha)\,d\alpha$, where $X(\alpha)$ is stationary.
If $\text{var}(Y(t)) = E[Y^2(t)] - [E(Y(t))]^2$ approaches 0 as $T \to \infty$, then the process is mean-ergodic.

Notice that $E[Y(t)) = \eta$, since the process is stationary.

$$E[Y^2(t)] = E\left[\left(\frac{1}{2T}\right)^2 \left(\int_{-T}^{T} X(\alpha)\,d\alpha\right)^2\right]$$

$$= \frac{1}{(2T)^2} E\left[\int_{-T}^{T}\int_{-T}^{T} X(\alpha)X(\beta)\,d\beta\,d\alpha\right]$$

$$= \frac{1}{(2T)^2}\int_{-T}^{T}\int_{-T}^{T} E[X(\alpha)X(\beta)]\,d\beta\,d\alpha$$

$$= \frac{1}{(2T)^2}\int_{-T}^{T}\int_{-T}^{T} R_{xx}(\beta-\alpha)\,d\beta\,d\alpha$$

Let $z_2 = \beta - \alpha$ $\quad z_1 = \alpha$

Then $dz_1\,dz_2 = \frac{1}{\begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix}}\,d\beta\,d\alpha = 1\cdot d\beta\,d\alpha$

$z_2 = \beta - \alpha$

Thus, the integral becomes:

$$E[Y^2(t)] = \frac{1}{(2T)^2}\int_{0}^{2T}\int_{-T}^{T-z_2} R_{xx}(z_2)\,dz_1\,dz_2 + \frac{1}{(2T)^2}\int_{-2T}^{0}\int_{-T-z_2}^{T} R_{xx}(z_2)\,dz_1\,dz_2$$

$$= \frac{1}{(2T)^2}\left[\int_{0}^{2T} R_{xx}(z_2)(2T - z_2)\,dz_2 + \int_{-2T}^{0} R_{xx}(z_2)(2T + z_2)\,dz_2\right]$$

$$= \frac{1}{(2T)^2}\int_{-2T}^{2T} R_{xx}(z_2)(2T - |z_2|)\,dz_2$$

$$E[Y^2(t)] = \frac{1}{t}\int_0^{2t} R_{xx}(z)\left(1 - \frac{z}{2t}\right)dz$$

Thus, if $E[Y^2(t)] - (E[X(t)])^2 = \frac{1}{t}\int_0^{2t} R_{xx}(z)\left(1-\frac{z}{2t}\right)dz - n^2 \to 0$ as $t \to \infty$, we get mean ergodicity.

(Same as saying that $\frac{1}{t}\int_0^{2t} C_{xx}(z)\left(1-\frac{z}{2t}\right)dz \to 0$)

Example: Say that $E[X(t)] = 0$, and that $R_{xx}(z) = e^{-|z|}$, $-\infty < z < \infty$. Is this process mean-ergodic?

- The condition above is a bit difficult to test in some cases. The following condition is equivalent:
  - A process $X(t)$ is mean-ergodic iff $\frac{1}{T}\int_0^T C(\tau)\,d\tau \to 0$ as $T \to 0$.
- The proof of this theorem, known as Slutsky's theorem, is in your text. Conceptually, the theorem makes sense since, if the autocovariance gets progressively weaker over time, we should

be able to average $X(t)$ at different times and get the ensemble mean (from a CLT-type idea). This condition is much easier to test.

There is also a spectral condition for ergodicity; we shall skip this theorem for now, perhaps returning to it after we study linear systems driven by noise.

A random process is said to be a Markov process if the future trajectory of the process is independent of the past, given the current value of the process. That is, a random process is Markov if, for any $t_1 \le \dots \le t_n$, if $P(X(t_n) \le x_n \mid X(t) \; t \le t_{n-1}) = P(X(t_n) \le x_n \mid X(t_{n-1}))$.

This also implies that

$$f(x_n \mid x_{n-1}, \dots, x_1) = f(x_n \mid x_{n-1}), \text{ for } t_1 \le \dots \le t_n$$

and that

$$E[X_n \mid X_{n-1}, \dots, X_1] = E[X_n \mid X_{n-1}].$$

Here are a couple interesting properties of Markov processes:

1. The process is also Markov reversed in time, i.e. $f(x_n \mid x_{n+1}, \dots, x_{n+k}) = f(x_n \mid x_{n+1})$.

2. If $k < m < n$, $f(x_n, x_k \mid x_m) = f(x_n \mid x_m) f(x_k \mid x_m)$, i.e. the past and future are independent given the present.

Let's prove these results:

Many many processes in the world around us are Markov. Let us think about a couple examples:

White Noise

A random process is said to be <u>white</u> (or is called white noise) if the state at any two times are uncorrelated, i.e. the autocovariance $C(t_i, t_j) = 0$ for $t_i \neq t_j$. A white process is what we typically think of as noise, because its value at one time is not correlated with its value at any other time.

Let us consider continuous-time white noise first. For simplicity, let us assume that the mean value of the process $X(t)$ is zero at each time t. We claim that this process is only a meaningful one if $C_{XX}(t_1, t_2) = R_{XX}(t_1, t_2) = q(t_1)\delta(t_1 - t_2)$, i.e. the autocovariance is a spike when $t_1 = t_2$, and 0 otherwise. The reason why we need for $R_{XX}(t_1, t_2)$ to be a spike at the origin (rather than a finite constant) is that otherwise the integral of the process over an interval is trivial. In particular, consider $V(t) = \int_0^t X(\alpha)\, d\alpha$. Notice that

$$E[V^2(t)] = E\left[\int_0^t \int_0^t X(\alpha)X(\beta)\, d\beta\, d\alpha\right] = \int_0^t \int_0^t R_{XX}(\alpha, \beta)\, d\beta\, d\alpha$$

$$E[V^2(t)] = \int_0^t \int_0^t q(\alpha)\delta(\beta - \alpha)\, d\beta\, d\alpha$$

$E[V^2(t)] = \int_0^t q(\alpha)\, d\alpha$, which grows gradually with t, something that is sensible.

In contrast, if $C_{XX}(t_1, t_2)$ were finite for $t_1 = t_2$, $E[V^2(t)]$ would equal 0 for all t which is not sensible. Let us pictorially think about why this makes sense:

Gaussian Random Processes

Earlier in the class, we decided that Gaussian random variables are specially interesting, because 1) many real uncertainties are Gaussian, and 2) Gaussian random variables are specially tractable. In similar fashion, engineers can model many random processes as Gaussian ones (we'll explain what this means in a minute). These Gaussian random processes are specially tractable. In this lecture, we'll define a Gaussian random process, and expose it's special tractability.

Jointly Gaussian Random Variables

In order to define Gaussian random processes, we first need to decide what it means for a group of more than two random variables to be jointly Gaussian. Here's the definition:

We say that a group of n random variables $X_1, \ldots, X_n$ are jointly Gaussian if

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \frac{1}{\sqrt{2\pi^n |C|}} e^{-\frac{1}{2}(\vec{x}-E[\vec{X}])^T C^{-1} (\vec{x}-E[\vec{X}])} \quad (*)$$

where $\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$, $E[\vec{X}] = \begin{bmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{bmatrix}$, and C = covariance matrix of $X_1, \ldots, X_n$

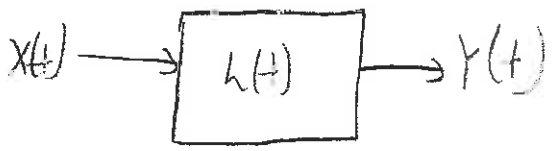
There is one property of jointly Gaussian random variables that is worth mentioning: linear combinations of jointly Gaussian random variables are Gaussian (and multiple such linear combinations are jointly Gaussian).

