Brief Review of Probability and Random Processes

John MacLaren Walsh, Ph.D.

ECES 632, Winter Quarter, 2010

1 Probability Space (Leon-Garcia 2.1-2.2)

A probability space is the collection $(\Omega, \mathscr{F}, \mathbb{P})$ of:

- a sample space Ω,
- a sigma algebra $\mathscr{F}$, which is a collection of subsets of Ω,
- and a probability measure $\mathbb{P}$ assigning probabilities to each set $\mathscr{A} \in \mathscr{F}$

The sigma algebra $\mathscr{F}$ must contain the empty set $\emptyset$ (as well as the entire sample space Ω) and be closed under complements and countably infinite unions and intersections, i.e.

- $\emptyset \in \mathscr{F}$
- $\mathscr{A} \in \mathscr{F} \implies \mathscr{A}^c \in \mathscr{F}$
- $\mathscr{A}_{\mathtt{i}} \in \mathscr{F},\ \mathtt{i} \in \{1, 2, \ldots\}, \implies (\bigcup_{\mathtt{i}=1}^{\infty} \mathscr{A}_{\mathtt{i}}) \in \mathscr{F}$
- $\mathscr{A}_{\mathtt{i}} \in \mathscr{F},\ \mathtt{i} \in \{1, 2, \ldots\}, \implies (\bigcap_{\mathtt{i}=1}^{\infty} \mathscr{A}_{\mathtt{i}}) \in \mathscr{F}$

The probability measure $\mathbb{P} : \mathscr{F} \to [0, 1]$, i.e. is a map between the sigma algebra $\mathscr{F}$ and the interval $[0, 1]$ which assigns probabilities to sets in $\mathscr{F}$, and must satisfy $\mathbb{P}(\Omega) = 1$ and *countable additivity*, which states that for any sequence of disjoint sets $\mathscr{A}_{\mathtt{i}} \in \mathscr{F},\ \mathtt{i} \in \mathbb{N}$, so that $\mathscr{A}_{\mathtt{j}} \bigcap \mathscr{A}_{\mathtt{i}} = \emptyset$ for all $\mathtt{j} \neq \mathtt{i}$ we have

$$\mathbb{P}\left(\bigcup_{\mathtt{i}=1}^{\infty} \mathscr{A}_{\mathtt{i}}\right) = \sum_{\mathtt{i}=1}^{\infty} \mathbb{P}\left(\mathscr{A}_{\mathtt{i}}\right)$$

From these requirements follow other familiar properties of a probability measure.

1.1 Total Probability

Suppose $\mathscr{F}$ contains sets $\{\mathscr{A}_i | i \in \{1, 2, \ldots\}\}$ which form a partition of Ω, so that $\Omega = \bigcup_{i=1}^{\infty} \mathscr{A}_i$, and $\mathscr{A}_i \bigcap \mathscr{A}_j = \emptyset\ i \neq j$. Then $\mathbb{P}(\mathscr{B}) = \sum_{i=1}^{\infty} \mathbb{P}(\mathscr{A}_i \bigcap \mathscr{B})$.

2 Random Variables

A random variable X is a map from the sample space to the real numbers $X : \Omega \to \mathbb{R}$ such that for all $x \in \mathbb{R}$, $\{\omega \in \Omega | X(\omega) \leq x\} \in \mathscr{F}$. Notation: capital (red) letters for the random variable, lower case (black) letters for the value it takes. (We will not provide explicit constructions of sigma algebras for continuous random variables for this course, however, the interested reader can look up material on the Borel sigma algebra, which is the sigma algebra generated by the open sets on $\mathbb{R}$ and the idea of a measurable map. A good graduate level mathematics intensive book tailored to operations research audiences is *Probability Essentials, 2nd ed.* by Jean Jacod and Philip Protter (Springer).)

2.1 Distribution Function Descriptions (Leon-Garcia 3.1-3.3, 4.5-4.6)

Owing to the complexity and situation dependency of the generic abstract sample space definition of a random variable, it is common to work with random variables' distribution functions.

