E231 Mathematical Methods in Engineering
Notes on Probability Theory and Random Sequences

Professor Kameshwar Poolla
Departments of Mechanical Engineering and EECS
University of California at Berkeley

---

1 Preliminaries

1.1 Basic Set Theory

1.1.1 Unions, Intersections and Complements

Let us first fixed a given set X and consider the set $\mathcal{S}(\mathcal{X})$ consisting of all subsets of X.

Let A and B be subsets of X.

1. **Intersection of A and B**, denoted by $A \cap B$ is the set of elements that belong to both A and B.

   $$A \cap B = \{x \mid x \in A \textbf{ and } x \in B\}$$

2. **Union of A and B**, denoted by $A \cup B$ is the set of elements that are in either A or B.

   $$A \cup B = \{x \mid x \in A \textbf{ or } x \in B\}$$

3. **Complement of A, A^c (relative to X)**, is the set of elements that are in X but not in A

   $$A^c = \{x \mid x \in X : x \notin A\}$$

   Other frequently used symbols for A^c are $\tilde{A}$ and $\sim A$.

4. **Complement of A relative to B Also known as difference), $A\backslash B$,** is the set of elements that are in B but not in A

   $$A\backslash B = \{x \mid x \in B \textbf{ and } x \notin A\}$$

   $$\mathcal{S}_1\backslash\mathcal{S}_2 = \{x \mid x \in \mathcal{S}_2 \textbf{ and } x \notin \mathcal{S}_1\}$$

   Another frequently used symbol for $A\backslash B$ is $A \sim B$.

5. **Symmetric difference between A and B, $A \triangle B$,** is defined by

   $$A \triangle B = (A\backslash B) \cup (B\backslash A)$$

   $$\mathcal{S}_1 \triangle \mathcal{S}_2 = \{x \mid x \notin \mathcal{S}_2 \textbf{ and } x \notin \mathcal{S}_1\}$$

6. **De Morgan's laws:**

$$\begin{aligned}(A \cup B)^c &= A^c \cap B^c \\ (A \cap B)^c &= A^c \cup B^c\end{aligned}$$

1.2 Countability

A set A is *countable* if it is in the range of some sequence (could be infinite) and it is finite if it is in the range of a finite sequence. In other words, A is countable if one can enumerate its elements $A = \{a_1, a_2, a_3, \cdots\}$.

- A subset of a countable set is also countable.
- If the sets A_n are countable for $n \geq 1$ so its their union

$$A = \cup_{n=1}^{\infty} A_n := \{a \mid a \in A_i \text{ for some } i \geq 1\}$$

- The cartesian product $A \times B = \{(a, b) \mid a \in A \text{ and } b \in B\}$ of countable sets is also countable.

Example 1.1 The set of rational numbers is countable. However, the set $[0, 1]$ is not countable.

1.3 Algebra of Sets

A collection $\mathcal{A}$ of subsets of X is called an algebra of sets is

(i) $A \cup B$ is in $\mathcal{A}$ whenever A and B are.

(ii) A^c is is in $\mathcal{A}$ whenever A is.

(iii) (Follows from Morgan's laws) $A \cap B$ is in $\mathcal{A}$ whenever A and B are.

Proposition 1.2 *Given any collection $\mathcal{C}$ of subsets of X, there is a smallest algebra $\mathcal{A}$ that contains $\mathcal{C}$; that is, there is an algebra $\mathcal{A}$ containing $\mathcal{C}$ and such that if $\mathcal{B}$ is any algebra containing $\mathcal{C}$, then $\mathcal{B}$ contains $\mathcal{A}$.*

The smallest algebra containing $\mathcal{C}$ is called the algebra generated by $\mathcal{C}$.

Proposition 1.3 *Let $\mathcal{A}$ be an algebra of subsets and $< A_i >$ a sequence of sets in $\mathcal{A}$. Then there is a sequence $< B_i >$ of sets in $\mathcal{A}$ such that $B_n \cap B_m = \varnothing$ for $n \neq m$ and*

$$\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i$$

Proof: Set $B_i = A_i$, and for each $n > 1$ define

$$B_n \;=\; A_n \setminus [\, A_1 \cup A_2 \cup \cdots \cup A_{n-1} \,] = A_n \cap A_1^c \cap A_2^c \cap \cdots \cap A_{n-1}^c$$

Since the complements and intersections of sets in the algebra $\mathcal{A}$ are also in $\mathcal{A}$, we have that $B_n \in \mathcal{A}$. We also have that $B_n \subset A_n$. Let B_n and B_m be two such sets and suppose that $m < n$. Then $B_m \subset A_m$, and therefore

$$\begin{aligned} B_m \cap B_n \subset A_m \cap B_n &= A_m \cap \left(A_n \cap A_1^c \cap \cdots \cap A_m^c \cap \cdots \cap A_{n-1}^c\right) \\ &= (A_m \cap A_m^c) \cap \cdots \\ &= \varnothing \end{aligned}$$

Since $B_i \subset A_i$, we have

$$\bigcup_{i=1}^{\infty} B_i \subset \bigcup_{i=1}^{\infty} A_i$$

Let $x \in \bigcup_{i=1}^{\infty} A_i$. Then x must belong to at least on e of the A_i's. Let n be the smallest value of i such that $x \in A_n$. Then $x \in B_n$ and so $x \in \bigcup_{n=1}^{\infty} B_n$. Thus,

$$\bigcup_{n=1}^{\infty} A_n \subset \bigcup_{n=1}^{\infty} B_n$$

and

$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n$$

$\diamondsuit$.

Definition: *An algebra $\mathcal{A}$ of sets is called a* **σ-algebra** *or a* **Borel field** *if every union of countable collection of sets in $\mathcal{A}$ is again in $\mathcal{A}$. That is, if $< A_i >$ a sequence of sets in $\mathcal{A}$, then*

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$$

$$\bigcap_{i=1}^{\infty} A_i \in \mathcal{A} \text{ (from De Morgan's laws)}$$

Proposition 1.4 *Given any collection $\mathcal{C}$ of subsets of X, there is a smallest σ-algebra that contains $\mathcal{C}$; that is, there is an σ-algebra $\mathcal{A}$ containing $\mathcal{C}$ and such that if $\mathcal{B}$ is any σ-algebra containing $\mathcal{C}$, then $\mathcal{A} \subset \mathcal{B}$.*

Proof: Let $\mathcal{F}$ be the family of all σ-algebras of subsets of X that contain $\mathcal{C}$ and let

$$\mathcal{A} = \bigcap \{\mathcal{B} \mid \mathcal{B} \in \mathcal{F}\}$$

Then $\mathcal{C}$ is a subcollection of $\mathcal{A}$, since each $\mathcal{B}$ in $\mathcal{F}$ contains $\mathcal{C}$. Moreover, $\mathcal{A}$ is a σ-algebra for if A and B are in $\mathcal{A}$, then for each $\mathcal{B} in \mathcal{F}$ we have $A \in \mathcal{B}$ and $B \in \mathcal{B}$. Since $\mathcal{B}$ is a *sigma*-algebra,

$A \cup B \in \mathcal{B}$. SInce this is true for every $\mathcal{B} \in \mathcal{F}$, we have $A \cup B \in \bigcap\{\mathcal{B} \,|\, \mathcal{B} \in F\}$. Similarly, we see that if $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$. Thus, from the definition of $\mathcal{A}$, it follows that is $\mathcal{B}$ is a σ-algebra containing $\mathcal{C}$, then $\mathcal{A} \subset \mathcal{B}$. $\diamondsuit$.

The smallest σ-algebra containing $\mathcal{C}$ is called the *sigma*-algebra generated by $\mathcal{C}$.

Let $\{B_n\}_1^\infty$ be a collection of sets. Suppose $B_{n+1} \supseteq B_n$ for all n. Then it is natural to write

$$\lim_{n\to\infty} B_n = \cup_{n=1}^\infty B_n$$

Note that $\omega \in \lim_{n\to\infty} B_n$ if and only if $\omega \in B_n$ for some n. Similarly if $\{C_n\}_1^\infty$ is a collection of sets nested as $C_{n+1} \subseteq C_n$, we shall write

$$\lim_{n\to\infty} C_n = \cap_{n=1}^\infty C_n$$

Note that $\omega \in \lim_{n\to\infty} C_n$ if and only if $\omega \in C_n$ for some n. Next, let $\{A_n\}_1^\infty$ be a collection of sets and define

$$B_n = \cap_{k=n}^\infty A_k \quad C_n = \cup_{k=n}^\infty A_k$$

It is clear that we have the containments $B_{n+1} \supseteq B_n$ and $C_{n+1} \subseteq C_n$ for all n. We *define*

$$\begin{aligned} \limsup A_n &= \lim_{n\to\infty} B_n = \cup_{n=1}^\infty \cap_{k=n}^\infty A_k \\ \liminf A_n &= \lim_{n\to\infty} C_n = \cap_{n=1}^\infty \cup_{k=n}^\infty A_k \end{aligned}$$

It is easy to show that $\limsup A_n \supseteq \liminf A_n$. The sequence of sets $\{A_n\}_1^\infty$ is said to *converge to the set* A, written $A_n \to A$ if

$$\limsup A_n = \liminf A_n = A$$

Exercise 1.5 Show that the sets lim sup and lim inf defined above have the following interpretation:

$$\begin{aligned} \limsup A_n &= \{\omega :\ \omega \in A_n \text{ eventually }\} \\ \liminf A_n &= \{\omega :\ \omega \in A_n \text{ infinitely often }\} \end{aligned}$$

Exercise 1.6 Show that $\limsup A_n \supseteq \liminf A_n$.

2 Probability Spaces

Consider an *experiment* $\mathcal{E}$ that can be conducted many times. Each *trial* of $\mathcal{E}$ results in a (possibly different) *outcome* ω.

- The *sample space* Ω is the set of all possible outcomes ω.

- We declare certain subsets of Ω to be *events* and we will assign "probabilities" to these special subsets. The set of events is $\mathcal{A}$. More precisely, $\mathcal{A}$ is a collection of subsets of Ω, which is closed under countable set operations - such collection is called a σ-algebra. The elements of $\mathcal{A}$ are called events.

- A *probability space* is a triple $(\Omega, \mathcal{A}, Prob)$ where $Prob : \mathcal{A} \to \mathbb{R}$ is a function, called the *probability* that satisfies the following axioms:

  (a) $0 \leq Prob(A) \leq 1$ for all $A \in \mathcal{A}$

  (b) $Prob(\Omega) = 1$

  (c) If $A \cap B = \phi$, then $Prob(A \cup B) = Prob(A) + Prob(B)$

  (d) If $\{A_n\}_1^\infty$ is a sequence of events (i.e. $A_n \in \mathcal{A}$) such that $A_n \to A$, then

  $$Prob(A_n) \to Prob(A)$$

Some useful probability function properties

Consider a probability space $(\Omega, \mathcal{A}, Prob)$ and let A, B and C be events. The following properties can be deduced from the axioms:

$$Prob(A \cup B) = Prob(A) + Prob(B) - Prob(A \cap B) \tag{1}$$

$$\begin{aligned} Prob(A \cup B \cup C) &= Prob(A) + Prob(B) + Prob(C) \\ &- Prob(A \cap B) - Prob(A \cap C) - Prob(B \cap C) \\ &+ Prob(A \cap B \cap C) \end{aligned} \tag{2}$$

Examples and Exercises

Example 2.1 Consider an experiment $\mathcal{E}$ that involves tossing two *fair* coins. The set of outcomes or the sample space is

$$\Omega = \{HH, HT, TH, TT\}$$

Let us declare the set of events $\mathcal{A}$ to be the set of all subsets, or the *power set*, of Ω. Thus $\mathcal{A}$ has $2^4 = 16$ events:

$$\begin{aligned} \mathcal{A} = \{ \quad & HH, HT, TH, TT \\ & \{HH, HT\}, \{HH, TH\}, \{HH, TT\}, \{HT, TH\}, \{HT, TT\}, \{TH, TT\} \\ & \{HH, HT, TH\}, \{HH, HT, TT\}, \{HH, TH, TT\}, \{HT, TH, TT\} \\ & \{HH, HT, TH, TT\}, \emptyset\} \end{aligned}$$

With the function $Prob : \mathcal{A} \to \mathbb{R}$ defined by

$$Prob(HH) = \ Prob(HT) = Prob(TH) = Prob(TT) = \frac{1}{4}$$

the triple $(\Omega, \mathcal{A}, Prob)$ forms a probability space. Note that we have not defined $Prob$ on all of $\mathcal{A}$ because, we can deduce this from the partial definition together with the axioms that $Prob$ must satisfy.

