Lab 08 – Discrete-Time Random Processes: 1

1. Objectives

Probability is the mathematical framework for modeling and reasoning about uncertain situations in which the outcome of a process varies with some randomness when the process is repeated. Many of the phenomena that occur in natural or engineered systems have inherent uncertainty and are best characterized through probabilistic tools such as random variables and random processes. For example, the Normal (or Gaussian) random variables are often used in signal processing and communication systems to model noise and distortion of signals.

This series of two labs will cover some basic methods of analyzing random processes, culminating with the application of these methods to the radar detection problem. Note that these labs assume an introductory background in probability theory.

2. Random Variables

The following section contains an abbreviated review of some of the basic definitions associated with random variables. It also discusses the concept of an *observation* of a random event, and introduces the notion of an *estimator*.

2.1 Basic Definitions:

A *random variable* is a function that maps a set of possible outcomes of a random experiment into a set of real numbers. The probability of an event can then be interpreted as the probability that the random variable will take on a value in a corresponding subset of the real line. This allows a fully numerical approach to modeling probabilistic behavior.

A very important function used to characterize a random variable is the *cumulative distribution function (CDF),* defined as

$$F_X(x) = P(X \leq x) \qquad x \in (-\infty, \infty) \tag{1}$$

Here, X is the random variable, and $F_X(x)$ is the probability that X will take on a value in the interval $(-\infty, x]$.

The derivative of the cumulative distribution function, if it exists, is known as the *probability density function*, denoted as $f_X(x)$. The probability density has the following property:

$$\begin{aligned} \int_{t_0}^{t_1} f_X(x)dx &= F_X(t_1) - F_X(t_0) \\ &= P(t_0 < X \leq t_1) \ . \end{aligned} \tag{2}$$

Since the probability that X lies in the interval $(-\infty, \infty)$ equals one, the entire area under the density function must also equal one.

The mean μ_X and variance $\sigma^2{}_X$ of a random variable X are defined by:

$$\mu_X = E[X] = \int_{-\infty}^{\infty} x f_X(x)dx \tag{3}$$

$$\sigma_X^2 = E[(X - \mu_X)^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)dx \ . \tag{4}$$

A very important type of random variable is the *Gaussian* or *normal* random variable. A Gaussian random variable has a density function of the following form:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{1}{2\sigma_X^2}(x - \mu_X)^2\right) \ . \tag{5}$$

Note that a Gaussian random variable is completely characterized by its mean and variance. This is not necessarily the case for other types of distributions. Sometimes, the notation $X \sim N(\mu, \sigma^2)$ is used to identify X as being Gaussian with mean μ and variance σ^2.

2.2 Samples of a Random Variable:

Suppose some random experiment may be characterized by a random variable X whose distribution is is unknown. For example, suppose we are measuring a deterministic quantity v, but our measurement is subject to a random measurement error ε. We can then characterize the observed value, X, as a random variable, $X = v + \varepsilon$. If the distribution of X does not change over time, we may gain further insight into X by making several independent observations $\{X_1, X_2, \ldots, X_N\}$. These observations X_i, also known as *samples*, will be independent random variables and have the same distribution $F_X(x)$. In this situation, the X_i's are referred to as *i.i.d.*, for *independent* and *identically distributed*. We also sometimes refer to $\{X_1, X_2, \ldots, X_N\}$ collectively as a sample, or observation, of size N.

Suppose we want to use our observation $\{X_1, X_2, \ldots, X_N\}$ to *estimate* the mean and variance of X. Two important *estimators* are the *sample mean* and *sample variance* defined by:

$$\hat{\mu}_X = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{6}$$

$$\hat{\sigma}_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \hat{\mu}_X)^2 \,. \tag{7}$$

It is important to realize that these sample estimates are functions of random variables, and are therefore themselves random variables. Therefore, we can also talk about the mean and variance of of the estimators. For example, we can compute the mean and variance of the sample mean $\hat{\mu}_X$.

$$E[\hat{\mu}_X] = E\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \frac{1}{N}\sum_{i=1}^{N} E[X_i] = \mu_X \tag{8}$$

$$\begin{aligned} Var[\hat{\mu}_X] &= Var\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \frac{1}{N^2} Var\left[\sum_{i=1}^{N} X_i\right] \\ &= \frac{1}{N^2}\sum_{i=1}^{N} Var[X_i] = \frac{\sigma_X^2}{N} \end{aligned} \tag{9}$$

In both (8) and (9) we have used the i.i.d. assumption.