We use the **cumulative distribution function (CDF)**:

$$\mathsf{F}_X(x) := \mathbb{P}[X \le x] = \mathbb{P}[\{\omega \in \Omega \,|\, X(\omega) \le x\}]$$

and, when it exists, the probability density function (PDF)

$$\mathsf{f}_X(x) := \frac{\partial \mathsf{F}_X}{\partial x}$$

(We can use weighted Dirac delta functions to handle a finite number of discontinuities in the CDF.) For discrete random variables, we can use the **probability mass function**:

$$\mathsf{p}_X(\mathtt{i}) := \mathbb{P}[X = \mathtt{i}] = \mathbb{P}[\{\omega \,|\, X(\omega) = \mathtt{i}\}]$$

(This distinction is necessary because the probability that a continuous random variable takes a particular value is zero.)

Multiple random variables can be similarly handled using the joint CDF

$$\begin{aligned} \mathsf{F}_{X_1,X_2,\ldots,X_N}(x_1,x_2,\ldots,x_N) &:= \mathbb{P}[X_1 \le x_1,\ X_2 \le x_2,\ldots,X_N \le x_N] \quad (1) \\ &= \mathbb{P}[\{\omega \in \Omega \,|\, X_1(\omega) \le x_1,\ X_2(\omega) \le x_2,\ldots,X_N(\omega) \le x_N\}] \quad (2) \end{aligned}$$

and the joint PDF

$$\mathsf{f}_{X_1,X_2,\ldots,X_N}(x_1,x_2,\ldots,x_N) = \frac{\partial^N \mathsf{f}_{X_1,X_2,\ldots,X_N}(x_1,x_2,\ldots,x_N)}{\partial x_1 \partial x_2 \cdots \partial x_N}$$

Note that

$$\mathsf{F}_{X_1}(x_1) = \mathsf{F}_{X_1,X_2,\ldots,X_N}(x_1,\infty,\ldots,\infty), \quad \mathsf{f}_{X_1}(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathsf{f}_{X_1,X_2,\ldots,X_N}(x_1,x_2,\ldots,x_N)\mathsf{d}x_2 \cdots \mathsf{d}x_N$$

$X_1, X_2, \ldots, X_N$ are **independent** if

$$\begin{aligned} \mathsf{F}_{X_1,X_2,\ldots,X_N}(x_1,x_2,\ldots,x_N) &= \mathsf{F}_{X_1}(x_1)\mathsf{F}_{X_2}(x_2)\cdots\mathsf{F}_{X_N}(x_N) \quad (3) \\ \mathsf{f}_{X_1,X_2,\ldots,X_N}(x_1,x_2,\ldots,x_N) &= \mathsf{f}_{X_1}(x_1)\mathsf{f}_{X_2}(x_2)\cdots\mathsf{f}_{X_N}(x_N) \quad (4) \end{aligned}$$

Notation: Will collect $X_1, X_2, \ldots, X_N$ into a vector $\mathbf{X}$. We will use $\mathbf{X} \sim \mathsf{F}_{\mathbf{X}}$ to specify that $\mathbf{X}$ is distributed according to the cumulative distribution function $\mathsf{F}_{\mathbf{X}}$ and $\mathbf{X} \sim \mathsf{f}_{\mathbf{X}}$ to specify that $\mathbf{X}$ is distributed according to the probability density function $\mathsf{f}_{\mathbf{X}}$.

2.2 Common RVs (Leon-Garcia 3.4, 4.8)

The PDFs, CDFs, means, and variances of several common random variables can be found in your textbook sections 3.4 and 4.8. Please take a moment to look over those sections and familiarize yourself with the random variables now.

2.3 Expectation (Leon-Garcia 3.6, 4.7)

Expected value of a random variable X

$$\mathbb{E}[X] := \int x \mathsf{f}_X(x)\mathsf{d}x$$

Expected value of a function of a random variable, e.g. $g(X)$ is similarly

$$\mathbb{E}[g(X)] := \int g(x)\mathsf{f}_X(x)\mathsf{d}x$$

Probability can be considered to be expectation of an indicator function.