Example 2.2 Consider an experiment $\mathcal{E}$ that consists of measuring some voltage. The sample space is $\mathbb{R}$. Let us declare the set of events to be

$$\mathcal{A} = \{A : \ A \text{ is the union of finitely many intervals } \ [a, b)\}$$

Note that $\mathcal{A}$ is *not* the power set of $\mathbb{R}$, i.e. there are certain sets of outcomes for which we will not assign probabilities. Define the probability function by

$$Prob([a, b)) = \min\{b, 1\} - \max\{a, 0\}$$

Then, the triple $(\Omega, \mathcal{A}, Prob)$ forms a probability space.

Example 2.3 Establish Eq. (1) using the axioms defining a probability space. Notice that, since

$$(A \cup B) = A \cup (A^c \cap B) \qquad \text{and} \qquad B = (A \cap B) \cup (A^c \cap B)$$

$$A \cap (A^c \cap B) = \emptyset \qquad \text{and} \qquad (A \cap B) \cap (A^c \cap B) = \emptyset\,,$$

then

$$Prob(A \cup B) = Prob(A) + Prob(A^c \cap B) \qquad \text{and} \qquad Prob(B) = Prob(A \cap B) + Prob(A^c \cap B)$$

and the result follows.

Exercise 2.4 Establish Eqs. (2) using the axioms defining a probability space.

Exercise 2.5 Consider an experiment consisting of tossing N fair coins. Compute the probability of the event $A = \{\omega : \ k \text{ coins land heads}\}$.

Exercise 2.6 Two independent, uniformly distributed cuts are made on a meter stick. Find the probability that the resulting three pieces can be arranged to form a triangle.

Exercise 2.7 (Buffon) A one inch needle is dropped on an infinite two-dimensional sieve with thin horizontal wires that are spaced one inch apart. Assume the needle is dropped from a random, level orientation and that no subsequent rotation occurs. Determine the probability that the needle falls through the grating without touching any of the wires.

Basic Combinatorial Analysis

(a) **Permutations:** Let there be n distinct objects and define a permutation of these objects as a simple rearrangement of them. Then the number of permutation of these n objects is

$$Per(n) = n!$$

(b) **k-Permutations of n objects $P(n,k)$:** Let there be n distinct objects and let k be a positive integer such that $k < n$. The number of k-permutations of n distinct objects, i.e. the number of distinct ways that we can pick k elements out of n and arrange them in a sequence is

$$Per(n,k) = \frac{n!}{(n-k)!}$$

Example 2.8 The number of 2-permutations from the set $\{W, X, Y, Z\}$ is

$$P(4,2) = \frac{4!}{(4-2)!} = 12$$

namely WX, XW, WY, YW, WZ, ZW, XY, YX XZ, ZX, YZ and ZY.

(c) **Permutations with repeated elements:** Let there be k *types* of objects of which n_1 are of type 1, n_2 are of type 2, etc. Then the number of ways in which these $n_1 + n_2 + \cdots + n_k$ objects can be rearranged is

$$PerRep(n_1, \cdots, n_k]) = \frac{(\sum_{i=1}^{k} n_i)!}{\prod_{i=1}^{k}(n_i!)} = \frac{(n_1 + \cdots + n_k)!}{n_1! n_2! \cdots n_k!}$$

Example 2.9 The number of permutations of the letters in the word MASSACHUSETTS is obtained by grouping letters as follows: $n_{_M} = 1$, $n_{_A} = 2$, $n_{_S} = 4$, $n_{_C} = 1$, $n_{_H} = 1$, $n_{_U} = 1$, $n_{_T} = 2$ and computing

$$PerRep(n_{_M}, n_{_A}, n_{_S}, n_{_C}, n_{_H}, n_{_U}, n_{_T}]) = \frac{13!}{4!2!2!}$$

(d) **Combinations without repetitions:** Let there be n distinguishable objects and let $k \leq n$. A k-combination is a selection of k objects without regards to order. Then, the number of k-combinations of n distinguishable objects is

$$Com(n,k) = \left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$$

Example 2.10 The number of 2-combinations from the set $\{W, X, Y, Z\}$ is

$$Com(4,2) = \frac{4!}{2!(4-2)!} = 6$$

namely WX, WY, WZ, XY, XZ and YZ.

Example 2.11 A batch of 100 manufactured parts is checked by an inspector who examines 10 specimens selected at random. If none of the 10 specimens is defective, the inspector accepts the entire batch. Otherwise, the batch is examined further. What is the probability that a batch containing 20 defective parts will be accepted?

Define N as the total number of ways that we can pick 10 arbitrary samples from a batch of 100. This number is

$$N = Com(100, 10) = \begin{pmatrix} 100 \\ 10 \end{pmatrix} = \frac{100!}{10!\,90!}$$

Let M be the total number of ways of picking 10 components from 80 non-defective parts and no part from 20 defective parts. This number is

$$M = Com(80, 10) \times Com(20, 0) = Com(80, 10) = \frac{80!}{10!\,70!}$$

Since all parts are selected at random, the combinations are equiprobable. Thus, the probability of picking no defective parts from a sample of 10 parts is

$$Prob = \frac{M}{N} = \frac{80!}{10!\,70!} \times \frac{10!\,90!}{100!} = 0.0951$$

3 Conditioning and Independence

Let $(\Omega, \mathcal{A}, Prob)$ be a probability space and let $A, B \in \mathcal{A}$ be two events with $Prob(B) \neq 0$. We *define* the *conditional probability* of A given that B has occurred as

$$Prob(A|B) = \frac{Prob(A \cap B)}{Prob(B)} \tag{3}$$

This is called *Bayes' rule*, and $Prob(A|B)$ has the interpretation of "the chance event A has occurred given the information that event B has occurred".

We say that A and B are *independent events* if

$$Prob(A|B) = P(A) \;\text{ or equivalently, }\; Prob(A \cap B) = Prob(A)Prob(B) \tag{4}$$

This can be thought of as "the information that B has occurred does not alter the chance that A has also occurred".

Exercise 3.1 According to statistical data, it rains in Berkeley two out of three weekends. Forecasters predict the weather correctly with probability 0.8 if it rains and with probability 0.7 otherwise. If the forecast calls for rain, what is the probability that it will rain?

Remarks

- We can define the conditional probability $Prob_{|B}(A) = Prob(A|B)$ for any event A and it constitute a new probability measure. For instance, given three events A, B and C,

$$Prob_{|B}(A \cap C) = Prob_{|B}(A|C)\, Prob_{|B}(C) = Prob(A|B \cap C)\, Prob(C|B)$$

- Notice that two event being independent should not be confused with two event being disjoint. Indeed, if A and B are disjoint, then $Prob(A \cap B) = P(\emptyset) = 0.$, so that $Prob(A \cap B) = Prob(A)Prob(B)$ if $Prob(A) = 0$ and/or $Prob(B) = 0$. Therefore, knowing that A occurs implies that B does not not, unless B is impossible in the first place.

A collection of events $\{A_1, \cdots A_n, \cdots\}$ is called *independent* if for every finite sub-collection, we have

$$Prob\left(\cap_{i=1}^{m}\ A_{k_i}\right) = \Pi_{i=1}^{m}\ Prob\left(A_{k_i}\right) \tag{5}$$

Let B_i, $i = 1, \cdots, n$ be a collection of **disjoint** events, whose union is the sample space Ω and let A be another event. Therefore,

$$Prob(A) = \sum_{i=1}^{n} Prob(A \cap B_i) = \sum_{i=1}^{n} Prob(A|B_i)\, Prob(B_i)$$

Hence,

$$Prob(B_j|A) = \frac{Prob(B_j \cap A)}{P(A)} = \frac{Prob(A|B_j)\, Prob(B_j)}{\sum_{i=1}^{n} Prob(A|B_i)\, Prob(B_i)}$$

Example 3.2 The "Monty Hall" three-door problem asks a contestant to choose one of three doors, hoping to find the one door that conceals a prize. The other two doors conceal duds. After the contestant chooses, Monty Hall (the master of ceremonies of the Let's Make a Deal television show) opens one of the doors the player did not choose to reveal a dud. Then the contestants are permitted to stay with their original choice or switch to the other unopened door. Should the contestants switch their choice in order to increase their probability of wining?

The player's probability of getting the prize if she does not switch is 1/3. Lets now determine the player's probability of getting the prize if she switches. Let the doors be called x, y and x. Let C_x be the event that the prize is behind door x and so on. Let H_x be the event that the host opens door x and so on. Now lets assume that that the initial choice is door x, the probability that the contestant wins the prize if she then switches her choice is (why?)

$$Prob(H_z \cap C_y) + Prob(H_y \cap C_z) = Prob(H_z|C_y)\, Prob(C_y) + Prob(H_y|C_z)\, Prob(C_z)$$

Notice that $Prob(H_z|C_y)$ is the conditional probability that the hosts opens door z given that the prize is in door y. Since the initial choice is door x, $Prob(H_z|C_y) = 1$. By the same argument, $Prob(H_y|C_z) = 1$ and, since $Prob(C_x) = Prob(C_y) = Prob(C_z) = 1$, the probability of wining the prize by switching is

$$Prob(H_z \cap C_y) + Prob(H_y \cap C_z) = \frac{2}{3}.$$

4 Random Variables

We are particularly interested in experiments whose outcome is a real number, for example as in the measurement of a voltage. For experiments whose outcomes are not real numbers, like coin tossing, we can (perhaps unnaturally) assign real numbers to outcomes. For example, in the two-coin toss experiment we can declare $HH = 1, HT = 2, TH = 3, TT = 4$.
Let $(\Omega, \mathcal{A}, Prob)$ be a probability space. A *random variable* X is a *measurable* function

$$X : \Omega \to \mathbb{R}$$

Let $\mathcal{A}$ be a collection of events of Ω. (Recall that $\mathcal{A}$ is closed under countable set operations.) A function $[X : \Omega \to \mathbb{R}$ is $\mathcal{A}$-measurable is $X^{-1}((-\infty, a]) \in \mathcal{A}$ for all $a \in \mathbb{R}$. Therefore, we can define $Prob(X \leq a)$ for all $a \in \mathbb{R}$.