We are now ready to define a Gaussian random process:

> A random process X is said to be a Gaussian random process if $X(t_1), \ldots, X(t_n)$ are jointly Gaussian, for all sets of times $t_1, \ldots, t_n$ and all n.

Gaussian random processes are specially tractable, in several senses:

1. They are fully specified by first and second moments. That is, from $E(X(t))$ and $R_{XX}(t, \tau)$, we can compute the covariance matrix C for any set of random variables $X(t_1), \ldots, X(t_n)$. Hence, we can find the n^{th}-order PDF for $X(t)$ for any n, using $(*)$.
2. Let's say that we drive a linear system with a Gaussian random process:



Then the output $Y(t)$ is a Gaussian random process.

LECTURE 4, PART B: ASYMPTOTICS OF RANDOM PROCESSES

4. Asymptotics of Random Processes

In the previous section, we have given a general probabilistic and statistical description of a random process. When studying random processes, we sometimes do not need to concern ourselves with this full description, and only wish to understand the asymptotics of the process (the dynamics when the time variable becomes large).

For the sake of convenience, we will limit our study of asymptotics to processes with $t \in \mathbb{Z}^+$, i.e. sequences of random variables $X(0), X(1), X(2), \ldots$ (we'll sometimes also use the notation $X_0, X_1, X_2, \ldots$). However, the definitions and results present here can readily be adapted to the continuous-time case.

(Continued in Part 2)

Let's begin our discussion of convergence by reviewing the notion of convergence for deterministic signals. Recall that convergence of sequences is important when we are trying to evaluate the steady-state value of a signal. Formally, we say that a sequence of numbers $x(1), x(2), \ldots$ converges to a number x, if for every ε, there exists n such that

$$|x(k) - x| \leq \varepsilon \text{ for all } k \geq n.$$

Let us now develop several notions of stochastic convergence. Our first idea is an obvious generalization of the deterministic notion of convergence. We say that the random sequence converges everywhere (meaning for all experimental outcomes) to x, if $X(w, k)$ converges to x for every w. That is, the sequence converges everywhere if for every possible outcome of the experiment, the sequence converges to x.

Convergence everywhere seems like a sensible notion, but in fact it is usually way too stringent. Let us describe one example of a random process that common sense says is convergent, but is not convergent everywhere.

The example before suggests that we should permit the sequences for some initial conditions to not converge, as long as such sequences have probability 0 (notice this is possible because single outcomes have zero probability). This leads us to the following definition of <u>convergence almost everywhere</u>.

Let A be the event consisting of all outcomes ω such that $X(\omega,t)$ converges to x. If the probability of the event A is 1, then the random sequence is said to converge almost everywhere to x. In short, $X(\omega,t)$ converges almost everywhere to x if $P(\{X(t)\to x\}) = 1$.

Convergence almost everywhere is also known as <u>convergence with probability 1</u>.

Convergence almost everywhere turns out to be a quite sensible notion, but is sometimes can be difficult to test because it requires knowledge of the joint density of $X(t)$ over many times t. Also, we sometimes do not care whether entire sequences converge, but simply want to know whether the probability that the random process is near x gets bigger at larger times. We are thus motivated to pursue yet one more definition of convergence. A random sequence is said to converge <u>in probability</u> to x if $P(|X(n)-x|>\varepsilon)\to 0$ as $n\to\infty$ for all ε. That is, $X(n)$

converges in probability to x, if the probability that $X(n)$ differs from x by more than ε at time n approaches zero as n approaches infinity for any ε. We notice that convergence in probability only requires knowledge of the first-order distribution of $X(n)$, and in no way implies that the entire sequence converges for any outcome in the sample space.

We can easily see that convergence almost everywhere implies convergence in probability. Do you see why?

Also, let us give an example which shows that convergence in probability does not imply convergence with probability 1.

As we discussed in Part A, we sometimes may only be able to, or only wish to, characterize moments of a random process. In fact, first- and second-moment information are sufficient to meaningfully define the convergence of a random sequence: we can think of a sequence as converging if the spreads of the sequence values record a point decrease with increasing t. In particular, we say that $X(t)$ converges to x in a mean square sense, if

$$E[(X(t)-x)^2] \to 0 \quad \text{as } t \to \infty.$$

Let us check for mean square convergence in an example:

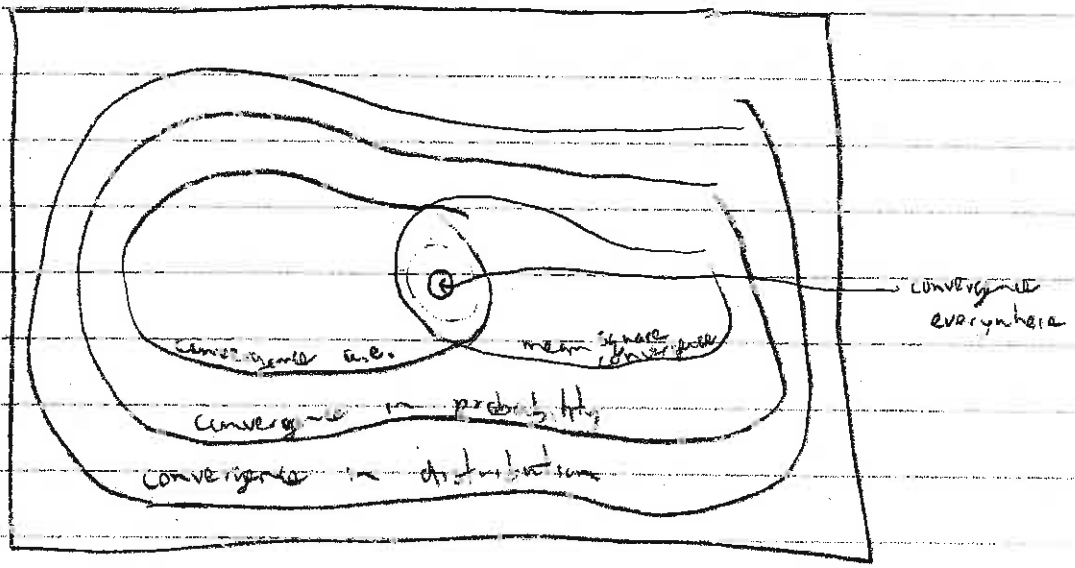
We note that mean square convergence is quite a strong notion of convergence, in that it implies convergence in probability. This can be proved using Chebyshev's inequality:

It's also worth noting that mean square convergence neither implies nor is implied by convergence almost everywhere. You will do some examples verifying this idea in your homework.

The condition for mean square convergence can sometimes be difficult to test, because we may not know the limit point to which the sequence converges. The Cauchy criterion (see your text) yields a simpler test for convergence. In particular, applying it to the mean square convergence test, we obtain that if $E[|x_{n+m} - x_n|^2] \to 0$ as $n \to \infty$ for all $m > 0$, then the sequence converges in a mean square sense. Similar tests for the other notions of convergence that do not depend on the limit point can be developed (though this idea is used most often in checking mean-square convergence).