The **mean** of a random variable is is expected value, while its **variance** $\mathsf{VAR}(X)$ is defined as $\mathbb{E}[(X - \mathbb{E}[X])^2]$. The covariance between two random variables X and Y is defined as

$$\mathsf{COV}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

the **correlation coefficient** between two random variables X and Y is then defined as

$$\rho_{X,Y} := \frac{\mathsf{COV}(X, Y)}{\sqrt{\mathsf{VAR}(X)\mathsf{VAR}(Y)}}$$

For a column random vector $\mathbf{X}$, define the **covariance matrix**

$$\mathbf{\Sigma}_{\mathbf{X},\mathbf{X}} := \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T\right]$$

so that the $\mathtt{i}, \mathtt{j}$th element of the matrix is

$$[\mathbf{\Sigma}_{\mathbf{X},\mathbf{X}}]_{\mathtt{i},\mathtt{j}} := \mathsf{COV}(X_{\mathtt{i}}, X_{\mathtt{j}})$$

The **Markov inequality** says that

$$\mathbb{P}\left[|X| \geq a\right] \leq \frac{\mathbb{E}[|X|]}{a}$$

The **Chebyshev inequality** says that for a random variable X with mean m, variance σ^2

$$\mathbb{P}\left[|X - m| \geq a\right] \leq \frac{\sigma^2}{a^2}$$

2.4 Transformations Between RVs (Leon-Garcia 3.5, 4.6)

Suppose we have a injective (1 to 1) transformation G from an open set $\mathscr{O} \subset \mathbb{R}^N$ to $\mathbb{R}^N$. Because it is injective, we may find a unique inverse transformation $\mathsf{G}^{-1} : \mathsf{G}(\mathscr{O}) \to \mathscr{O}$ so that $\mathsf{G}^{-1}(\mathsf{G}(\mathbf{x})) = \mathbf{x}$ for any $\mathbf{x} \in \mathscr{O}$. Suppose also that G is continuously differentiable, so that we can define a matrix of partial derivatives

$$\mathbf{J}(\mathbf{y}) := \left[\frac{\partial \mathsf{G}_{\mathtt{i}}^{-1}(\mathbf{y})}{\partial y_{\mathtt{j}}}\right]$$

and that the determinant of this matrix is not equal to zero anywhere on $\mathscr{O}$. Now suppose we define the random vector $\mathbf{Y}$ to be the result of operating on $\mathbf{X}$ which takes values in $\mathscr{O}$ with the transformation G, so that

$$\mathbf{Y} := \mathsf{G}(\mathbf{X})$$

We wish to determine the probability density function for $\mathbf{Y}$ from the probability density function for $\mathbf{X}$. It is

$$\mathsf{f}_{\mathbf{Y}}(\mathbf{y}) = \begin{cases} \mathsf{f}_{\mathbf{X}}(\mathsf{G}^{-1}(\mathbf{y}))|\det \mathbf{J}(\mathbf{y})| & \mathbf{y} \in \mathsf{G}(\mathscr{O}) \\ 0 & \text{otherwise} \end{cases}$$

When the transformation of interest G is not injective (for instance it maps a two dimensional vector to a one dimensional vector), sometimes it can be made into an injection by properly augmenting it $\mathbf{y}' = (\mathsf{G}(\mathbf{x}), x_{i_1}, \ldots, x_{i_U})$. Then one can use the theorem above to calculate the distribution for $\mathbf{y}'$ and then integrate out $x_{i_1}, \ldots, x_{i_U}$ to get the desired distribution. When this does not work either, there is still another theorem that one can use when there are a finite number of smooth solutions. See your book.