5 Distribution and Density Functions

Let $(\Omega, \mathcal{A}, Prob)$ be a probability space and let X be a random variable. The *probability distribution function* $P_X(x)$ associated with X is

$$P_X(x) = Prob\{\omega : \ X(\omega) < x\} = Prob\{X < x\}$$

It is easy to show that

Lemma 5.1 *(a)* $\lim_{x \to -\infty} \ P_X(x) = 0$

(b) $\lim_{x \to \infty} \ P_X(x) = 1$

(c) $P_X(x)$ *is monotone nondecreasing in* x

(d) $P_X(x)$ *is left-continuous in* x

In the event $P_X(x)$ is differentiable everywhere, we define the *probability density function* $p_X(x)$ associated with X as

$$p_X(x) = \frac{d}{dx} P_X(x)$$

The interpretation of the density function is

$$p_X(x)\Delta x \approx Prob\{X \in [x, x + \Delta dx]\}$$

This approximation becomes exact in the limit $\Delta x \to 0$. It is easy to verify that

$$p_X(x) \geq 0 \ \text{ and } \ \int_{-\infty}^{x} p_X(\lambda) d\lambda = P_X(x)$$

6 Expectation

Let X be a random variable. The *expected value* of X is defined as

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x p_{_X}(x) dx \tag{6}$$

and has the interpretation of the "average" value of the random variable X. It is also called the *mean* of X and is written $\overline{x}$ or m_x.
Let f be a real valued function and X be a random variable. Consider the random variable $Y = f(X)$. Analogous to (6) we can write

$$\mathbf{E}[Y] = \int_{-\infty}^{\infty} y p_{_Y}(y)\, dy$$

which can be shown to be (see #8)

$$\mathbf{E}[Y] = \mathbf{E}[f(X)] = \int_{-\infty}^{\infty} f(x) p_{_X}(x)\, dx \tag{7}$$

The m^{th} *moment* of X is $\mathbf{E}[X^m]$. Of particular importance are the 1st moment (or mean) and 2nd moment (or variance). It is easy to verify that

$$\mathbf{E}\left[(X - m_x)^2\right] = \mathbf{E}\left[X^2\right] - m_x^2$$

The expectation is a *linear* operator. If X and Y are two random variables

$$\mathbf{E}[\alpha X + \beta Y] = \alpha \mathbf{E}[X] + \beta \mathbf{E}[Y] \tag{8}$$

7 Examples

(a) Consider an experiment $\mathcal{E}$ that involves tossing a "fair" coin N times. The sample space consists of all vectors of length N, each component of which is 0 (heads) or 1 (tails) or $\Omega = \{0,1\}^N$. Since each outcome is equally likely, we have $Prob\{\omega\} = 2^{-N}$ for all $\omega \in \Omega$. Consider the random variable

$$X : \Omega \to \mathbb{R} : \omega \to \#\text{ of } 1's \text{ in } \omega$$

This random variable takes integer values $0, 1, \cdots, N$ and

$$Prob\{X = k\} = \frac{1}{2^N}\binom{n}{k}$$

(b) Consider an experiment $\mathcal{E}$ that involves measurement of some voltage known *a priori* to be in the sample space $\Omega = [0, 1)$. We declare the set of events $\mathcal{A}$ to consist of countable unions of semi-open intervals $[a, b)$ of the real line. With the function $Prob : \mathcal{A} \to \mathbb{R}$ defined by

$$Prob\{[a,b)\} = \begin{cases} 0 & \text{if } b \le 0 \text{ or } a \ge 1 \\ \min\{1,b\} - \max\{0,a\} & \text{otherwise} \end{cases}$$

the triple $(\Omega, \mathcal{A}, Prob)$ forms a probability space. Consider the random variable $X : \Omega \to \mathbb{R} : \omega \to \omega$. It is easy to verify that the density function of X is

$$p_{_X}(x) = \begin{cases} 1 & x \in [0,1) \\ 0 & \text{else} \end{cases}$$

We say X is uniformly distributed on $[0,1)$, written $X \sim U[0,1)$. Consider the new random variable $Y = X^2$. More precisely, $Y : \Omega \to \mathbb{R} : \omega \to \omega^2$. Observe that

$$\begin{aligned} P_{_Y}(y) &= Prob\{Y < y\} = Prob\{|X| < \sqrt{y}\} = P_{_X}(\sqrt{y}) \\ p_{_Y}(y) &= \frac{d}{dy} P_{_Y}(y) = p_{_X}(\sqrt{y}) \frac{1}{2\sqrt{y}} = \begin{cases} 1/2\sqrt{y} & y \in [0,1) \\ 0 & \text{else} \end{cases} \end{aligned}$$

(c) Let X be a Poisson distributed random variable with

$$Prob\{X = k\} = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \cdots$$

It can be verified that $\mathbf{E}[X] = \lambda$ and $\mathbf{E}\left[(X - \overline{x})^2\right] = \lambda$

(d) Let $\sigma > 0$ and consider the random variable X with density function

$$p_{_X}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(x-m)^2}{2\sigma^2}\right\}$$

It is straightforward to verify that $\mathbf{E}[X] = m$ and $\mathbf{E}\left[(X-m)^2\right] = \sigma^2$. We say that X is Gaussian or Normal with mean m and covariance σ^2 and this is written $X\mathbb{N}\left(m, \sigma^2\right)$. Gaussian random variables play a fundamental role in probability theory and practice.

8 Random Vectors

Let $(\Omega, \mathcal{A}, \mathcal{P})$ be a probability space and let $X_1, \cdots, X_n$ be random variables over this space, ie. $X_k : \Omega \to \mathbb{R}$ for $k = 1, \cdots, n$
Define the *random vector* $X = [X_1 \cdots X_n]'$ and let $x = [x_1 \cdots x_n]' \in \mathbb{R}^n$. The *joint distribution function* of X is

$$P_{_X}(x) = \text{Prob}\{X_1, < x_1 \ \& \ X_2 < x_2 \ \& \cdots \& \ X_n < x_n\}$$

and the *joint density function* of X is

$$p_{_X}(x) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} P_{_X}(x) \tag{9}$$

These functions have the usual interpretations and we can introduce the notion of expectation in the obvious manner. In particular, if X is a random n vector and Y is a random m vector, and $Z = [X \ Y]'$ then

$$P_{_Y}(y) = \int_{-\infty}^{\infty} P_{_{XY}}(x, y) dx \tag{10}$$

The *covariance matrix of* X and the *cross-covariance matrix of* X *and* Y are defined respectively as

$$\begin{aligned}\Lambda_{XX} &= \mathbf{E}\left[(X-\overline{x})(X-\overline{x})'\right] \in \mathbb{R}^{N\times N} \\ \Lambda_{XY} &= \mathbf{E}\left[(X-\overline{x})(Y-\overline{y})'\right] \in \mathbb{R}^{N\times N}\end{aligned} \tag{11}$$

Theorem 8.1 *Let* Z *be a random vector. Then* $\Lambda_{ZZ} \geq 0$.

Proof: It is immediate that $\Lambda_{ZZ} = \Lambda'_{ZZ}$. Next, fix $\xi \in \mathbb{R}^N$ and observe that

$$\xi'\Lambda_{ZZ}\xi = \mathbf{E}\left[\xi'(X-\overline{x})(X-\overline{x})'\xi\right] = \mathbf{E}\left[Q^2\right] \geq 0$$

where the random *variable* $Q = (X-\overline{x})'\xi$, proving the claim. □

Example 8.2 Define the random variables $X \sim \cup[0,1], Y = X^2$ and consider the random vector $Z = [X\ Y]'$. It is a straightforward exercise to verify that $P_Z(z) = P_{XY}(x,y) = P_X(\min(x,\sqrt{y}))$ and

$$\Lambda_{ZZ} = \begin{bmatrix} 1/3 & 1/4 \\ 1/4 & 1/5 \end{bmatrix}$$

Theorem 8.3 (Cauchy-Schwartz Inequality) *For any random variables* X, Y

$$\mathbf{E}^2[XY] \leq \mathbf{E}\left[X^2\right]\mathbf{E}\left[Y^2\right]$$

Proof: Define $Z = [X\ Y]'$ and note that by an argument identical to previous theorem.

$$M = \mathbf{E}\left[ZZ'\right] = \begin{bmatrix} \mathbf{E}\left[X^2\right] & \mathbf{E}[XY] \\ \mathbf{E}[XY] & \mathbf{E}\left[Y^2\right] \end{bmatrix} \geq 0$$

This implies that $det\ M = \mathbf{E}\left[X^2\right]\mathbf{E}\left[Y^2\right] - \mathbf{E}^2[XY] \geq 0$, proving the claim. □
The random variables X, Y are called *uncorrelated* if

$$\Lambda_{XY} = \mathbf{E}\left[(X-\overline{x})(Y-\overline{y})'\right] = 0 \tag{12}$$

and the collection of random variables $\{X_1, \cdots, X_N\}$ is called *uncorrelated* if the collection is pairwise uncorrelated, i.e.

$$\Lambda_{X_iY_j} = 0 \text{ for } i \neq j$$

or, equivalently, if Λ_{XX} is *diagonal*. These definitions extend *verbatim* to random *vectors*. Note that the collection $X = \{X_1, \cdots, X_N\}$ of random vectors is uncorrelated if and only if Λ_{XX} is *block-diagonal*.

9 Functions of a Random Variable

Let X be a random variable over the probability space $(\Omega, \mathcal{A},\ \text{Prob})$ and let $f : \mathbb{R} \to \mathbb{R}$. Then, $Y = f(X)$ is also a random variable over the same probability space $(\Omega, \mathcal{A},\ \text{Prob})$.
Strictly speaking this is not always true. In order for Y to qualify as a random variable, we require that for all $y \in \mathbb{R}$, the sets $S_y = \{\omega : Y(\omega) < y\}$ must be events, i.e. $S_y \in \mathcal{A}$. This is necessary so that we may assign probabilities to S_y and thereby define $P_y(y)$. With this caveat, we can write

$$P_{_Y}(y) = Prob\{Y < y\} = Prob\{f(X) < y\} = \int_S p_{_X}(x)dx \tag{13}$$

where $S = \{x : f(x) < y\}$.

Example 9.1 Consider the random variable $X\mathbb{N}(0, 1)$ and let $Y = X^2$. We will compute the density function $p_{_Y}(y)$. First observe that

$$P_{_Y}(y) = Prob\{X^2 < y\} = Prob\{-\sqrt{y} \le X \le \sqrt{y}\} = P_{_X}(\sqrt{y}) - P_{_X}(-\sqrt{y})$$

Differentiating this expression yields $p_{_Y}(y) = 1/\sqrt{2\pi y}\exp\{-y^2/2\}$.

Now suppose in the neighborhood of some point y° the function f is invertible with a differentiable inverse. More precisely, there exists an open neighborhood $\mathcal{N}$ containing y° on which the function $g(y) = x$ is well-defined, continuous, and differentiable where $y = f(x)$. Next observe that

$$\begin{aligned} p_{_Y}(y^\circ)\Delta y &\approx Prob\{Y \in [y^\circ, y^\circ + \Delta y]\} = Prob\{X \in [g(y^\circ), g(y^\circ + \Delta y)]\} \\ &\approx Prob\{X \in [g(y^\circ), g(y^\circ) + \frac{dg}{dy}(y^\circ)\Delta y]\} \\ &\approx p_{_X}(g(y^\circ))\left|\frac{dg}{dy}(y^\circ)\right|\Delta y \end{aligned}$$

Taking the limit $\Delta y \to 0$, we obtain

$$p_{_Y}(y^\circ) = p_{_X}(g(y^\circ))\left|\frac{dg}{dy}(y^\circ)\right| \tag{14}$$

Example 9.2 Consider the random variable $X\mathbb{N}(0, 1)$ and let $Y = f(X) = \exp(X)$. Observe that $f(x) = \exp(x)$ is invertible for all $y > 0$ and its inverse is $g(y) = \ln(y)$. Next, for $y^\circ > 0$, we have

$$p_{_Y}(y^\circ) = p_{_X}(g(y^\circ))\left|\frac{dg}{dy}(y^\circ)\right| = p_{_X}(\ln(y^\circ))/y^\circ = \frac{1}{y^\circ\sqrt{2\pi}}\exp\{-(\ln y^\circ)^2/2\}$$

Also, since the random variable Y is always positive, $p_{_Y}(y) = 0$ for $y \le 0$.