An estimate $\hat{a}$ for some parameter a which has the property E $[\hat{a}] = a$ is said to be an *unbiased* estimate. An estimator such that Var $[\hat{a}] \to 0$ as $N \to \infty$ is said to be *consistent*. These two properties are highly desirable because they imply that if a large number of samples are used, the estimate will be close to the true parameter.

Suppose X is a Gaussian random variable with mean 0 and variance 1. Use the MATLAB function *randn* to generate 1000 samples of X, denoted as $X_1, X_2, \ldots, X_{1000}$. See the online help for the *randn* function. Plot these samples using the MATLAB function *plot*. We will assume our generated samples are i.i.d. Write MATLAB functions to compute the sample mean and sample variance through equations (6) and (7) without using the predefined *mean* and *var* functions of MATLAB. Use your sample mean and variance MATLAB functions to compute the sample mean and sample variance of the samples you just generated.

Task 1:

For the above mentioned 1000 samples of a Gaussian random variable X with mean 0 and variance 1,

i. Submit the plot of samples of X.

ii. Submit the MATLAB code of functions that calculate the sample mean and sample variance.

iii. Submit the sample mean and sample variance you calculated.

iv. Why are these sample mean and sample variance not equal to the true mean (0) and true variance (1)?

2.3 Linear Transformation of a Random Variable:

A linear transformation of a random variable X has the following form:

$$Y = aX + b \qquad (10)$$

where a and b are real numbers, and $a \neq 0$. A very important property of Normal/Gaussian random variables and linear transformations is as follows:

Normality is Preserved by Linear Transformations

If X is a normal random variable with mean μ and variance σ^2, and if $a \neq 0$, b are scalars, then the random variable

$$Y = aX + b$$

is also normal, with mean and variance

$$\mathbf{E}[Y] = a\mu + b, \qquad \text{var}(Y) = a^2\sigma^2.$$

For example, in (10), if X is Gaussian then Y will also be Gaussian, but not necessarily with the same mean and variance.

Consider a linear transformation Y of a Gaussian random variable X with mean 0 and variance 1. Calculate the constants a and b which make the mean and the variance of Y 4 and 9, respectively.

Generate 1000 samples of X, and then calculate 1000 samples of Y by applying the linear transformation in equation (10), using the a and b that you just determined. Plot the resulting samples of Y, and use your functions to calculate the sample mean and sample variance of the samples of Y.

Task 2:

For the above mentioned 1000 samples of a Gaussian random variable X with mean 0 and variance 1,

i. Submit the linear transformation you used to generate 1000 samples of a Gaussian random variable Y with mean 4 and variance 9.
ii. Submit the MATLAB code to generate samples of Y from samples of X.
iii. Submit the plot of samples of Y.
iv. Submit the sample mean and sample variance of Y you calculated using your MATLAB functions from Task 1.

3. Joint Distributions

In this section, we will study the concept of a joint distribution. We will see that joint distributions characterize how two random variables are related to each other. We will also see that *correlation*, *covariance*, and *correlation coefficient* simpler measures (compared to joint distributions) of the dependencies between random variables, which can be very useful for analyzing both random variables and random processes.

3.1 Background on Joint Distributions

Sometimes we need to account for not just one random variable, but several. In this section, we will examine the case of two random variables–the so-called *bivariate* case–but the theory is easily generalized to accommodate more than two random variables.