3 Conditional Probability (Leon-Garcia 2.4, 4.4)

Suppose that we know that a particular event $\mathscr{A} \in \mathscr{F}$ occurred, given this knowledge, the probability that another event $\mathscr{B} \in \mathscr{F}$ occurred is called the conditional probability $\mathbb{P}[\mathscr{B}|\mathscr{A}]$ and is defined by

$$\mathbb{P}[\mathscr{B}|\mathscr{A}] := \begin{cases} \frac{\mathbb{P}[\mathscr{A} \bigcap \mathscr{B}]}{\mathbb{P}[\mathscr{A}]} & \mathbb{P}[\mathscr{A}] \neq 0 \\ \text{undefined} & \mathbb{P}[\mathscr{A}] = 0 \end{cases}$$

3.1 Bayes' Theorem

Suppose $\mathscr{A}_{\mathtt{i}} \in \mathscr{F}$, $\mathtt{i} \in \{1, \ldots, N\}$ form a partition of the sample space Ω, so that $\bigcup_{\mathtt{i}=1}^{N} \mathscr{A}_{\mathtt{i}} = \Omega$ and $\mathscr{A}_{\mathtt{i}} \bigcap \mathscr{A}_{\mathtt{j}} = \emptyset$ for all $\mathtt{j} \neq \mathtt{i}$, then by total probability

$$\mathbb{P}[\mathscr{A}_{\mathtt{i}} | \mathscr{B}] = \frac{\mathbb{P}[\mathscr{B} | \mathscr{A}_{\mathtt{i}}] \mathbb{P}[\mathscr{A}_{\mathtt{i}}]}{\sum_{\mathtt{j}=1}^{N} \mathbb{P}[\mathscr{B} | \mathscr{A}_{\mathtt{j}}] \mathbb{P}[\mathscr{A}_{\mathtt{j}}]}$$

3.2 Conditional Distribution

X, Y random variables with joint density $\mathsf{f}_{X,Y}$. The conditional distribution for X given $Y = y$ is

$$\mathsf{F}_{X|Y}(x|y) := \lim_{\Delta \downarrow 0} \frac{\mathbb{P}\left[X \leq x, y < Y \leq y + \Delta\right]}{\mathbb{P}\left[y < Y \leq y + \Delta\right]}$$

The conditional density is then

$$\mathsf{f}_{X|Y}(x, y) := \frac{\mathsf{f}_{X,Y}(x, y)}{\mathsf{f}_{Y}(y)}$$

Example: Suppose $\mathbf{Z} := [\mathbf{X}, \mathbf{Y}]$ is a jointly Gaussian random vector with covariance matrix $\begin{bmatrix} \mathbf{\Sigma_{X,X}} & \mathbf{\Sigma_{X,Y}} \\ \mathbf{\Sigma_{Y,X}} & \mathbf{\Sigma_{Y,Y}} \end{bmatrix}$. Then $\mathsf{f}_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y})$ viewed as a function of $\mathbf{x}$ is a jointly Gaussian probability density with mean

$$\mathbb{E}\left[\mathbf{X} | \mathbf{Y} = \mathbf{y}\right] := \mathbb{E}\left[\mathbf{X}\right] + \mathbf{\Sigma_{X,Y}} \mathbf{\Sigma_{Y,Y}^{-1}} \left(\mathbf{y} - \mathbb{E}\left[\mathbf{Y}\right]\right)$$

and covariance matrix

$$\mathbf{\Sigma_{X|Y}} := \mathbf{\Sigma_{X,X}} - \mathbf{\Sigma_{X,Y}} \mathbf{\Sigma_{Y,Y}^{-1}} \mathbf{\Sigma_{Y,X}}$$

4 Random Processes

4.1 Basic Definitions (Leon-Garcia 6.1-6.4)

A random process (also known as a stochastic process) is a map $X : \mathscr{T} \times \Omega \to \mathbb{R}$ between the sample space and signals, so that to every outcome $\omega \in \Omega$ there is a signal (a function of time)

$$X(t, \omega) \quad t \in \mathscr{T}$$

As in deterministic signals and systems $\mathscr{T}$ can be the real numbers $\mathbb{R}$, in which case X is a continuous time random process, or the integers $\mathbb{Z}$, in which case X is a discrete time random process.