In order to generalize this development to random vectors, we shall need the following simple Lemma:

Lemma 9.3 *Let $T \subseteq \mathbb{R}^N, A \in \mathbb{R}^{N\times N}$ and define $S = \{Ax : x \in T\} \subseteq \mathbb{R}^N$. Then, vol(S) = $|detA|vol(T)$, where $vol(S) = \int_{x\in S} dx$.* □

Now let X be a random N-vector and let $Y = f(X)$ where $f : \mathbb{R}^N \to \mathbb{R}^N$ is some function. Suppose $g = f^{-1}$ exists on some neighborhood $\mathcal{N}$ of $y^\circ \in \mathbb{R}^N$ and suppose $g : \mathcal{N} \to \mathbb{R}^N$ has continuous first partial derivatives. Let us write

$$g(y) = g(y_1, y_2, \cdots, y_N) = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix}' = \begin{bmatrix} g_1(y) & g_2(y) & \cdots & g_N(y) \end{bmatrix}'$$

The *Jacobian* of g at y° is the $N \times N$ matrix

$$J_g(y^\circ) = \left[\begin{array}{ccc} \frac{\partial g_1}{\partial y_1} & \cdots & \frac{\partial g_1}{\partial y_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_N}{\partial y_1} & \cdots & \frac{\partial g_N}{\partial y_N} \end{array} \right] \Bigg|_{y^\circ} \tag{15}$$

Using the Taylor expansion for vector-valued functions, we can write

$$g(y^\circ + \Delta y) = g(y^\circ) + J_g(y^\circ)\Delta y + \text{higher order terms}$$

We can approximate g by neglecting the higher order terms, and this approximation is good for small $\|\Delta y\|$. We can now compute

$$\begin{aligned} p_{_Y}(y^\circ)\Delta y_1 \cdots \Delta y_N &\approx \text{Prob}\{Y \in \{y^\circ + \Delta y : \ \Delta y \in \text{hypercube } T \text{ with sides } \Delta y_1, \cdots, \Delta y_N\}\} \\ &= \text{Prob}\{X \in \{g(y^\circ + \Delta y) : \ \Delta y \in T\}\} \end{aligned}$$

When $\|\Delta y\|$ is small, we can use the Taylor approximation to write

$$\{g(y^\circ + \Delta y) : \ \Delta y \in T\} \approx \{g(y^\circ) + J_g(y^\circ)\Delta y : \ \Delta y \in T\}$$

Also note that $p_{_X}(x) \approx p_{_X}(g(y^\circ)) =$ constant on this set. Continuing, using Lemma (9.3) we get

$$\begin{aligned} p_{_Y}(y^\circ)\Delta y_1 \cdots \Delta y_N &\approx \text{Prob}\{X \in \{g(y^\circ) + J_g(y^\circ)\Delta y : \ \Delta y \in T\}\} \\ &\approx p_{_X}(g(y^\circ))\text{vol}\ (J_g(y^\circ)(T)) \\ &= p_{_X}(g(y^\circ))|\text{det}J_g(y^\circ)|\text{vol}(T) \\ &= p_{_X}(g(y^\circ))|\text{det}J_g(y^\circ)|\Delta y_1 \cdots \Delta y_N \end{aligned}$$

Finally, taking the limit $\|\Delta y\| \to 0$ we get

$$p_{_Y}(y^\circ) = p_{_X}(g(y^\circ))|\text{det}J_g(y^\circ)| \tag{16}$$

Example 9.4 Let X be a random N-vector with density $p_{_X}(x)$. Define the random N-vector $Y = AX$ where $A \in \mathbb{R}^{N\times N}$ is invertible. We can then compute the density for Y as $p_{_Y}(y) = p_{_X}(A^{-1}y)|\text{det } A|$.

Exercise 9.5 Let X be uniformly distributed on $[0, 1]$. Find a function $f : \mathbb{R} \to \mathbb{R}$ so that the new random variable $Y = f(X)\mathbb{N}\left(O, \sigma^2\right)$.

Exercise 9.6 Let X, Y be random variables with the joint density

$$p_{XY}(x,y) = \frac{1}{2\pi}\exp\{-(x^2+y^2)/2\}$$

Define the new random variables $R > 0$ and $\Theta \in [0, 2\pi)$ by

$$X = R\cos\Theta \quad , \quad Y = R\sin\Theta$$

Compute the joint density $p_{R\Theta}(r,\theta)$, and the densities $p_R(r), p_\Theta(\theta)$.

10 Statistical Independence

A collection of random variables $\{X_1, X_2, \cdots, X_N\}$ is called *statistically independent* if

$$P_X(x) = \Pi_{k=1}^N P_{X_k}(x_k) \quad \text{or equivalently,} \quad p_X(x) = \Pi_{k=1}^N p_{X_k}(x_k) \tag{17}$$

Lemma 10.1 *If $\{X_1, \cdots, X_N\}$ are independent then $\{f_1(X_1), \cdots, f_N(X_N)\}$ are also independent. Moreover*

$$\mathbf{E}\left[\Pi_1^N f(X_k)\right] = \Pi_1^N \mathbf{E}\left[f(X_k)\right]$$

Lemma 10.2 *If a collection of random variables $\{X_1, \cdots, X_N\}$ are independent then they are uncorrelated.*

Proof: Using Lemma (10.1), we have for $i \neq j$,

$$\Lambda_{X_i X_j} = \mathbf{E}\left[(X_i - \overline{x}_i)(X_j - \overline{x}_j)'\right] = \mathbf{E}\left[X_i - \overline{x}_i\right]\mathbf{E}\left[X_j - \overline{x}_j\right]' = 0$$

The converse of this lemma is *not true* true in general. □

Exercise 10.3 Is statistical independence transitive, i.e. is it true that X, Y are independent and Y, Z are independent implies that X, Z are also ? Give proof or counterexample.

11 Moments and Characteristic Functions

Let X be a random variable. Recall that the n^{th} moment of X is $\mathbf{E}\left[X^n\right]$.
For the random variable X, define its *characteristic function* or *moment generating function* by

$$\Phi_X(\omega) \stackrel{\triangle}{=} \mathbf{E}\left[e^{j\omega X}\right] = \int_{x=-\infty}^{\infty} e^{j\omega x} P_X(x)dx \tag{18}$$

Observe that the moment generating function $\Phi_X(\omega)$ is simply the Fourier Transform of the density function $p_X(x)$. Using the Inverse Fourier Transform we can write

$$p_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi_X(\omega)d\omega \tag{19}$$

The moments of X can be readily found from $\Phi_X(\omega)$ as follows.

Theorem 11.1 *Let X be a random variable with moment generating function $\Phi_X(\omega)$. Then, $\mathbf{E}[X^n] = j^{-n}\Phi_X^{(n)}(0)$, where $^{(n)}$ denotes the n^{th} derivative with respect to ω.*

Proof: We make the following calculation

$$\begin{aligned}\mathbf{E}[X^n] &= \int_{-\infty}^{\infty} x^n p_X(x)dx = \int_{-\infty}^{\infty} x^n e^{j\omega x} p_X(x)dx\Big|_{\omega=0} \\ &= \frac{1}{j^n}\int_{-\infty}^{\infty} \frac{d^n}{d\omega^n} e^{j\omega x} p_X(x)dx\Big|_{\omega=0} = \frac{1}{j^n}\frac{d^n}{d\omega^n}\left(\int_{-\infty}^{\infty} e^{j\omega x} p_X(x)dx\right)_{\omega=0} \\ &= \frac{1}{j^n}\frac{d^n}{d\omega^n}\Phi_X(\omega)|_{\omega=0}\end{aligned}$$

Example 11.2 Let $X\mathbb{N}\left(m,\sigma^2\right)$. The moment generating function of X is

$$\Phi_X(\omega) = \int_{\infty}^{-\infty} e^{j\omega x}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-m)^2/2\sigma^2}dx = \exp\{j\omega m - \omega^2\sigma^2/2\}$$

Using this we can compute $\mathbf{E}\left[X^4\right] = j^{-4}\Phi_X^{(4)}|_{\omega=0} = 3\sigma^4 + \sigma^2 m^2 + m^4$.

Moment generating functions can be extended to random *vectors* in the following fashion. Let $X = [X_1, \cdots, X_n]'$ be a random vector. Define the moment generating function of X as

$$\Phi_X(\omega) = \mathbf{E}\left[e^{j\omega' X}\right] = \int_{x_1=-\infty}^{\infty}\cdots\int_{x_n=-\infty}^{\infty} e^{j(\omega_1 x_1+\cdots+\omega_n x_n)}p_X(x)dx_1\cdots dx_n \tag{20}$$

where $\omega = [\omega_1 \cdots \omega_n]'$.
Moment generating functions are particularly useful when working with sums of random variables. We will see this later in the context of the Central Limit Theorem (section # 21). For now, observe that if X and Y are *independent* random variables and $S = X + Y$, then

$$\Phi_S(\omega) = \mathbf{E}\left[e^{j\omega S}\right] = \mathbf{E}\left[e^{j\omega X}e^{j\omega Y}\right] = \mathbf{E}\left[e^{j\omega X}\right]\mathbf{E}\left[e^{j\omega Y}\right] = \Phi_X(\omega)\Phi_Y(\omega) \tag{21}$$

Here we have used Lemma (10.1). Since Fourier domain multiplication is "time" domain convolution we have

$$p_S = p_X * p_Y$$

12 Conditioning

Let X_1, X_2 be random variables, and A_1, A_2 be subsets of $\mathbb{R}$. Using Bayes' rule, we can write

$$Prob\{X_1 \in A_1 | X_2 \in A_2\} = \frac{Prob\{X_1 \in A_1 \ \&\ X_2 \in A_2\}}{Prob\{X_2 \in A_2\}} = \frac{\int_{A_1\times A_2} p_{X_1X_2}(x_1x_2)dx_1dx_2}{\int_{A_2} p_{X_2}(x_2)dx_2}$$

Let $A_2 = [x_2^0, x_2^0 + \Delta x_2]$. For small $\Delta x_2, p_{X_1,X_2}(x_1,x_2) \approx p_{X_1,X_2}(x_1,x_2^0)$ on the rectangle $A_1 \times A_2$ and $p_{X_2}(x_2) \approx p_{X_2}(x_2^0)$ on A_2. Then,

$$\begin{aligned}Prob\{X_1 \in A_1 | X_2 \in A_2\} &\approx \frac{\int_{A_1\times A_2} p_{X_1X_2}(x_1,x_2^0)dx_1dx_2}{\int_{A_2} p_{X_2}(x_2^0)dx_2} \\ &= \frac{\int_{A_1} p_{X_1X_2}(x_1,x_2^0)dx_1\Delta x_2}{p_{X_2}(x_2^0)\Delta x_2} = \int_{A_1}\frac{p_{X_1X_2}(x_1,x_2^0)}{p_{X_2}(x_2^0)}dx_1\end{aligned}$$

In the limit $\Delta x_2 \to 0$ we obtain

$$Prob\{X_1 \in A_1 | X_2 = x_2^0\} = \int_{A_1} p_{X_1|X_2}(x_1, x_2^0) dx_1$$

where

$$p_{X_1|X_2=x_2^0}(x_1, x_2^0) = \frac{p_{X_1X_2}(x_1, x_2^0)}{p_{X_2}(x_2^0)} \tag{22}$$

The function $p_{X_1|X_2=x_2^0}$ is called the *conditional density of* X_1 *given* $X_2 = x_2^0$. The interpretation of this density is

$$p_{X_1|X_2=x_2^0}(x_1, x_2^0)\Delta x_1 \approx Prob\{X_1 \in [x_1, x_1 + \Delta x_1] | X_2 = x_2^0\}$$

The extension of the above discussion to random vectors is *verbatim*.

Example 12.1 Let X, Y be random variables with joint density $p_{XY}(x, y) = \exp\{-(x^2 + y^2)/2\}/2\pi$. Define new random variables R, Θ by $X = R\cos(\Theta), Y = R\sin(\theta)$. Recall that (see Exercise 9.6)

$$p_{R\Theta}(r, \theta) = \begin{cases} r\exp\{-r^2/2\}/2\pi & \text{for } \ r > 0, \theta \in [0, 2\pi) \\ 0 & \text{otherwise} \end{cases}$$

We will find the conditional density $p_{X|R}(x, r)$. For this we first need the joint density $p_{XR}(x, r)$. Observe that

$$\begin{bmatrix} X \\ R \end{bmatrix} = \begin{bmatrix} X \\ \sqrt{X^2 + Y^2} \end{bmatrix} = f\left(\begin{bmatrix} X \\ Y \end{bmatrix}\right)$$

The function f is invertible for all X and $R > 0$. We can write its inverse as

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} X \\ \sqrt{R^2 - X^2} \end{bmatrix} = g\left(\begin{bmatrix} X \\ R \end{bmatrix}\right)$$

The Jacobian of g is found to be

$$J_g(x, r) = \begin{bmatrix} \frac{\partial X}{\partial X} & \frac{\partial X}{\partial R} \\ \frac{\partial \sqrt{R^2 - X^2}}{\partial X} & \frac{\partial \sqrt{R^2 - X^2}}{\partial R} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ * & R/\sqrt{R^2 - X^2} \end{bmatrix}$$

We can now calculate using (15)

$$p_{XR}(x, r) = p_{XY}(x, \sqrt{r^2 - x^2})|\det\ J_g(x, r)| = \frac{1}{2\pi}\exp\{-r^2/2\} r (r^2 - x^2)^{-1/2}$$

Finally, the conditional density may be found as

$$p_{X|R}(x, r) = \frac{p_{XR}(x, r)}{p_R(r)} = \frac{r}{\sqrt{r^2 - x^2}}$$

13 Conditional Expectation

Let X be a random variable, and suppose the constant c serves as an "estimate" of X. Let us choose c to be the *minimum variance estimator*, i.e. c is chosen to minimize

$$\mathbf{E}\left[(X-c)^2\right] = E[(X - m_X + m_X - c)^2] = \sigma_X^2 + (m_X - c)^2$$