Finally, there are many occasions where we a random process does not converge to a value, but the first-order CDF of the random process does converge to a fixed function, i.e. $F(x,n) \to F(x)$ for all x as $n \to \infty$. Convergence in distribution is a much weaker notion of convergence, and is applicable to any problem where the CDF/PDF reach a steady-state with time. We note that all the other notions of convergence automatically imply convergence in distribution.

To summarize, let us present a Venn diagram relating the notions of convergence (Figure 8-3 in your text):



Convergence of Some Special Sequences

There are some special sequences that we encounter over and over and so their convergence is of special interest. In these cases, analysts have gained great insight into the sequences' convergence; let us take a moment to study these special cases.

Special Case 1

We can pose the problem of finding the fraction of trials of an experiment that are successful in terms of random processes. In particular, let's consider a Bernoulli process X_i,

i.e. a process where each X_i is an independent random variable that equals 1 with probability p and equals 0 otherwise. We note that $X_i = 1$ can be interpreted as a success on i^{th} trial, and $X_i = 0$ a failure. We can thus interpret the random process

$$Y_i = \frac{\sum_{j=1}^{i} X_j}{i}$$

as representing the fraction of trials that are successes.

We claim that Y_i approaches p in probability. This result is known as the <u>law of large numbers</u>.

Proof:

Note that $E[Y_i] = \frac{E[\sum_{j=1}^{i} X_j]}{i} = \frac{\sum_{j=1}^{i} E(X_j)}{i} = \frac{pi}{i} = p.$

Thus, from Chebyshev's inequality, we know that $P(|Y_i - p| < \varepsilon) \geq 1 - \frac{\text{var}(Y_i)}{\varepsilon^2}$

$$\text{var}(Y_i) = \text{var}\left(\frac{\sum_{j=1}^{i} X_j}{i}\right) = \frac{1}{i^2} \sum_{j=1}^{i} \text{var}(X_j) = \frac{1}{i^2}(ip(1-p)) = \frac{p(1-p)}{i}$$

↑ Since independent

Thus, $P(|Y_i - p| < \varepsilon) \geq 1 - \frac{p(1-p)}{i\varepsilon^2} \to 1$ as $i \to \infty$, for each ε.

Thus, we have proved convergence in probability. □

In fact, it can be shown that Y_i approaches p not only in probability but almost everywhere. This

stronger result is known as the strong law of large numbers. We will not take the time to prove the strong law here.

Special Case 2

Consider a sequence of independent random variables $X_1, X_2, \ldots$, and consider the random variable $S = X_1 + X_2 + \ldots + X_n$ (i.e., the sum of the first n elements of this sequence). Notice that $E[S] = \sum_{i=1}^{n} E(X_i)$, while $\text{Var}(S) = \sum_{i=1}^{n} \text{var}(X_i)$.

As long as the following conditions hold:

$$\begin{cases} 1. \quad \sigma_1^2 + \ldots + \sigma_n^2 \xrightarrow[n \to \infty]{} \infty \\ 2. \quad \int_{-\infty}^{\infty} x^{\alpha} f_i(x)\, dx < K < \infty \text{ for some } \alpha > 2 \text{ and some } K. \end{cases}$$

the CDF of the R.V. S approaches the CDF of a Gaussian random variable with mean $E[S]$ and variance $\text{var}(S)$.

Proof: the proof is rather messy and complicated, so I'll skip it for now.

LECTURE 4, PART 3

In this third part of our introduction to random processes, we will identify several interesting classes of random process.

Our motivation for spending so much time in exploring special classes of random processes is that, philosophically, our general definition of a random process is too broad. An arbitrary random process is quite difficult to characterize (CDFs/PDFs/statistics are hard to find, let alone use to compute quantities of interest). Luckily, most random processes are far more structured, and we can exploit this structure to achieve much sharper characterizations of these processes. Our goal here is to identify some useful yet tractable special classes of random processes.

Here are the special classes that we shall consider:

1. Stationary processes
2. Ergodic processes
3. Markov Processes
4. White processes (White noise)
5. Independent increments processes
6. Gaussian processes
7. Independent processes

For each of these types, we shall define the notion, think about how we can check whether a process is of the type, and think about the general analyses that are possible

Stationary Processes

Many random processes of interest maintain a similar probabilistic description over time, or (more precisely) are probabilistically identical with respect to shifts of the origin. For instance, we might expect the intensity of ambient noise affecting a circuit to remain relatively constant. Similarly, if we are keeping track of the times required for service of customers at a checkout counter at Dissmore's, the statistical description of the process should not change with time. Processes of this sort are known as stationary processes.

Formally, we define two types of stationarity, one based on CDFs/PDFs and one based on statistics.

1. We say that a process is strong sense stationary (SSS) if the n^{th}-order PDF is shift invariant for each n, i.e. if

$$f(x_1, \ldots, x_n; t_1, \ldots, t_n) = f(x_1, \ldots, x_n; t_1+c, \ldots, t_n+c),$$

for each $n, t_1, \ldots, t_n$, and c.

Example: For a strong sense stationary process we expect the joint distribution of $X(1)$, $X(3)$, and $X(8)$ to be the same as the joint distribution of $X(4)$, $X(6)$, and $X(11)$.

Example 2: Consider the process that is generated as follows. We flip a fair coin. If the coin shows head, $X(t)=1$ at all times, and if the coin shows tails $X(t)=0$ at all times.

We can show that this process is strong sense stationary, by finding the n^{th}-order PDF for $X(t)$. In particular, notice that

$$f(x_1,\dots,x_n;t_1,\dots,t_n)$$
$$= f(x_1,\dots,x_n;t_1,\dots,t_n|H)P(H)$$
$$+ f(x_1,\dots,x_n;t_1,\dots,t_n|T)P(T)$$
$$= \delta(x_1-1)\cdots\delta(x_n-1)\cdot\frac{1}{2} + \delta(x_1)\cdots\delta(x_n)\cdot\frac{1}{2}$$

Notice that $f(x_1,\dots,x_n;t_1+c,\dots,t_n+c)$ is exactly the same as $f(x_1,\dots,x_n;t_1,\dots t_n)$ for any c, since the n^{th}-order density does not depend on t. Thus, the random process is SSS.

Although strong-sense-stationarity makes analysis simpler (as we shall see shortly), it is in practice both restrictive and unwieldy: it is a bit much to expect that PDFs of all orders are stationary, and anyway it's unlikely that we can find the n^{th} order PDFs to check.

2. We are thus motivated to define stationary based only on second-order statistics of a random process. We say that a random process is wide sense stationary (WSS), if $E[X(t)]$ is a constant for all time, (i.e. $E[X(t)]=m$) and $R_{XX}(t_1,t_2)$ only depends on t_2-t_1 (i.e., $R_{XX}(t_1+c,t_2+c)=R_{XX}(t_1,t_2)$

In this case, notice that $R_{XX}(t+\tau, t) = E[X(t+\tau)X(t)]$ depends only on $t+\tau-t$ or τ, so $R_{XX}(t+\tau, t) = R(\tau)$ ← (some function of τ). This function $R(\tau) = E[X(t+\tau)X(t)] = E[X(t+\frac{\tau}{2})X(t-\frac{\tau}{2})]$ specifies the autocorrelation of $X(t)$, and so we equivalently refer to $R(\tau)$ as the autocorrelation.