According to the **Kolmogorov extension theorem** (see, for example *Probability: Theory and Examples, 3rd. Ed.* by Rick Durrett) a discrete time random process is uniquely specified by its finite dimensional distributions. We will also specify a continuous time random process by its finite dimensional distributions. Here, by finite dimensional distributions we mean the set of all joint cumulative distribution functions $\mathsf{F}_{X(t_1), X(t_2), \ldots, X(t_N)}(x(t_1), x(t_2), \ldots, x(t_N))$ for all time instants $t_1, t_2, \ldots, t_N \in \mathscr{T}$ and all $N \in \mathbb{N}$.

We will often work with the mean of a random process

$$m_X(t) := \mathbb{E}\left[X(t)\right]$$

as well as the **auto-correlation** function

$$\mathsf{R}_X(t_1, t_2) := \mathbb{E}\left[X(t_1) X(t_2)\right] = \int xy \mathsf{f}_{X(t_1), X(t_2)}(x, y) \mathsf{d}x \mathsf{d}y$$

and the **auto-covariance** function

$$\mathsf{C}_X(t_1, t_2) := \mathbb{E}\left[(X(t_1) - m_X(t_1))(X(t_2) - m_X(t_2))\right] = \mathsf{R}_X(t_1, t_2) - m_X(t_1) m_X(t_2)$$

4.2 Stationary and Wide-Sense Stationary Random Processes (Leon-Garcia 6.5)

A **stationary** random process is one for which the joint distribution for the random process at any set of times $t_1, t_2, \ldots, t_N$ is invariant when replaced by the joint distribution at times $t_1+\tau, t_2+\tau, \ldots, t_N+\tau$ for any $\tau \in \mathscr{T}$.

$$\mathsf{F}_{X(t_1),X(t_2),\ldots,X(t_N)}(x_1, x_2, \ldots, x_N) = \mathsf{F}_{X(t_1+\tau),X(t_2+\tau),\ldots,X(t_N+\tau)}(x_1, x_2, \ldots, x_N) \quad (5)$$

Thus a stationary random process must have a mean which is time invariant.

More generally, a **wide sense stationary** (WSS) random process is one for which the mean is constant $m_X(t) = m \forall t \in \mathscr{T}$ and auto-correlation function $\mathsf{R}_X(t_1, t_2)$ which is a function of $t_2 - t_1$ alone, in which case we write it as $\mathsf{R}_X(\tau),\ \tau \in \mathscr{T}$.

Many processes are not stationary, but rather **cyclo-stationary** so that (5) only holds for τ as some multiple of a fundamental period T, i.e. for all $\tau = mT,\ m \in \mathbb{Z}$.

4.2.1 Power Spectral Density (Leon-Garcia 7.1)

For a wide sense stationary random process, we define the the Fourier transform of the auto-correlation sequence (discrete time fourier transform if $\mathscr{T}$ is $\mathbb{Z}$ or continuous time Fourier transform if $\mathscr{T}$ is $\mathbb{R}$) to be the **Power Spectral Density**. That is, for a discrete time random process $X(t)$, the power spectral density (PSD) is

$$\mathsf{S}_X(f) = \sum_{t=-\infty}^{\infty} \mathsf{R}_X(t) e^{-j2\pi f t}$$

which is periodic in 1 and is often plotted for $f \in [0, 1]$ or $f \in [-\frac{1}{2}, \frac{1}{2}]$. For a continuous time random process, the PSD is

$$\mathsf{S}_X(f) := \int_{-\infty}^{\infty} \mathsf{R}_X(t) e^{-j2\pi f t} \mathsf{d}t$$

with $f \in \mathbb{R}$.