The random variables X and Y have cumulative distribution functions (CDFs) $F_X(x)$ and $F_Y(y)$, also known as *marginal* CDFs. Since there may be an interaction between X and Y, the marginal statistics may not fully describe their behavior. Therefore we define a *bivariate*, or *joint* CDF as

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y).$$

If the joint CDF is sufficiently "smooth", we can define a joint probability density function,

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y).$$

Conversely, the joint probability density function may be used to calculate the joint CDF:

$$F_{X,Y}(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(s,t)\,ds\,dt.$$

The random variables X and Y are said to be *independent* if and only if their joint CDF (or PDF) is a separable function, which means

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

Informally, independence between random variables means that one random variable does not tell you anything about the other. As a consequence of the definition, if X and Y are independent, then the product of their expectations is the expectation of their product.

$$E[XY] = E[X]E[Y]$$

While the joint distribution contains all the information about X and Y, it can be very complex and is often difficult to calculate. In many applications, a simple measure of the dependencies of X and Y can be very useful. Three such measures are the *correlation*, *covariance*, and the *correlation coefficient*.

- Correlation

$$E[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{X,Y}(x,y) dx\, dy$$

- Covariance

$$E[(X - \mu_X)(Y - \mu_Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f_{X,Y}(x,y) dx dy$$

- Correlation coefficient

$$\rho_{XY} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} = \frac{E[XY] - \mu_X \mu_Y}{\sigma_X \sigma_Y}$$

The correlation coefficient may be viewed as a normalized version of the covariance. When cov(X,Y) = 0 (which also implies that correlation coefficient is zero), we say that X and Y are **uncorrelated**. A positive or negative covariance indicates that the value of $X - E[X]$ and $Y - E[Y]$, obtained in a single experiment, “tend” to have the same or the opposite sign respectively. Thus, the sign of covariance provides an important qualitative indicator of the relationship between X and Y. Also, notice that the independence implies uncorrelatedness. However, the converse is generally not true.

3.2 Samples of Two Random Variables

In the following experiment, we will examine the relationship between the scatter plots for pairs of random samples (X_i, Z_i) and their correlation coefficient. We will see that the correlation coefficient determines the shape of the scatter plot.

Let X and Y be independent Gaussian random variables, each with mean 0 and variance 1. We will consider the correlation between X and Z, where Z is equal to the following:

1. $Z = Y$
2. $Z = (X + Y)/2$
3. $Z = (3 * X + Y)/5$
4. $Z = (90 * X + Y)/100$

Notice that since Z is a linear combination of two Gaussian random variables, Z will also be Gaussian.

Use MATLAB to generate 1000 i.i.d. samples of X, denoted as $X_1, X_2, \ldots, X_{1000}$. Next, generate 1000 i.i.d. samples of Y, denoted as $Y_1, Y_2, \ldots, Y_{1000}$. For each of the four choices of Z, perform the following tasks:

1. Create samples of Z using your generated samples of X and Y.
2. Generate a scatter plot of the ordered pair of samples (X_i, Z_i). Do this by plotting points $(X_1, Z_1), (X_2, Z_2), \ldots, (X_{1000}, Z_{1000})$. In order to plot points without connecting them with lines, use the *plot* command with the '.' format: *plot(X, Z,'.')*. Use the command subplot(2,2,n) with (n=1,2,3,4) to plot the four cases for Z in the same figure. Be sure to label each plot using the *title* command.
3. Empirically compute an estimate of the correlation coefficient using your samples X_i and Z_i and the following formula.

$$\hat{\rho}_{XZ} = \frac{\sum_{i=1}^{N}(X_i - \hat{\mu}_X)(Z_i - \hat{\mu}_Z)}{\sqrt{\sum_{i=1}^{N}(X_i - \hat{\mu}_X)^2 \sum_{i=1}^{N}(Z_i - \hat{\mu}_Z)^2}}$$

Task 3:

For the above mentioned 1000 samples of Gaussian random variables X, Y, and Z:

i. Submit your scatter plots of (Xi, Zi) for the four cases. Note the empirically calculated $\hat{\rho}_{XZ}$ on each plot.

ii. Explain how the scatter plots are related to correlation coefficients.

References

[1] A. Papoulis, *Probability, Random Variables, and Stochastic Processes, 3rd ed.*, McGraw-Hill, New York, 1991.

[2] C. Bouman, Lab Assignment for the Purdue University Course: "Digital Signal Processing with Applications"

[3] D. Berstekas, J, Tsitsiklis, *Introduction to Probability, 2nd ed.*, Athena Scientific, 2008.