It is clear from the above expression that the optimal choice of c is $c = m_X$. Equivalently, the mean of X is the minimum variance estimate of X.
Now let X, Y be random variables over the same probability space. Observe that

$$\mathbf{E}\left[X|Y=y\right] = \int_{-\infty}^{\infty} x p_{X|Y}(x,y)dx = f(y) = \text{some function of } y \text{ alone}$$

Lemma 13.1 *With $f(y)$ defined as above, for any function $g(y)$*

$$\mathbf{E}\left[g(Y)f(Y)\right] = \mathbf{E}\left[g(Y)X\right]$$

Proof: A simple calculation reveals that

$$\begin{aligned}
\mathbf{E}\left[g(Y)f(Y)\right] &= \int_{y=-\infty}^{\infty} g(y)f(y)p_Y(y)dy \\
&= \int_{y=-\infty}^{\infty} g(y) \int_{x=-\infty}^{\infty} x p_{X|Y}(x|y)dx p_Y(y)dy \\
&= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} g(y) x p_{XY}(x,y)dxdy = \mathbf{E}\left[g(Y)X\right]
\end{aligned}$$

proving the claim. □

Theorem 13.2 *The conditional mean $f(y) = \mathbf{E}\left[X|Y=y\right]$ is the minimum variance estimate of X given $Y = y$, i.e. for any other estimator $g(y)$,*

$$\mathbf{E}\left[(X-g(Y))^2\right] \geq \mathbf{E}\left[(X-f(Y))^2\right]$$

Proof: Observe that

$$\begin{aligned}
\mathbf{E}\left[(X-g(Y))^2\right] &= \mathbf{E}\left[(X - f(Y) + f(Y) - g(Y))^2\right] \\
&= \mathbf{E}\left[(X-f(Y))^2\right] + \mathbf{E}\left[(f(Y)-g(Y))^2\right] + 2\mathbf{E}\left[(X-f(Y))(f(Y)-g(Y))\right] \\
&\geq \mathbf{E}\left[(X-f(Y))^2\right]
\end{aligned}$$

where the final inequality above follows from Lemma 13.1. □
The above theorem asserts that the conditional mean is the *minimum variance estimator*, and is optimal over *all* (even nonlinear) estimators.

Lemma 13.3 *(a) If X, Y are independent $\mathbf{E}\left[X|Y=y\right] = \mathbf{E}\left[X\right]$*

(b) $\mathbf{E}\left[g(Y)X|Y=y\right] = g(y)\mathbf{E}\left[X|Y=y\right]$

Example 13.4 The above lemma is useful for quickly computing conditional expectations. To illustrate this, let X, Y, R, Θ be as in Example 12.1. Then,

$$\begin{aligned}\mathbf{E}[X|R=r^\circ] &= \mathbf{E}[R\cos\Theta|R=r^\circ] \\ &= r^\circ \mathbf{E}[\cos\Theta|R=r^\circ] \\ &= r^\circ \mathbf{E}[\cos\Theta] = 0\end{aligned}$$

Here we have used the fact that R, Θ are independent.

Exercise 13.5 In the example above, compute $\mathbf{E}[X_1|\Theta=\theta^\circ]$.

Exercise 13.6 Let $X_1, \cdots, X_n$ be independent identically distributed random variables.

(a) Compute $\mathbf{E}[X_1|X_1+X_2+\cdots X_n=s]$.

(b) Compute for $k \leq n$,

$$\mathbf{E}\left[\frac{X_1+\cdots+X_k}{X_1+\cdots+X_n}\right].$$

Exercise 13.7 Let Y, X_1, X_2 be (not necessarily independent) random variables. Find the best linear estimator of Y given X_1 and X_2, i.e. find the linear estimator whose error variance is minimal.

14 Gaussian Random Vectors

Recall that a random variable X is called *Gaussian* if

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\} \tag{23}$$

This is written $X \sim N(M, \sigma^2)$. A Gaussian random variable is *completely* specified by the parameters m, σ^2. It happens that these parameters are the mean and covariance respectively, i.e.

$$\mathbf{E}[X] = m, \mathbf{E}\left[(X-m)^2\right] = \sigma^2$$

Thus the 2nd order statistics of a Gaussian random variable completely specify the random variable. A random *vector* $X = [X_1 \cdots X_n]'$ is called Gaussian if there exists a nonsingular matrix Λ_{XX} such that the joint density function $p_X(x)$ takes the form

$$p_X(x) = \frac{1}{[(2\pi)^n \det\Lambda_{XX}]^{\frac{1}{2}}} \exp\left\{-\frac{(x-m)'\Lambda_{XX}^{-1}(x-m)}{2}\right\} \tag{24}$$

So X is completely specified by the parameters $m \in \mathbb{R}^N, \Lambda_{XX} \in \mathbb{R}^{N\times N}$ (invertible). It happens that

$$m = \mathbf{E}[X], \Lambda_{XX} = \mathbf{E}\left[(X-m)(X-m)'\right] > 0$$

Theorem 14.1 *Let $X \sim N(m_x, \Lambda_{XX})$ and let $Y = AX + b$ where $A \epsilon \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$. Suppose A has full row rank, i.e. rank $(A) = m$. Then,*

$$Y \sim N(m_y, \Lambda_{YY}) \quad \textit{where} \quad m_y = Am_x + b,\ \Lambda_{YY} = A\Lambda_{XX}A' \tag{25}$$

Sketch of Proof: First conduct tedious algebraic manipulation to show that $p_Y(y)$ has the appropriate form for it to qualify as Gaussian. Then

$$\begin{aligned} \mathbf{E}[Y] &= m_y = Am_x + b \\ \mathbf{E}\left[(Y - m_y)(Y - m_y)'\right] &= \mathbf{E}\left[A(X - m_x)(X - m_x)'A'\right] = A\Lambda_{XX}A' \end{aligned}$$

Lemma 14.2 *Let X be a Gaussian random vector. Then, X is uncorrelated $\Rightarrow$ X is independent.*

Proof: Since X is uncorrelated Λ_{XX} is diagonal. So we can write $\Lambda_{XX} = \text{diag}\{\sigma_1^2, \cdots, \sigma_n^2\}$. Then, the joint density p_X factors into individual densities as

$$p_X(x) = \Pi_{k=1}^N \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\{-(x_k - m_k)^2/2\sigma_k^2\}$$

proving the claim. □

Theorem 14.3 *Let X be a random n-vector, Y be a random m-vector, and let*

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix} \mathbb{N}(m_z, \Lambda_{ZZ}) = N\left(\begin{bmatrix} m_x \\ m_y \end{bmatrix}, \begin{bmatrix} \Lambda_{XX} & \Lambda_{XY} \\ \Lambda_{YX} & \Lambda_{YY} \end{bmatrix}\right)$$

Then, $X|Y = y^0 \mathbb{N}\left(m_{x|y^0}, \Lambda_{XX|Y=y^0}\right)$ where

$$m_{x|y^0} = m_x + \Lambda_{XY}\Lambda_{YY}^{-1}\left(y^0 - m_y\right) \tag{26}$$

$$\Lambda_{XX|Y=y^0} = \Lambda_{XX} - \Lambda_{XY}\Lambda_{YY}^{-1}\Lambda_{XY}' \tag{27}$$

Proof: First note that

$$\begin{bmatrix} A & D \\ C & B \end{bmatrix}^{-1} = \begin{bmatrix} \Delta^{-1} & -\Delta^{-1}F \\ -E\Delta^{-1} & B^{-1} + E\Delta^{-1}F \end{bmatrix}$$

where $\Delta = A - DB^{-1}C$ (the Schur complement), $E = B^{-1}C, F = DB^{-1}$, and provided all inverses above exist. Applying this to Λ_{ZZ} yields

$$\Lambda_{ZZ}^{-1} = \begin{bmatrix} \Lambda_{XX} & \Lambda_{XY} \\ \Lambda_{YX} & \Lambda_{YY} \end{bmatrix}^{-1} = \begin{bmatrix} \Delta^{-1} & -\Delta^{-1}F \\ -F'\Delta^{-1} & \Lambda_{XX}^{-1} + F'\Delta^{-1}F \end{bmatrix}$$

where $F = \Lambda_{XY}\Lambda_{YY}^{-1}$ and $\Delta = \Lambda_{XX} - \Lambda_{XY}\Lambda_{YY}^{-1}\Lambda_{XY}$. Next define $v = x - m_x$ and $w = y^0 - m_y$, and examine the quantity

$$\begin{aligned} Q &= \begin{bmatrix} v' & w' \end{bmatrix} \Lambda_{ZZ}^{-1} \begin{bmatrix} v \\ w \end{bmatrix} - w'\Lambda_{XX}^{-1} w \\ &= \begin{bmatrix} v' & w' \end{bmatrix} \begin{bmatrix} \Delta^{-1} & -\Delta^{-1}F \\ -F'\Delta^{-1} & F'\Delta^{-1}F \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} \\ &= (v - Fw)'\Delta^{-1}(v - Fw) \end{aligned}$$

We can now compute the conditional density

$$p_{X|Y=y^0} = \frac{p_{XY}(x, y^0)}{p_Y(y^0)} = c\exp\{-Q/2\} = c\exp\{-(v - Fw)'\Delta^{-1}(v - Fw)/2\}$$

where c is some constant. This density has the requisite form for $X|Y = y^0$ to be Gaussian. Also, we obtain the mean and variance of $X|Y = y^0$ simple by examining this density function, and these expressions are found to be in agreement with the theorem statement. □

Remark 14.4 The conditional mean $m_{X|Y=y^0}$ is to be regarded as the "best" (minimum error variance) estimate of X given that $Y = y^0$, and $\Lambda_{XX|Y=y^0}$ is the corresponding minimal error variance.
Observe that if X and Y are uncorrelated, $\Lambda_{XY} = 0$ so $m_{X|Y=y^0} = m_X$. We should expect this because for Gaussian variables, uncorelated implies independent. Thus X and Y are independent and the knowledge that $Y = y^0$ gives us no information about X. As a consequence the best estimate of X given $Y = y^0$ is its (unconditional) mean m_X.
Also observe that $\Lambda_{XX} \geq \Lambda_{XX|Y=y^0}$ (as a matrix inequality). Thus the variance of $X|Y = y^0$ is less than that of X. This is consistent with our intuition that the information $Y = y^0$ reduces the "uncertainty" in X.

Exercise 14.5 Let $X\mathbb{N}(m, \Lambda)$ where

$$m = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad , \quad \Lambda = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix}$$

(a) Find a matrix Q such that $X = QZ$ with Z being independent, zero-mean, unit-variance, Gaussian random variables.