Example:

Consider a WSS process $X(t)$. Let us find $E[X(t_1)X(t_2)]$ from the autocorrelation $R(\tau)$. To do so, notice that $E[X(t_1)X(t_2)] = E[X(t_1)X(t_1+t_2-t_1)]$ $= R(t_2-t_1)$. We thus also automatically recover that $E[X(t_1)^2] = R(0)$.

Notice that checking wide sense stationarity (WSS) is a lot simpler than checking SSS: all we have to do is verify the first- and second-moment conditions, which are often easy to obtain from a description of the random process. Nevertheless, WSS processes are especially amenable to analysis (as we shall argue in a moment) and hence have been widely studied. Also, they are a broader class than SSS (specifically, SSS implies WSS), and turn out to be widely applicable.

There are several reasons why WSS processes are specially tractable:

1. Since the autocorrelation only depends on the delay between the two time points, it can be obtained much more easily from data than the autocorrelation for a nonstationary process.

2. As we shall see during the next few weeks, certain operations on WSS processes are especially amenable to analysis. For example, the outputs of WSS processes driven by noise can be characterize easily.

3. The special structure of a WSS process permits us to meaningfully develop the notion of a spectrum for random processes. That is, we can think of a WSS random process as having components at different frequencies.

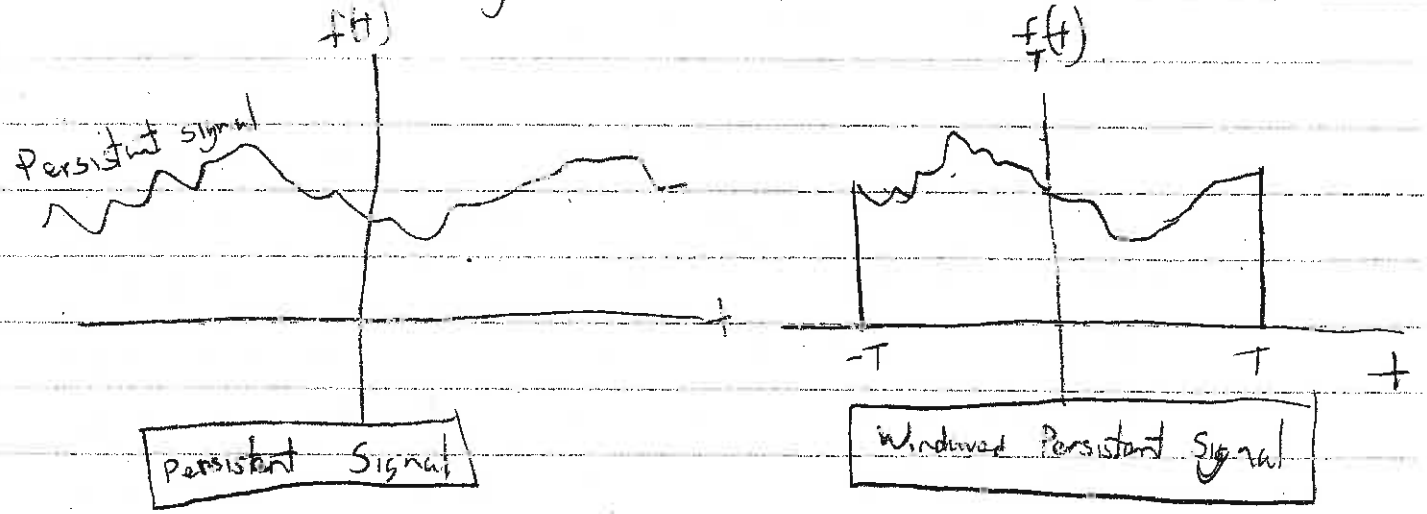
Let us pursue the third point above in some more detail, by developing the notion of a power spectrum for random processes, and finding a convenient expression for the power spectrum for WSS processes. We should begin by reviewing the idea of a spectrum (Fourier transform), energy, and power. In particular, recall that many signals $f(t)$ can be written as a sum of sinusoids:

$f(t) = \int_{-\infty}^{\infty} F(w) e^{jwt}\, dt$, where $F(w)$ is known as the Fourier transform of $f(t)$ and can be found as $F(w) = \int_{-\infty}^{\infty} f(t) e^{-jwt}\, dt$.

The energy contained in the signal $f(t)$ is given by $\int_{-\infty}^{\infty} f^2(t)\,dt$. We can straightforwardly show that this energy is equal to $\int_{-\infty}^{\infty} \frac{|F(w)|^2}{2\pi}\,dw$ (this is called Parseval's theorem).

Thus, we can think of the energy contained in a signal $f(t)$ as the sum (integral) of the energies contained in all the frequency bands. With this in mind we refer to $\Phi(w) = \int_{-\infty}^{\infty} \frac{|F(w)|^2}{2\pi}$ as the energy spectrum of a signal.

Unfortunately, most noise signals are persistent signals (they "go on" for ever, see below). The energy in a persistent signal is infinite, and further the signal doesn't have a Fourier transform.



For such persistent signals, it makes sense to compute the energy of a windowed version of the signal (i.e., the signal over an interval of time). Specifically, let us define $f_T(t) = \begin{cases} f(t), & -T \le t \le T \\ 0, & \text{otherwise} \end{cases}$

The energy in this signal is

$$\int_{-\infty}^{\infty} f_T^2(t)\,dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} |F_T(\omega)|^2\,d\omega,$$

where $F_T(\omega)$ is the Fourier transform of $f_T(t)$.

We can consider the energy in this signal at each frequency, i.e. we can consider the energy spectrum $\Phi_T(\omega) = \frac{|F_T(\omega)|^2}{2\pi}$.

The energy spectrum for windowed versions of a persistent signal tell us a lot about its frequency content. However, it is inconvenient that the energy spectrum grows as the windowing interval T grows. Thus, it is sensible for us to consider the energy per unit time or power, i.e.

$$\left(\frac{J\cdot rad}{s}\right) \rightarrow \frac{2\pi}{2T}\int_{-\infty}^{\infty} f_T^2(t)\,dt = \frac{1}{2T}\int_{-\infty}^{\infty} |F_T(\omega)|^2\,d\omega.$$

Based on this expression for the power, we can sensibly think about the power spectrum of the windowed signal, i.e. $P_T(\omega) = \frac{1}{2T}|F_T(\omega)|^2$. Unlike the energy spectrum, we expect the power spectrum to stay bounded — and in fact, in many cases, reach a limit — as T becomes larger and larger.

Finally, we are ready to consider the power spectrum for a random process. For a random process, we note that the power spectrum $P_T(\omega)$ is different for each outcome in the sample space (which maps to a different signal). In this case, it is very reasonable for us

to consider the expected value of the power at each frequency, i.e. $E[P_T(\omega)]$. We are especially interested in the expected power in the limit as T goes to ∞, i.e. the expected power at each frequency ω over the duration of the signal. We call this quantity $S(\omega) = \lim_{T\to\infty} E[P_T(\omega)]$, the power spectrum.