4.2.2 LTI Filtering of WSSRPs

Wide sense stationary random processes are special because the response of a linear time invariant system to a WSSRP is a WSSRP. To see this, let us focus on the discrete time case. Consider the WSSRP $X[n]$ with mean m_X and autocorrelation function $\mathsf{R}_X[\tau]$ entering into a LTI system with real valued impulse response $h[n]$. The output of the system will be

$$Y[n] := \sum_{k=-\infty}^{\infty} h[k] X[n-k]$$

The mean of the random process $Y[n]$ is then

$$\mathbb{E}[Y[n]] = \sum_{k=-\infty}^{\infty} h[k] m_X = m_Y$$

which is constant, as it should be if $Y[n]$ is to be a WSSRP. The autocorrelation of $Y[n]$ is then

$$\begin{aligned}
\mathbb{E}[Y[n]Y[m]] &= \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} h[k]h[l]\mathbb{E}\left[X[n-k]X[m-l]\right] && (6)\\
&= \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} h[k]h[l]\mathsf{R}_X[n-k-(m-l)] && (7)\\
&= \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} h[k]h[l]\mathsf{R}_X[n-m+l-k] && (8)\\
&= \mathsf{R}_Y[n-m] && (9)
\end{aligned}$$

which shows that the signal $Y[n]$ is a WSSRP. Next note that the sum in (8) is actually the convolution of R_X (shown in class) with $h[k]$ followed by the convolution with $h[-k]$. Recall that for a real valued $h[k]$, the DTFT of $h[-k]$ is the complex conjugate of the DTFT of $h[k]$. Because convolution in time corresponds to multiplication in frequency we have the relation

$$\mathsf{S}_Y(f) = |H(f)|^2 \mathsf{S}_X(f)$$

4.3 Markov Processes (Leon-Garcia 6.2)

A **Markov random process** is one for which the future is independent of the past given the present. Thus, for arbitrary times $t_1 < t_2 < \cdots < t_N$

$$\mathbb{P}[a < X(t_N) \le b | X(t_1) = x_1, X(t_2) = x_2, \ldots, X(t_{N-1}) = x_{N-1}] = \mathbb{P}[a < X(t_N) \le b | X(t_{N-1}) = x_{N-1}]$$

When such a density exists, this translates to

$$\mathsf{f}_{X(t_N)|X(t_1),X(t_2),\ldots,X(t_{N-1})}(x_N | x_1, x_2, \ldots, x_{N-1}) = \mathsf{f}_{X(t_N)|X(t_{N-1})}(x_N | x_{N-1})$$

4.4 Gaussian Processes (Leon-Garcia 6.2)

A random process $X(t)$ is a Gaussian random process if its finite dimensional distributions are all joint Gaussian distributions. Thus, a Gaussian random process is completely specified by its mean function $m(t) := \mathbb{E}[X(t)]$ and auto-covariance $\mathsf{C}_X(t_1, t_2)$.

4.5 AR, MA, and ARMA Time Series Models (Marple, Chapter 6, Leon-Garcia 7.3)

Several classes of models which have held longstanding historical importance in digital signal processing are the **auto-regressive**, **moving average**, and **auto-regressive moving average** models.

In short, an **auto-regressive moving average** model of order (p, q) (denoted ARMA(p,q)) is the output of a IIR filter with rational transfer function

$$H(z) := \frac{B(z)}{A(z)}$$

where

$$A(z) := \sum_{t=0}^{p} a_t z^{-t}, \quad B(z) := \sum_{t=0}^{q} b_t z^{-t}, \quad a_0 = 1, b_0 = 1$$

driven by an input signal that is white noise (Gaussian RP with mean 0 and correlation function $\mathsf{R}(t_1, t_2) = \sigma^2 \delta(t_1 - t_2)$). We will require the polynomials $B(z)$ and $A(z)$ to have all of their zeros within the unit circle to require $H(z)$ to be a stable casual minimum-phase filter (e.g. $H(z)$ is SCAMP = stable causal monic (h[0]=1) and minimum phase).

An AR(p) model is an ARMA(p) model and a MA(q) model is a ARMA(0,q) model.