(b) Find $\mathbf{E}[X_1|X_2, X_3]$

(c) Find Prob $\{X_1 \geq 0$ and $X_2 \geq 0\}$

15 Inequalities

Theorem 15.1 (Chebychev) *Let X be a random variable. Then, for any $r > 0$ and $\epsilon > 0$,*

$$Prob\{|X| \geq \epsilon\} \leq \frac{\mathbf{E}[|X|^r]}{\epsilon^r} \tag{28}$$

Of particular importance is the case $r = 2$ when

$$Prob\{|X| \geq \epsilon\} \leq \frac{\mathbf{E}\left[X^2\right]}{\epsilon^2} \tag{29}$$

Proof: The result follows from the following sequence of inequalities

$$\begin{aligned}
\mathbf{E}\left[|X|^r\right] &= \int_{-\infty}^{\infty} |x|^r p_{_X}(x)dx \geq \int_{|x|\geq\epsilon} |x|^r p_{_X}(x)dx \\
&\geq \epsilon^r \int_{|x|\geq\epsilon} p_{_X}(x)dx = \epsilon^r Prob\{|X| \geq \epsilon\}
\end{aligned}$$

Theorem 15.2 (Chebychev) *Let X be a random variable with mean m and covariance σ^2 For $\epsilon > 0$,*

$$Prob\{X - m \geq \epsilon\} \leq \frac{\sigma^2}{\sigma^2 + \epsilon^2} \tag{30}$$

Proof: Without loss of generality, assume $m_{_X} = 0$. Define the new random variable

$$I_\epsilon(X) = \begin{cases} 1 & \text{if} \quad X \leq \epsilon \\ 0 & \text{if} \quad X > \epsilon \end{cases}$$

Notice that

$$\begin{aligned}
\int_{-\infty}^{\infty} (\epsilon - x) I_\epsilon(x) p_{_X}(x)dx &= \int_{-\infty}^{\epsilon} (\epsilon - x) p_{_X}(x)dx \\
&= \int_{-\infty}^{\infty} (\epsilon - x) p_{_X}(x)dx - \int_{\epsilon}^{\infty} (\epsilon - x) p_{_X}(x)dx \\
&= \epsilon + \int_{\epsilon}^{\infty} (x - \epsilon) p_{_X}(x)dx \geq \epsilon
\end{aligned}$$

Squaring both sides of the final inequality above (note that $\epsilon > 0$) and using the Cauchy-Schwartz inequality yields

$$\begin{aligned}
\epsilon^2 &\leq \mathbf{E}^2\left[(\epsilon - X) I_\epsilon(X)\right] \\
&\leq \mathbf{E}\left[(\epsilon - X)^2\right] \mathbf{E}\left[I_\epsilon^2(X)\right] \\
&= \mathbf{E}\left[(\epsilon - X)^2\right] \mathbf{E}\left[I_\epsilon(X)\right] \\
&= (\sigma^2 + \epsilon^2) \mathbf{E}\left[I_\epsilon(X)\right]
\end{aligned}$$

Finally observe that

$$\begin{aligned}
Prob\{X \geq \epsilon\} &= \int_{\epsilon}^{\infty} p_{_X}(x)dx = 1 - \int_{-\infty}^{\epsilon} p_{_X}(x)dx \\
&= 1 - \mathbf{E}\left[I_\epsilon(X)\right] \leq 1 - \frac{\epsilon^2}{\sigma^2 + \epsilon^2} = \frac{\sigma^2}{\sigma^2 + \epsilon^2}
\end{aligned}$$

proving the claim. □

Theorem 15.3 (Hoeffding) *Let* $(X_1, \cdots, X_N)$ *be* independent, bounded *random variables with* $X_k \in [0,1]$. *Define* $S = (X_1 + \cdots + X_N)/N$. *Then,*

$$Prob\{S - m_{_S} \geq \epsilon\} \leq \exp\{-2N\epsilon^2\} \tag{31}$$

These inequalities are very useful to prove convergence in probability or in quadratic mean of a sequence of random variables (see # 17). Another useful inequality is

Theorem 15.4 (Jensen's Inequality) *Let* X *be a random vector confined to some* convex *set* $\Omega \subseteq \mathbb{R}^n$ *and let* $f : \Omega \to \mathbb{R}$ *be a* convex *function. Then*

$$E[f(X)] \geq f(E[X]) \tag{32}$$

Proof: A property that convex functions satisfy is that

$$f(\sum_1^N \alpha_i X_i) \leq \sum_1^N \alpha_i f(X_i)$$

where $\alpha_i \in [0,1]$ and $\sum_1^N \alpha_i = 1$. The infinitesimal version of this property is

$$f\left(\int_\Omega x w(x) dx\right) \leq \int_\Omega f(x) w(x) dx$$

for functions $w : \Omega \to \mathbb{R}$ such that $w(x) \geq 0$ and $\int_\Omega w(x)dx = 1$. The result follows on choosing $w(x) = p_{_X}(x)$. □

Exercise 15.5 In this problem we consider a simple randomized algorithm for volume computation called Rejection sampling.
Let $S \subseteq T \subseteq \mathbb{R}^n$. We wish to compute the relative volume of S in T, i.e. we want to find $v = vol(S)/vol(T)$. We do this as follows:
We generate L "darts" uniformly distributed on T and count the number of darts M which also lie in S. Then we declare that

$$\hat{v} = \frac{M}{L}$$

approximates the relative volume v. Determine upper and lower bounds on the number of darts L that must be thrown in order that

$$Prob\left\{|\hat{v} - v| \leq \epsilon v\right\} \geq 1 - \delta$$

Note that this method of volume computation is a *randomized* algorithms in that the output of the algorithm $\hat{v}$ is a random variable. The parameter ϵ is called the accuracy, and $1 - \delta$ is called the confidence and we are *probably, approximately* computing the volume v.

16 Random Sequences

A *random sequence* is a collection of random variables $X = \{X_1, X_2, \cdots\} = \{X_n\}_{n=1}^{\infty}$ defined over the *same* probability space $(\Omega, \mathcal{A}, \text{Prob})$. We specify the statistics of X by providing the countable collection of joint densities

$$p_X^N \stackrel{\triangle}{=} p_{X_1 \cdots X_N}(x_1, \cdots, x_N) \quad N = 1, 2, 3, \cdots$$

This is an extraordinary amount of redundant information. In most interesting cases we can have a more succinct statistical description.

Example 16.1 *IID random sequences.* Let $X = \{X_n\}_1^{\infty}$ be a random sequence with each X_k being *identically distributed* and the collection X being statistically *independent*. Then, the statistics of X are completely specified by the the density function $p_{X_k}(\cdot) = f(\cdot)$ as

$$p_X^N(x_1, \cdots, x_N) = \Pi_{k=1}^N \ f(x_k)$$

Example 16.2 Random sequences can be constructed easily from IID sequences. For example, consider the random sequence $Z = \{Z_n\}_1^{\infty}$ where

$$Z_1 = X_1, \quad Z_k = X_k - X_{k-1} \quad \text{for } k \geq 2$$

and $X = \{X_n\}_1^{\infty}$ is IID with $f(\cdot) = p_{X_K}(\cdot)$. We can verify that

$$p_Z^N = \Pi_{k=1}^N \ f(z_k - z_{k-1}) \quad \text{with } z_0 \equiv 0.$$

17 Convergence Notions

Let $\{r_n\}_1^{\infty}$ be a sequence of real numbers. We say that $\{r_n\}$ *converges* to r if for any $\epsilon > 0$, there exists N_ϵ such that

$$|r_n - r| \leq \epsilon \text{ for } n \geq N_\epsilon \tag{33}$$

This will be written $\lim_{n\to\infty} r_n = r$ or simply $\{r_n\} \to r$.
The test for convergence (33) unfortunately involves prior knowledge of the limit r. A more convenient test for convergence is the

Theorem 17.1 (Cauchy Convergence Test) *A sequence of real number $\{r_n\}$ converges if and only if*

$$\lim_{n,k\to\infty} |r_{n+k} - r_n| = 0$$

or equivalently, if and only if given any $\epsilon > 0$, $\exists N_\epsilon$:

$$|r_{n+k} - r_n| \leq \epsilon \quad \textit{for } n \geq N_\epsilon$$

Example 17.2 Consider the sequence $\{r_n\}_1^\infty$ where

$$r_n = \sum_{m=1}^{n} \frac{1}{m^2}.$$

Observe that

$$|r_{n+k} - r_n| = \sum_{m=n+1}^{n+k} \frac{1}{m^2} \leq \int_n^{n+k} \frac{dx}{x^2} = \frac{k}{n(n+k)} \leq \frac{1}{n}$$

As a consequence, $\lim_{n,k\to\infty} |r_{n+k} - r_n| = 0$ and thus $\{r_n\}$ converges. A more delicate argument can be used to show that $\{r_n\} \to \pi^2/6$.

We are interested in convergence aspects for sequences of random variables for which there are several different convergence notions.

Definition 17.3 Let $\{X_n\}_1^\infty$ be a random sequence and Y be a random variable, all over the *same* probability space $(\Omega, \mathcal{A},$ Prob$)$.
$\{X_n\}$ *converges to* Y *in distribution*, written $X_n \xrightarrow{d} Y$, if

$$p_{X_n}(x) \to p_Y(x) \quad \text{for all } x$$

$\{X_n\}$ *converges to* Y *in probability*, written $X_n \xrightarrow{p} Y$ if for all $\epsilon > 0$

$$\text{Prob } \{|X_n - Y| \geq \epsilon\} \to 0$$

$\{X_n\}$ *converges to* Y *in quadratic mean*, written $X_n \xrightarrow{qm} Y$ if

$$E[|X_n - Y|^2] \to 0$$

$\{X_n\}$ *converges to* Y *almost surely* or *with probability* 1, written $\{X_n\} \xrightarrow{as} Y$ if Prob $\{S\} = 1$ where the event $S \in \mathcal{A}$ is

$$S = \{w : \{X_n(\omega)\} \to Y(\omega)\}$$

The following result elucidates the relationships between these various notions of convergence.

Theorem 17.4 *Let* $\{X_n\}$ *be a random sequence and* Y *be a random variable, all over the* same *probability space. Then*

(a) almost sure convergence $\Longrightarrow$ *convergence in probability*

(b) convergence in probability $\Longrightarrow$ *convergence in distribution*

(c) convergence in quadratic mean $\Longrightarrow$ *convergence in probability*

(d) if $\{X_n\}$ *is* bounded, *convergence in probability* $\Longrightarrow$ *convergence in quadratic mean*

18 Borel-Cantelli Lemma

Let $\{X_n\}_1^\infty$ be a random sequence. Fix $\epsilon > 0$ and define the events

$$\begin{aligned} A_{n,\epsilon} &= \{\omega : |X_k(\Omega)| \leq \epsilon \quad \text{for } k \geq n\} \\ T_\epsilon &= \cup_{n=1}^\infty A_{n,\epsilon} \end{aligned}$$

Observe that if $n_1 \geq n_2$ then $A_{n_1,\epsilon} \supseteq A_{n_2,\epsilon}$. Thus,

$$T_\epsilon = \cup_{n=1}^\infty A_{n,\epsilon} = \lim_{n\to\infty} \; A_{n,\epsilon}$$

Also, if $\epsilon_1 \geq \epsilon_2 > 0$ then $A_{n,\epsilon_1} \supseteq A_{n,\epsilon_2}$. Thus,

$$\cup_{n=1}^\infty A_{n,\epsilon_1} \supseteq \cup_{n=1}^\infty A_{n,\epsilon_2} \quad \text{or} \quad T_{\epsilon_1} \supseteq T_{\epsilon_2}$$

Consequently,

$$\cap_{\epsilon>0} \cup_{n=1}^\infty A_{n,\epsilon} = \cap_{\epsilon>0} T_\epsilon = \lim_{\epsilon\to 0} \; T_\epsilon = \lim_{\epsilon\to 0} \lim_{n\to\infty} \; A_{n,\epsilon}$$

Define the event

$$S = \{\omega : \{X_k\}(\omega) \to 0\}$$

Proposition 18.1 *$S = \cap_{\epsilon>0} \cup_{n=1}^\infty A_{n,\epsilon}$ and therefore $Prob(S) = \lim_{\epsilon\to 0} \; \lim_{n\to\infty} \; Prob(A_{n,\epsilon})$.*

Proof: Let $\omega \in S$. Then $X_k(\omega) \to 0$ or $\forall \epsilon > 0, \exists N(\epsilon,\omega)$ such that $|X_k(\omega)| \leq \epsilon$ for $k \geq N(\epsilon,\omega)$. which proves that $\omega \in \cap_{\epsilon>0} T_\epsilon$.
Next, if $\omega \in \cap_{\epsilon>0} T_\epsilon$ then $\forall \epsilon > 0, \Omega \in T_\epsilon$. It follows that $\forall \epsilon > 0, \exists N(\epsilon,\omega)$ such that $\omega \in A_{N(\epsilon,\omega),\epsilon}$. As a result, $\forall \epsilon > 0, \exists N(\epsilon,\Omega)$ such that $|X_k(\Omega)| \leq \epsilon$ for $k \geq N(\epsilon,\Omega)$ or equivalently, $\omega \in S$, proving the claim. □
We now prove the main tool to establish almost sure convergence.

Theorem 18.2 (Borel-Cantelli Lemma) *If for all $\epsilon > 0$,*

$$\sum_{k=1}^\infty Prob\ \{|X_k - Y| > \epsilon\} < \infty \tag{34}$$

then $X_n \xrightarrow{\text{as}} Y$.