Definition: Power spectrum (Power Spectral Density): $\lim_{T\to\infty} E[P_T(\omega)]$

Let us come up with a useful expression for the power spectrum of a WSS random process $Y(t)$.

To do so, note that

$$S(\omega) = \lim_{T\to\infty} E[P_T(\omega)] = \lim_{T\to\infty} \frac{1}{2T} E[|F_T(\omega)|^2]$$

$$S(\omega) = \lim_{T\to\infty} \frac{1}{2T} E[F_T(\omega) F_T^*(\omega)]$$

$$S(\omega) = \lim_{T\to\infty} \frac{1}{2T} E\left[\int_{-\infty}^{\infty} f_T(t) e^{-j\omega t}\, dt \int_{-\infty}^{\infty} f_T(\tau) e^{j\omega\tau}\, d\tau\right]$$

$$S(\omega) = \frac{1}{2} E\left[\int_{-T}^{T}\int_{-T}^{T} f(t) f(\tau) e^{-j\omega(t-\tau)}\, d\tau\, dt\right]$$

$$S(\omega) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \int_{-T}^{T} E[f(t) f(\tau)] e^{-j\omega(t-\tau)}\, d\tau\, dt$$

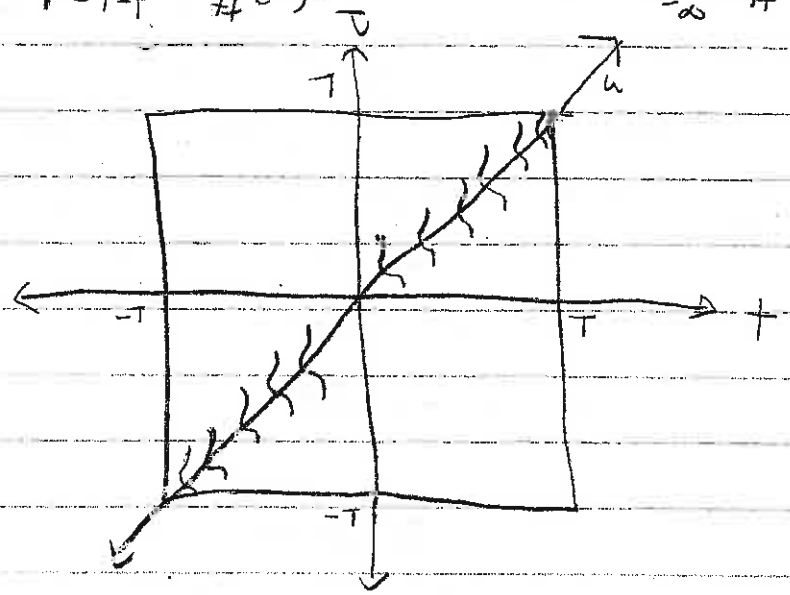
$$S(\omega) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T}\int_{-T}^{T} R_{ff}(t-\tau) e^{-j\omega(t-\tau)}\, d\tau\, dt.$$

$$= \int_{-T}^{T} \int_{t-T}^{t+T} R_{ff}(u) e^{-j\omega u}\, du\, dt$$

Let us assume that $R_{ff}(\alpha)$ becomes small as α becomes large. This is sensible (correlation dies out with time), and anyway the integral does not exist otherwise.

Then the following picture shows that, for large T,

$$\int_{-T}^{T}\int_{t-T}^{t+T} R_{ff}(u)e^{-jwu}\,du\,dt \approx 2T\int_{-\infty}^{\infty} R_{ff}(u)e^{-jwu}\,du$$



Thus, we recover that

$$S(w) = \lim_{T\to\infty} \frac{1}{2T}\cdot 2T\int_{-\infty}^{\infty} R_{ff}(u)e^{-jwu}\,du$$

$$\boxed{S(w) = \int_{-\infty}^{\infty} R_{ff}(u)e^{-jwu}\,du}$$

In words, the power spectral density is the Fourier transform of the the autocorrelation.

This property is a key feature of stationary processes.

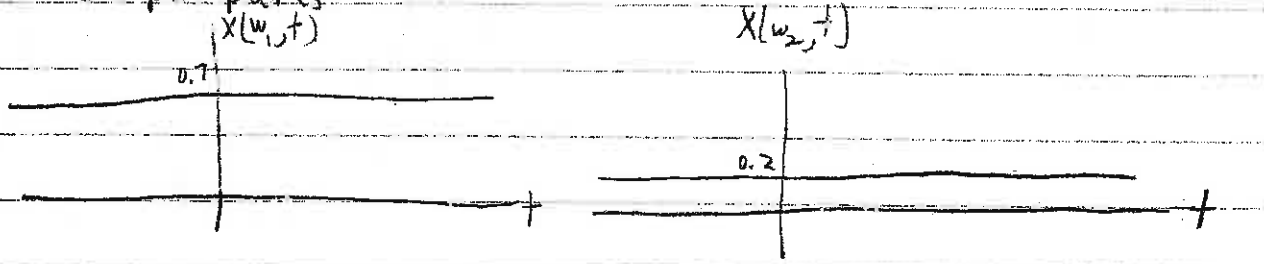
Ergodic Processes

Often, we need to infer statistical properties of a random process from a single instance of the process (i.e., the signal associated with one outcome of the underlying experiment). In general, there is no hope of doing this, because the statistical properties of the process change with time. However, for many stationary processes, statistical properties of the process can be found from a single instance, because the instance displays the average value over time. Random processes where averages can be found from single instances are called ergodic processes.

Specifically, a mean-ergodic process is a stationary process in which the time-average of the signal approaches the mean of the process. That is, consider the random variable $\eta_T = \frac{1}{2T}\int_{-T}^{T} X(t)\,dt$, where $X(t)$ is a random process. Note that $E[\eta_T] = \eta = E[X(t)]$. If $E[(\eta_T - \eta)^2] = \text{var}(\eta_T)$ approaches 0 as $T \to \infty$, the random process is said to mean-ergodic. In other words, a mean-ergodic process is one in which the time average of the signal approaches the ensemble average (the average over sample paths or equivalently outcomes). Processes that are mean-ergodic are particularly easy to characterize from samples, because only one sample of the process (of sufficient duration) is needed to compute the mean.

Example 1

Consider a random process $X(t)$ that equals a constant C at all times, where C is a random variable that is uniform on $[0,1]$. Here are some sample paths:



The sample paths above suggest that the process is not mean-ergodic. To prove this formally, note that $m_T = \frac{1}{2T}\int_{-T}^{T} X(t)\,dt = \frac{1}{2T}\int_{-T}^{T} C\,dt = \frac{2CT}{2T} = C$

Thus, $\text{var}(m_T) = \frac{1}{12}$ for all T, which does not approach 0.

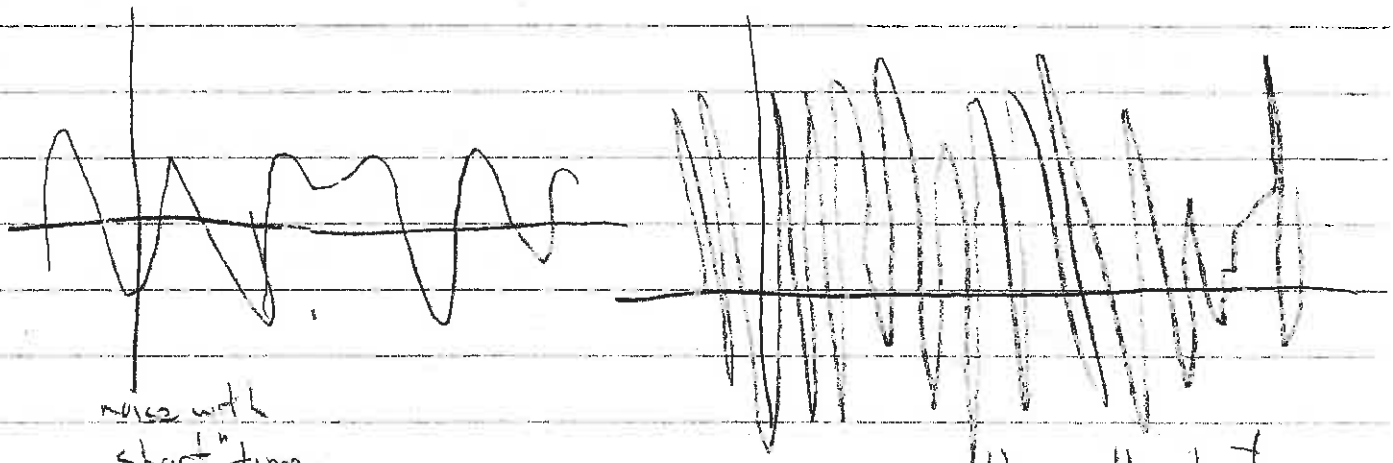
Example 2

Consider a DT random process $X(t)$ that is defined as follows:

$$X(t) = \begin{cases} 0, & \text{w.p. } 0.4 \\ 1, & \text{w.p. } 0.6 \end{cases}, \quad t = -\infty, \ldots, -1, 0, 1, 2, \ldots \infty$$

where $X(t)$ at each time is independent.

In our discussion of convergence, we proved that $\frac{1}{2T}\sum_{t=-(T-1)}^{T} X(t)$ approaches 0.6 in a mean square sense. Hence, this signal is mean-ergodic.



noise with short "time constant"

noise with really short "time constant"

- The shorter the time constant, the larger we need the amplitude of the signal to be to get the same integral over a particular period. Thus, in the limiting case where the time constant is arbitrarily small, we need arbitrarily large amplitude to get meaningful interpretations for integrals of the signal.

The time-constant interpretation of white noise also makes clear that white noise models are appropriate when the speed of the noise process is fast compared to other signals and system responses.

Discrete-time: $C_{xx}(t,\tau) = \begin{cases} q(t), & \text{for } t=\tau \\ 0, & \text{otherwise} \end{cases}$

Computing $\text{var}(\eta_T)$ to check mean-ergodicity

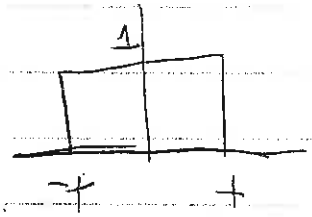
In general, let $Y(t) = \frac{1}{2T}\int_{-T}^{T} X(\alpha)\,d\alpha$, where $X(\alpha)$ is stationary.
If $\text{var}(Y(t)) = E[Y^2(t)] - [E(Y(t))]^2$ approaches 0 as $t \to \infty$, then the process is mean-ergodic.

Note that $E[Y(t)] = \eta$, since the process is stationary.

$$E[Y^2(t)] = E\left[\left(\frac{1}{2T}\right)^2 \left(\int_{-T}^{T} X(\alpha)\,d\alpha\right)^2\right]$$

$$= \frac{1}{(2T)^2} E\left[\int_{-T}^{T}\int_{-T}^{T} X(\alpha)X(\beta)\,d\beta\,d\alpha\right]$$

$$= \frac{1}{(2T)^2}\int_{-T}^{T}\int_{-T}^{T} E[X(\alpha)X(\beta)]\,d\beta\,d\alpha$$



$$= \frac{1}{(2T)^2}\int_{-T}^{T}\int_{-T}^{T} R_{xx}(\beta-\alpha)\,d\beta\,d\alpha$$

Let $z_2 = \beta - \alpha$ $\quad z_1 = \alpha$

Then $dz_1\,dz_2 = \dfrac{1}{\begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix}}\,d\beta\,d\alpha = 1\cdot d\beta\,d\alpha$

$z_2 = \beta - \alpha$

$z_2 = \beta - \alpha$

Thus, the integral becomes:

$$E[Y^2(t)] = \frac{1}{(2T)^2}\int_{0}^{2T}\int_{-T}^{T-z_2} R_{xx}(z_2)\,dz_1\,dz_2$$

$$+ \frac{1}{(2T)^2}\int_{-2T}^{0}\int_{-T-z_2}^{T} R_{xx}(z_2)\,dz_1\,dz_2$$

$$= \frac{1}{(2T)^2}\left[\int_{0}^{2T} R_{xx}(z_2)(2T - z_2)\,dz_2 + \int_{-2T}^{0} R_{xx}(z_2)(2T + z_2)\,dz_2\right]$$

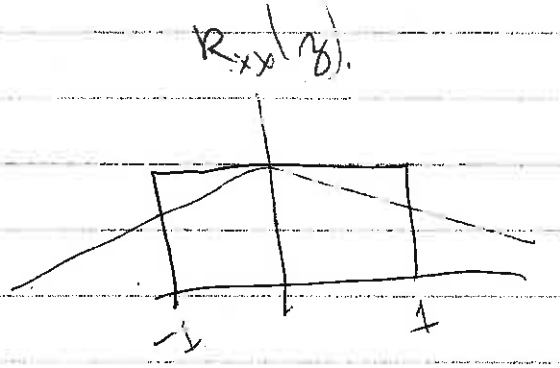
$$= \frac{1}{(2T)^2}\int_{-2T}^{2T} R_{xx}(z_2)(2T - |z_2|)\,dz_2$$

$$E(Y^2(t)) = \frac{1}{t}\int_0^{2t} R_{xx}(z)\left(1-\frac{z}{2t}\right)dz \leq \frac{2t}{t} \to 0$$

Thus, if $E(Y^2(t)) - (E[Y(t)])^2 = \frac{1}{t}\int_0^{2t} R_{xx}(z)\left(1-\frac{z}{2t}\right)dz - m^2 \to 0$ as $t \to \infty$, we get mean ergodicity.

(Same as saying that $\frac{1}{t}\int_0^{2t} C_{xx}(z)\left(1-\frac{z}{2t}\right)dz \to 0$)

Example: Say that $E[X(t)] = 0$, and that $R_{xx}(z) = e^{-|z|}$, $-\infty < z < \infty$. Is this process mean-ergodic?



- The condition above is a bit difficult to test in some cases. The following condition is equivalent:
  - A process $X(t)$ is mean-ergodic iff $\frac{1}{T}\int_0^T C(\tau)d\tau \to 0$ as $T \to 0$.
- The proof of this theorem known as Slutsky's theorem, is in your text. Conceptually the theorem makes sense since, if the autocovariance gets progressively weaker over time, we should

be able to average $X(t)$ at different times and get the ensemble mean (from a CLT-type idea). This condition is much easier to test.