Proof: Define a new random sequence $\{Z_n\}_1^\infty$ by $Z_n = X_n - Y$ and consider the complimentary events

$$\begin{aligned} A_{n,\epsilon} &= \{\omega : |Z_k(\omega)| \leq \epsilon \quad \text{for all } k \geq n\} \\ B_{n,\epsilon} &= \{\omega : |Z_k(\omega)| > \epsilon \quad \text{for some } k \geq n\} \end{aligned}$$

Then,

$$Prob(B_{n,\epsilon}) \leq \sum_n^\infty Prob\{|Z_n| > \epsilon\}$$

From (34) it must happen that $\forall \epsilon > 0$,

$$\lim_{n\to\infty} \sum_{n}^{\infty} \text{Prob } \{|Z_K| > \epsilon\} = 0$$

Thus, for all $\epsilon > 0$, we have that $Prob(B_{n,\epsilon}) \longrightarrow 0$ as $n \to \infty$. Consequently, from Proposition 18.1 we have

$$ProbS = \lim_{\epsilon\to 0} \lim_{n\to\infty} Prob(A_{n,\epsilon}) == 1 - \lim_{\epsilon\to 0} \lim_{n\to\infty} Prob(B_{n,\epsilon}) = 1$$

proving the claim. □

19 Proofs and Examples

Proof of Theorem 17.4(c): Suppose $\{X_n\}$ converges to Y in quadratic mean, i.e. $\mathbf{E}\left[|X_n - Y|^2\right] \to 0$. We show that $\{X_n\}$ also converges to Y in probability. Fix $\epsilon > 0$. From the Chebychev inequality (29) we have

$$Prob\{|X_n - Y| \geq \epsilon\} \leq \frac{\mathbf{E}\left[|X_n - Y|^2\right]}{\epsilon^2} \to 0$$

proving the claim. □

Example 19.1 We provide an example to illustrate that the converse of the above result is false. Consider the random sequence $\{X_n\}$ defined as

$$X_n = \begin{cases} n & \text{with probability } \frac{1}{n} \\ 0 & \text{otherwise} \end{cases}$$

and let $Y \equiv 0$. Observe that for any $\epsilon > 0$,

$$Prob\{|X_n - Y| \geq \epsilon\} = Prob\{X_n \geq \epsilon\} = \frac{1}{n} \to 0$$

proving that $\{X_n\} \longrightarrow 0$ in probability. However,

$$\mathbf{E}\left[|X_n - Y|^2\right] = \frac{1}{n} n^2 \not\longrightarrow 0$$

so we do not have convergence in quadratic mean.
For bounded random sequences however, convergence in probability and in quadratic mean are equivalent. This is shown next.

Proof of Theorem 17.4(d): Suppose $\{X_n\}$ is bounded and converges to Y in probability. We show that $\{X_n\}$ also converges in quadratic mean to Y. Define a new random sequence

$\{Z_n\}$ by $Z_n = X_n - Y$. Since X_n and Y are uniformly bounded, we can write $|Z_n| \leq M$, for some real number M. Next observe that for any $\epsilon > 0$,

$$\begin{aligned}
\mathbf{E}\left[Z_n^2\right] &= \int_{-M}^{M} z^2 p_{Z_n}(z)dz = \int_{|z|\leq\epsilon} z^2 p_{Z_n}(z)dz + \int_{\epsilon<|z|} z^2 p_{Z_n}(z)dz \\
&\leq \epsilon^2 \int_{|z|\leq\epsilon} p_{Z_n}(z)dz + M^2 \int_{\epsilon<|z|} p_{Z_n}(z)dz \\
&\leq \epsilon^2 + M^2 Prob\{|X_n - Y| > \epsilon\}
\end{aligned}$$

Then, since $Z_n \to 0$ in probability, we have

$$0 \leq \lim_{n\to\infty} \mathbf{E}\left[Z_n^2\right] \leq \epsilon^2$$

The above inequality holds for all $\epsilon > 0$, forcing $\lim_{n\to\infty} \mathbf{E}\left[Z_n^2\right] = 0$. Equivalently, $\{X_n\}$ converges to Y in quadratic mean. □

Example 19.2 Let $\{X_n\}$ be IID with $p_n(\cdot) = f(\cdot)$, and let Y be any random variable also with the same density function $f(\cdot)$. Clearly $\{X_n\}$ converges to Y in distribution. Convergence in distribution is very weak. It says nothing about the "sample paths" of $\{X_n\}$.

Example 19.3 Let $\{X_n\}$ be IID with $X_n \mathbb{N}(0, 1)$. Consider the random sequence $\{Z_n\}$ defined by

$$Z_n = \sum_{k=1}^{n} \frac{1}{k} X_k$$

It is easy to see that $\mathbf{E}[Z_n] = 0$. Since $\{X_n\}$ is IID we can make the following calculation

$$\sigma_n^2 = \mathbf{E}\left[Z_n^2\right] = \mathbf{E}\left[\sum_{j=1}^{n}\sum_{k=1}^{n} \frac{1}{jk} X_j X_k\right] = \sum_{k=1}^{n} \frac{1}{k^2} \longrightarrow \pi^2/6$$

The limit above can be computed using Fourier methods (see example (17.2). Since X_k are Gaussian, so is Z_n and we can immediately write its density because we know its second-order statistics :

$$p_{Z_n} = \frac{1}{\sqrt{2\pi}} \exp\{-\lambda^2/2\sigma_n^2\}$$

from which it is clear that $\{Z_n\}$ converges in distribution to a random variable with density $N(0, \pi^2/6)$.

Proof of Theorem 17.4(b): Suppose $\{X_n\}$ converges to Y in probability. We show that $\{X_n\}$ also converges in distribution. Fix $\lambda \in \mathbb{R}$ and $\epsilon > 0$. Let $P_n(\lambda) = Prob\{X_n < \lambda\}$ and define the event $A = \{\omega : |X_n(\omega) - Y(\omega)| \geq \epsilon\}$. Then, ...

Proof of Theorem 17.4(a): Suppose $\{X_n\}$ converges to Y almost surely. Then,

$$1 = Prob\{S\} = \lim_{\epsilon\to 0} \lim_{n\to\infty} Prob\{A_{n\epsilon}\}$$

where $A_{n,\epsilon} = \{\omega : |X_k(\omega)| \leq \epsilon \text{ for all } k \geq n\}$. This implies that

$$\lim_{\epsilon \to 0} \lim_{n \to \infty} \; Prob\{B_{n\epsilon} = 0$$

where $B_{n,\epsilon} = \{\omega : |X_k(\omega)| > \epsilon \text{ for some } k \geq n\}$.
Note that $B_{n,\epsilon} \supseteq \{\omega : |X_n(\omega)| > \epsilon\}$. Hence,

$$\lim_{\epsilon \to 0} \lim_{n \to \infty} \; Prob\{X_n| \geq \epsilon\} = 0$$

Fix $\delta > 0$. The above equation implies that $\exists \epsilon^0(\delta)$ and $N(\epsilon^0, \delta)$ such that

$$Prob\{|X_n| \geq \epsilon\} \leq \delta$$

for $n \geq N(\epsilon^0, \delta)$ and $\epsilon \leq \epsilon^0(\delta)$. However if $\epsilon \geq \epsilon^0(\delta)$,

$$Prob\{|X_n| \geq \epsilon\} \leq Prob\{|X_n| \geq \epsilon^\circ(\delta)\} \leq \delta$$

Thus, for all $n \geq N(\epsilon^0, \delta) = \widehat{N}(\delta)$ and *all* ϵ we have

$$0 \leq Prob\{|X_n| \geq \epsilon\} \leq \delta$$

or, $\lim n \to \infty \; Prob\{|X_n| \geq \epsilon\} = 0$ for all ϵ proving convergence in probability. □

Example 19.4 This standard example illustrates that convergence in probability $\not\Rightarrow$ almost sure convergence.
Let $(\omega, \mathcal{A}, Prob)$ be a probability space with $\omega = [0, 1]$ and consider the random variables $X_{n,m} : \omega \to \mathbb{R}$
This defines a countable collection of random variables $\{X_{n,m}\}$ for $n = 0, 1, 2, \cdots$ and $m = 1, 2, \cdots, 2^n$ which we could re-index as $\{X_k\}_1^\infty$ if we wanted to. Note that for any $0 < \epsilon < 1$,

$$Prob\{|X_{n,m}| \geq \epsilon\} = \frac{1}{2^n} \longrightarrow 0 \quad \text{as } m, n \to \infty$$

Thus, $\{X_{n,m}\} \longrightarrow 0$ in probability.
However, for *every* $\omega \in [0, 1]$,

$$X_{n,m}(\omega) \not\longrightarrow 0$$

because $X_{n,m}(\omega) = 1$ infinitely often. Thus, $\{X_{n,m}\} \not\longrightarrow 0$ almost surely. More precisely, $\{X_{n,m}\}$ *diverges* from 0 almost surely!

Exercise 19.5 Let $\{X_k\}_1^\infty$ be independent with

$$X_n = \begin{cases} n & \text{with probability } \frac{1}{n} \\ 0 & \text{otherwise} \end{cases}$$

Show that X does *not* converge to zero almost surely.

Exercise 19.6 Suppose $X_n \to c$ in probability where c is some constant. Let $f : \mathbb{R} \to \mathbb{R}$ be continuous at c. Show that $f(X_n) \to f(c)$ in probability.

Exercise 19.7 Let X be an *independent* random sequence and $c \in \mathbb{R}$. Show that either X converges to c almost surely or diverges from c almost surely. In other words, the event

$$S = \{\omega : \ X_n(\omega) \to c\}$$

either has $Prob\{S\} = 1$ or has $Prob\{S\} = 0$.

Exercise 19.8 Give an example of a random sequence $X = \{X_k\}_1^\infty$ and a random variable Y such that

$$0 < Prob\{X_n \to Y\} < 1$$

20 Large Number Laws

Theorem 20.1 (Weak Law of Large Numbers) *Let $\{X_k\}_1^\infty$ be uncorrelated. Suppose the means m_k, variances σ_k^2 exist and that*

$$\frac{1}{n^2}\sum_1^n \sigma_k^2 \to 0$$

Define $S_n = \sum_{k=1}^n (X_k - m_k)$. Then, $\{S_n/n\} \xrightarrow{qm} 0$.

Proof: Observe that $\mathsf{E}\left[\frac{1}{n^2}S_n^2\right] = \frac{1}{n^2}\sum_1^n \sigma_k^2 \to 0$, proving the claim. □

Example 20.2 Let $\{X_k\}_1^\infty$ be an IID random sequence with mean m and variance σ^2. Then, conditions of theorem are met, and

$$\left\{\frac{1}{n}(X_1 + \ldots + X_n)\right\} \xrightarrow{qm} m.$$

Equivalently, the "sample means"" converge in quadratic mean to the "ensemble" mean m.

Theorem 20.3 (Strong Law of Large Numbers) *Let $\{X_k\}_1^\infty$ be independent. Suppose*

$$\begin{aligned} E[X_k] &= m_k, \\ E[(X_k - m_k)^2] &= \sigma_k^2, \\ E[(X_k - m_k)^4] &= 4^{th} \text{ moment } \leq M \text{ for all } k \end{aligned} \tag{35}$$

i. e. the $1^{st}, 2^{nd}$ and 4^{th} moments exist and the 4^{th} moment is uniformly bounded. As before define $S_n = \sum_{k=1}^n (X_k - m_k)$. Then, $\{S_n/n\} \xrightarrow{a.s.} 0$.