There is also a spectral condition for ergodicity; we shall skip this theorem for now, perhaps returning to it after we study linear systems driven by noise.

A random process is said to be a Markov process if the future trajectory of the process is independent of the past, given the current value of the process. That is, a random process is Markov if, for any $t_1 < \cdots < t_n$, if $P(X(t_n) \le x_n \mid X(t_i),\ t_i \le t_{n-1}) = P(X(t_n) \le x_n \mid X(t_{n-1}))$.

This also implies that

$$f(x_n \mid x_{n-1}, \ldots, x_1) = f(x_n \mid x_{n-1}), \text{ for } t_1 \le \cdots \le t_n$$

and that

$$E[X_n \mid X_{n-1}, \ldots, X_1] = E[X_n \mid X_{n-1}].$$

Here are a couple interesting properties of Markov processes:

1. The process is also Markov reversed in time, i.e. $f(x_n \mid x_{n+1}, \ldots, x_{n+m}) = f(x_n \mid x_{n+1})$.

2. If $k < m < n$, $f(x_n, x_k \mid x_m) = f(x_n \mid x_m) f(x_k \mid x_m)$, i.e. the past and future are independent given the present.

Let's prove these results:

Many, many processes in the world around us are Markov. Let us think about a couple examples:

White Noise

A random process is said to be <u>white</u> (or is called white noise) if the state at any two times are uncorrelated, i.e. the autocovariance $C(t_i, t_j) = 0$ for $t_i \neq t_j$. A white process is what we typically think of as noise, because its value at one time is not correlated with its value at any other time.

Let us consider continuous-time white noise first. For simplicity, let us assume that the mean value of the process $X(t)$ is zero at each time t. We claim that this process is only a meaningful one if $C_{XX}(t_1, t_2) = R_{XX}(t_1, t_2) = q_1(t_1)\delta(t_1 - t_2)$, i.e. the autocovariance is a spike when $t_1 = t_2$, and 0 otherwise. The reason why we need for $R_{XX}(t_1, t_2)$ to be a spike at the origin (rather than a finite constant) is that otherwise the integral of the process over an interval is trivial. In particular, consider $W(t) = \int_0^t X(\alpha)\, d\alpha$. Notice that

$$E[W^2(t)] = E\left[\int_0^t \int_0^t X(\alpha)X(\beta)\, d\beta\, d\alpha\right] = \int_0^t \int_0^t R_{XX}(\alpha, \beta)\, d\beta\, d\alpha$$

$$E[W^2(t)] = \int_0^t \int_0^t q_1(\alpha)\delta(\beta - \alpha)\, d\beta\, d\alpha$$

$E[W^2(t)] = \int_0^t q_1(\alpha)\, d\alpha$, which grows gradually with t, something that is sensible.

In contrast, if $C_{XX}(t_1, t_2)$ were finite for $t_1 = t_2$, $E[W^2(t)]$ would equal 0 for all t which is not sensible.

Let us pictorially think about why this makes sense:

LECTURE 5: LINEAR SYSTEMS DRIVEN BY NOISE

Section 1: Introduction

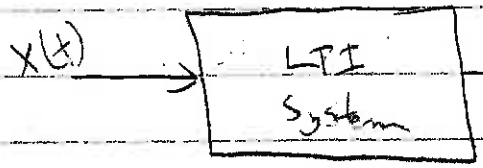
Now that we have gained an understanding of noise processes — and, in particular, stationary noise processes — we can begin to pursue one of the most significant applications of random process theory in engineering: the analysis of systems driven by random process inputs. That is, we would like to develop systematic ways for characterizing the outputs of systems whose inputs are random processes. Because many interesting natural phenomena as well as engineering designs are well described as linear systems, and because linear systems are specially tractable, we will focus specifically on analysis of linear systems driven by noise processes.

The lecture is organized as follows:

- In Section 2, we specify and motivate the problem of analyzing responses of linear systems driven by white noise.
- In Section 3, we solve the problem developed in Section 2.
- In Section 4, we address the special case where the input is stationary, which admits a spectral methodology for analysis.
- In Section 5, we will pursue the discrete time case,

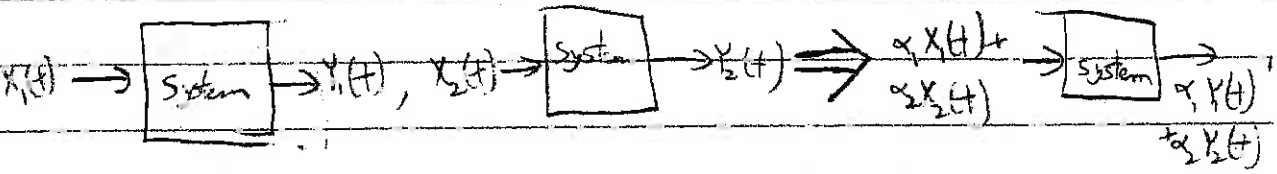
Section 2: The Problem

Let us consider a linear time-invariant system with a single input $X(t)$ and a single output $Y(t)$:

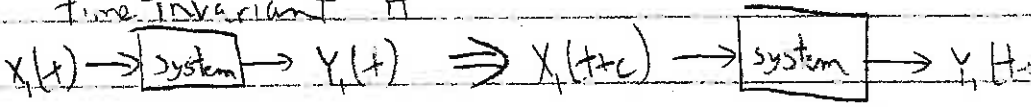


Assume Continuous Time

Recall that the system is considered "linear" if



and time-invariant if



For simplicity, we will assume that initial conditions of the system are 0. This is not limiting, since it is well-known that we can find the response to the initial conditions separately and add it to the response due to the input, to find the total response.

Note that the system will have initial conditions whenever it has some memory

It is classical that the response of a SISO LTI system can be found in two steps:

1. Find the response of the system when the input is a spike at time 0, i.e. $X(t)=\delta(t)$. Let's call this response the impulse response, and denote it $h(t)$.

2. Find the response of the system to an arbitrary input $X(t)$ by breaking the input into a train of impulses, and adding the response to each. This yields an output $Y(t) = \int_{-\infty}^{\infty} X(\alpha) h(t-\alpha)\, d\alpha = \int_{-\infty}^{\infty} X(t-\alpha) h(\alpha)\, d\alpha$. The expression for $Y(t)$ is the famed convolution integral. The convolution operation is often written in short as $Y(t) = X(t) * h(t)$.

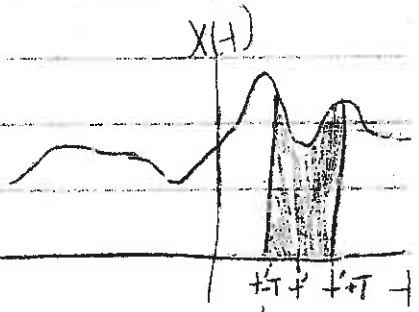
Note: For a few special linear systems, the impulse response is an unwieldy entity (e.g., $Y(t) = \frac{dX(t)}{dt}$). In these cases, the results derived here will not be directly useful, but similar analysis can be used.

Here, we shall study the outputs $Y(t)$ of LTI SISO systems, when the input $X(t)$ is a random process. In particular, we will assume that the mean and covariance of $X(t)$ are known, and that the impulse response $h(t)$ is known (or can easily be determined from the system description). Our goal will be to find the mean and covariance of the output $Y(t)$, as well as the cross-statistics of $X(t)$ and $Y(t)$.

Example 1

$X(t)$: White noise, $E[X(t)] = 0$, $R_{XX}(t, \tau) = 4\delta(t-\tau)$

System, Averager: $Y(t) = \int_{t-T}^{t+T} X(\alpha)\, d\alpha$



$Y(t') = \int_{t'-T}^{t'+T} X(t)\, dt$

- Note that the impulse response of this system is

$h(t) = 1, \quad -T \le t \le T$

Do you see why?

Intuition: response should be smoother

Example 2

- $X(t)$ is is a sinusoid with random phase
- The output $Y(t)$ satisfies the differential equation

$$\frac{dY(t)}{dt} + 6Y(t) = X(t)$$

- Impulse response: $h(t) = e^{-6t}, \quad t \geq 0$

do you see why?

Section 3: Finding statistics of $Y(t)$ and joint statistics of $X(t)$ and $Y(t)$

Let us first compute the expected value of $Y(t)$

Notice that

$$E[Y(t)] = E\left[\int_{-\infty}^{\infty} X(\alpha) h(t-\alpha)\, d\alpha\right]$$

$$E[Y(t)] = \int_{-\infty}^{\infty} E[X(\alpha)]\, h(t-\alpha)\, d\alpha$$

That is, the expected output $E[Y(t)]$ can be computed as the convolution of the expected input with the impulse response, or in short $E[Y(t)] = E(X(t)) * h(t)$

Next, let us attempt to find the autocorrelation $R_{YY}(t,\tau)$. In doing so, we'll find it convenient to first find the cross-correlation between $X(t)$ and $Y(t)$, and then find the autocorrelation of $Y(t)$.

Notice that

$$R_{XY}(t,\tau) = E[X(t)Y(\tau)]$$

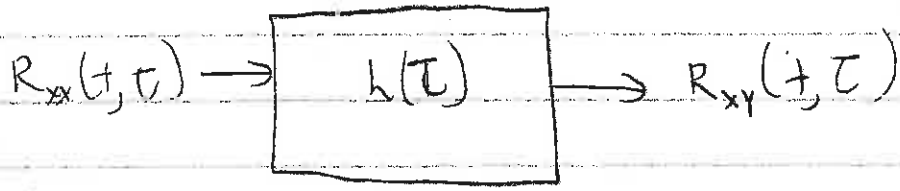
$$= E[X(t)Y(\tau)]$$

$$= E\left[X(t)\int_{-\infty}^{\infty} X(\tau-\alpha)h(\alpha)\,d\alpha\right]$$

$$= \int_{-\infty}^{\infty} E[X(t)X(\tau-\alpha)]\,h(\alpha)\,d\alpha$$

$$\boxed{R_{XY}(t,\tau) = \int_{-\infty}^{\infty} R_{XX}(t,\tau-\alpha)h(\alpha)\,d\alpha}$$

Notice that $R_{X,Y}(t,\tau)$ can actually be viewed as the output of a linear system with input $R_{X,X}(t,\tau)$ and impulse response $h(\tau)$, where t is just a parameter:



Now consider $R_{YY}(t,\tau)$.

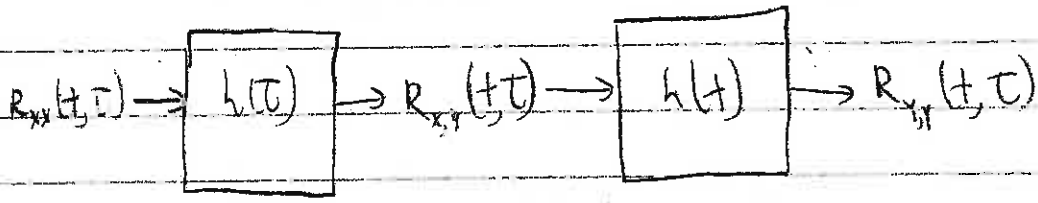
Notice that $R_{YY}(t,\tau) = E[Y(t)\,Y(\tau)]$

$$= E\left[\left(\int_{-\infty}^{\infty} X(t-\alpha)\,h(\alpha)\,d\alpha\right)(Y(\tau))\right]$$

$$= \int_{-\infty}^{\infty} E[X(t-\alpha)\,Y(\tau)]\,h(\alpha)\,d\alpha$$

$$\boxed{R_{YY}(t,\tau) = \int_{-\infty}^{\infty} R_{XY}(t-\alpha,\tau)\,h(\alpha)\,d\alpha}$$

Notice that $R_{YY}(t,\tau)$ can be viewed as the response of a linear system with input $R_{XY}(t,\tau)$ and impulse response $h(t)$, where τ is just a parameter:



Thus, we have developed a methodology for finding the first and second moments of outputs from linear systems. We can find higher moments in similar fashion from the moments of $X(t)$, but this is a bit beyond the scope of this course.

It's worth noting that the covariances $C_{XY}(t,\tau)$ and $C_{YY}(t,\tau)$ can be found in very similar fashion:

$$C_{XY}(t,\tau) = R_{XX}(t,\tau) * h(\tau)$$

$$C_{YY}(t,\tau) = R_{XY}(t,\tau) * h(t).$$

Linear Systems Driven by Stationary Processes

When a linear system is driven by a stationary process, the second moments of the output process become especially simple to compute. Also, since stationary processes have well defined spectra, we can hope to compute the spectrum of the output.

As before, let us consider a system with input $X(t)$, impulse response $h(t)$, and output $Y(t)$. Let us assume that $X(t)$ is a stationary random process. As before, let us compute joint statistics of X and Y, and then the statistics of Y.

In particular, let's consider

$$E[X(t+\tau)Y(t)]$$

$$= E\left[X(t+\tau)\int_{-\infty}^{\infty} X(t-\alpha)\,h(\alpha)\,d\alpha\right]$$

$$= \int_{-\infty}^{\infty} E[X(t+\tau)X(t-\alpha)]\,h(\alpha)\,d\alpha$$

$$= \int_{-\infty}^{\infty} R_{xx}(\tau+\alpha)\,h(\alpha)\,d\alpha$$ ← using stationarity

$$= \int_{-\infty}^{\infty} R_{xx}(\tau-\alpha)\,h(-\alpha)\,d\alpha$$ ← using change of variables

$$\boxed{R_{x,y}(\tau) = R_{xx}(\tau) * h(-\tau)}$$ ← notice that this is only a function of τ! Let's call this $R_{x,y}(\tau)$.

In similar fashion, we can show that $Y(t)$ is in fact stationary, and in fact that

$$\boxed{R_{Y,Y}(\tau) = R_{X,Y}(\tau) * h(\tau)}$$

Combining equations, we thus obtain that

$$\boxed{R_{Y,Y}(\tau) = R_{XX}(\tau) * h(-\tau) * h(\tau)}$$

Finally, taking Fourier transforms,

$$S_{YY}(\omega) = S_{XX}(\omega) H^*(j\omega) H(j\omega)$$

$$\boxed{S_{YY}(\omega) = S_{XX}(\omega) |H(j\omega)|^2}$$

Finally, let's do an example:

The discrete-time case is coming soon!
So is the special Gaussian R.P. case