Proof: First note that by the Cauchy-Schwartz inequality (Theorem 8.3),

$$\sigma_k^2 = \mathsf{E}\left[(X_k - m_k)^2 1^2\right] \leq \left[\mathsf{E}\left[(X_k - m_k)^4\right]\mathsf{E}\left[1^4\right]\right]^{1/2} \leq M^{1/2}. \tag{36}$$

Next, fix $\epsilon > 0$. Using the Chebychev inequality (Theorem 15.1) we have

$$Prob\{|\frac{1}{n}S_n| \geq \epsilon\} \leq \frac{\mathsf{E}\left[S_n^4/n^4\right]}{\epsilon^4} = \frac{\mathsf{E}\left[S_n^4\right]}{n^4\epsilon^4} \tag{37}$$

We can now expand

$$\mathbf{E}\left[S_n^4\right] = \mathbf{E}\left[\sum_{i,j,k,l=1}^{n} (X_i - m_i)(X_j - m_j)(X_k - m_k)(X_l - m_l)\right]$$

In the above expansion, there are n terms of the form $\mathbf{E}\left[(X_i - m_i)^4\right]$ (each of which are less than M) and $6n(n-1)/2$ terms of the form $\mathbf{E}\left[(X_i - m_i)^2(X_j - m_j)^2\right]$ with $i \neq j$. These terms are also less than M from (36).
The remaining terms look like

$$\mathbf{E}\left[(X_i - m_i)^3(X_j - m_j)\right] \quad \text{with} \quad i \neq j, \quad \text{or}$$
$$\mathbf{E}\left[(X_i - m_i)(X_j - m_j)(X_k - m_k)(X_l - m_l)\right] \quad \text{with} \quad i, j, k, l \text{ distinct}$$

Both of these types of terms are zero because $\{X_k\}_1^\infty$ is assumed independent. As a consequence (37) becomes

$$Prob\{|\frac{1}{n}S_n| \geq \epsilon\} \leq \left(\frac{6n(n-1)}{2}M + nM\right)\frac{1}{\epsilon^4 n^4} \leq (4n^2 M)\frac{1}{\epsilon^4 n^4} = \frac{4M}{\epsilon^4 n^2}$$

Next we have

$$\sum_1^\infty Prob\{|\frac{1}{n}S_n| \geq \epsilon\} \leq \sum_1^\infty \frac{4M}{\epsilon^4 n^2} < \infty$$

for all $\epsilon > 0$. We can then infer using the Borel-Cantelli Lemma (18.2) that

$$\left\{\frac{1}{n}S_n\right\} \xrightarrow{a.s.} 0$$

as claimed. □

Remark 20.4 The conditions (35) in the theorem statement can be relaxed to either

(a) $\{X_k\}_1^\infty$ IID and the mean $m = \mathbf{E}[X_k]$ exists

or the *Kolmogorov conditions*

(b) $\begin{cases} \{X_k\}_1^\infty \text{ independent and the means} m_k = \mathbf{E}[X_k] \text{ exist} \\ \text{and for some } M, \zeta > 0 \;\; \mathbf{E}\left[|X_k - m_k|^{1+\zeta}\right] \leq M \quad \text{for all } k \end{cases}$

Example 20.5 Let $\{X_k\}_1^\infty$ be IID, with mean m. Then, the strong law asserts that

$$\frac{1}{n}(X_1 + \ldots + X_n) \xrightarrow{as} m$$

This is a very strong statement: "sample path" means converge almost surely to the "ensemble" mean.

Example 20.6 Let $\{X_k\}_1^\infty$ be IID with

$$p_k(x) = \frac{1}{\pi(1+x^2)} \quad \text{(Cauchy density)}$$

It can be verified that $\Phi_{X_n}(\omega) = \mathbf{E}\left[e^{j\omega X_n}\right] = e^{-|\omega|}$. The Cauchy density is very badly behaved: the mean and variance do not exist. For example

$$\mathbf{E}[X_n] = \int_{-\infty}^{\infty} x \, \frac{dx}{\pi[1+x^2]} \stackrel{\triangle}{=} \lim_{T_1,T_2\to\infty} \int_{-T_1}^{T_2} \frac{xdx}{\pi[1+x^2]}$$

does not exist. We thus do not meet the requirements of either the strong or the weak number laws. In point of fact, since the density function is even we might cavalierly declare that $m = 0$. But

$$\Phi_{S_n/n}(\omega) = \mathbf{E}\left[e^{j\omega S_n/n}\right] = \mathbf{E}\left[e^{j\omega(X_1+\cdots+X_n)/n}\right] = \Pi_1^n \Phi_{X_k}(\omega/n) = \left(e^{\frac{-|\omega|}{n}}\right)^n = e^{-|\omega|}$$

Consequently, $p_{S_n/n}(\lambda) = 1/(\pi[1+\lambda^2])$ and clearly $S_n/n \not\longrightarrow 0$ in any sense.

Exercise 20.7 Let X be IID with each X_k being distribution according to the Cauchy density, i.e.

$$p_{X_k}(x) = \frac{1}{\pi(1+x^2)}$$

Using MATLAB, generate typical sample paths of X (for 3000 samples) and plot the running sample means. Observe that this does not "converge" to 0 because, as observed in class, the Large number laws do no apply to the Cauchy distribution.

21 The Central Limit Theorem

Essentially this very important result states that the sum of a large number of independent random variables asymptotically looks like it is normally distributed, regardless of the individual distributions.

Theorem 21.1 Central Limit Theorem *Let $\{X_k\}_1^\infty$ be IID and suppose the density function $p_{X_K}(\cdot) = f(\cdot)$ exists, and suppose that the mean m and variance σ^2 exist. Define as before*

$$S_n = (X_1 - m) + \cdots + (X_n - m)$$

Note that $\sigma_{S_n}^2 = \mathbf{E}\left[(X_1-m)^2 + \cdots + (X_n-m)^2\right] = n\sigma^2$ because X is IID. Then,

$$\left\{\frac{S_n}{\sigma\sqrt{n}}\right\} \xrightarrow{d} Y\mathbb{N}(0,1)$$

Proof: We compute the moment generating function of $S_n/\sigma\sqrt{n}$ as

$$\begin{aligned}\Phi_{\frac{S_n}{\sigma\sqrt{n}}} &= \mathbf{E}\left[\exp\{j\omega S_n/\sigma\sqrt{n}\}\right] = \Pi_{k=1}^{n}\mathbf{E}\left[\exp\{j\omega(X_n-m)/\sigma\sqrt{n}\}\right] \\ &\approx \Pi_1^n\mathbf{E}\left[1+\frac{j\omega(X_k-m)}{\sigma\sqrt{n}}-\frac{\omega^2(X_k-m)^2}{2n\sigma^2}\right] \\ &= \Pi_1^n\left(1-\frac{\omega^2}{2n}\right) = \left(1-\frac{\omega^2}{2n}\right)^n\end{aligned}$$

Note that for large n the approximation above becomes an equality. Thus we may write

$$\lim_{n\to\infty}\Phi_{S_n/\sigma\sqrt{n}} = \lim n\to\infty\left(1-\frac{\omega^2}{2n}\right)^n = \exp\{-\omega^2/2\}$$

Observe that the final expression above is the moment generating function for $Y\mathbb{N}(0,1)$. Therefore the corresponding densities also converge as claimed. We have assumed here that if a sequence of Fourier transforms converge, then the corresponding function also converge. □

Remark 21.2 The conditions on the Central limit theorem can be weakened to either

(*a*) $\{X_k\}$ are uniformly bounded and independent and $\sigma_{S_n}^2 \to \infty$

or the Lyapunov conditions

(*b*) $$\frac{1}{\sigma_{S_n}^{2+\zeta}}\sum_1^N\mathbf{E}\left[|X_k-m_k|^{2+\zeta}\right]\to 0$$

Note that in (a), (b) above we do not require that X_k admit a density function.

Example 21.3 (Bernoulli Trials) Consider an experiment $\mathcal{E}$ which consists of tossing countably many coins, each of which lands "heads" with probability p, and "tails" with probability $q = 1-p$. Define a random sequence $\{X_k\}_1^\infty$ by

$$X_k(\omega) = \begin{cases} 1 & \text{if the } k^{th} \text{ coin lands heads} \\ 0 & \text{else}\end{cases}$$

Note that $\{X_k\}_1^\infty$ is IID, $\mathbf{E}[X_k] = m = p$ and $\mathbf{E}\left[(X_k-p)^2\right] = pq$. Also, $\{X_k\}$ is uniformly bounded (by 1) and

$$\sigma_{S_n}^2 = \mathbf{E}\left[((X_1-m)+\cdots+(X_n-m))^2\right] = npq\to\infty$$

So we have met all the requirements (a) of the Central limit theorem, and even though X_k does not have a density. As a consequence we have

$$\left\{\frac{1}{pq\sqrt{n}}S_n\right\}\xrightarrow{d} N(0,1)$$

Does this sequence converge in quadratic mean ? How about $\{S_n/(npq)\}$?

22 Autocorrelation

At this point we will find it more convenient to deal with *two-sided* random sequences as $X = \{X_k\}_{-\infty}^{\infty}$. Let p_{t_k} and P_{t_k} denote the density and distribution functions respectively of X_k.
The random sequence X is *independent* if for any finite collection of indices $\{t_1, t_2, \cdots, t_N\}$ we have

$$P_{t_1,t_2,\cdots,t_N} = \Pi_{k=1}^{N} P_{t_k}$$

and X is *uncorrelated* if

$$\mathbf{E}\left[(X_i - m_i)\overline{(X_j - m_j)}\right] = 0, \quad \text{for} \quad i \neq j$$

Also, X is *Gaussian* if every finite collection $\{X_{t_1}, \cdots, X_{t_N}\}$ is jointly Gaussian. As before, X is independent $\Rightarrow$ X is uncorrelated. Also, X is Gaussian and uncorrelated $\Rightarrow$ X is independent.
The 2^{nd} *order statistics* of X are the means and the covariances:

$$m_t = \mathbf{E}[X_t], \quad R_{XX}(t,s) = \mathbf{E}\left[(X_t - m_t)\overline{(X_s - m_s)}\right]$$

The collection of covariances $R_{XX}(t,s)$ is called the *auto-correlation function* of the random sequence X.

Theorem 22.1 *(a)* $R_{XX}(t,s) = \overline{R_{XX}(s,t)}$

(b) Let $e = \{t_1, t_2, \cdots, t_N\}$ *be any collection of time indices and define*

$$\Lambda = [R_{XX}(t_i, t_j)] \in \mathbb{R}^{N \times N}$$

Then $\Lambda \geq 0$.

(c) $R_{XX}(t,t) \geq 0$ *for all* t.

(d) $R_{XX}(t,t) \cdot R_{XX}(s,s) \geq |R_{XX}(t,s)|^2$ *for all* t, s.

Proof: (a) $\overline{R_{XX}(s,t)} = \mathbf{E}\left[\overline{(X_s - m_s)}(X_t - m_t)\right] = R_{XX}(t,s)$.
(b) Define the random vector $Z = \begin{bmatrix} X_{t_1} - m_{t_1} & \cdots & X_{t_N} - m_{t_N} \end{bmatrix}'$. Then $\Lambda = \mathbf{E}[ZZ^*] \geq 0$ (see Theorem 8.1).
(c) follows from (b) on choosing $e = \{t\}$. (d) is the Cauchy-Schwartz inequality (see Theorem 8.3) applied to Z above with $e = \{t, s\}$. □
A random sequence X is called *strict-sense stationary* or SSS if for *any* finite collection of indices $\{t_1, \cdots, t_N\}$ and *all* integers T we have

$$P_{t_1,t_2,\cdots,t_N} = P_{t_1+T,t_2+T,\cdots,t_N+T} \tag{38}$$

In other words, the joint distribution of the random variables $\{X_{t_1}, X_{t_2}, \cdots, X_{t_N}\}$ is the same as for *any* "shift" of this collection of random variables.

A random sequence X is called *wide-sense-stationary* or WSS if

$$m_t = \mathbf{E}[X_t] = m = \text{ constant independent of } t \tag{39}$$

$$R_{XX}(t,s) = R_{XX}(t-s,0) \text{ for all } t,s \tag{40}$$

In this case, since $R_{XX}(t,s)$ depends only on $(t-s)$ it is more convenient to define R_{XX} via a single argument as

$$R_{XX}(t) \stackrel{\triangle}{=} \mathbf{E}\left[(X_{t+T}-m)\overline{(X_T-m)}\right] \quad \text{independent of } T \tag{41}$$

It is clear from these definitions that $SSS \Rightarrow WSS$. The converse is not true in general.
Given two random sequences X, Y, it is useful to define the *cross-correlation* function

$$R_{XY}(t,s) \stackrel{\triangle}{=} \mathbf{E}\left[(X_t - m_{X_t})\overline{(Y_s - m_{Y_s})}\right] \tag{42}$$

One of the most important WSS processes is *white-noise*. A random sequence $X = \{X_k\}_{-\infty}^{\infty}$ is called white noise if

(a) X is WSS, (b) $m = 0$ (zero mean), (c) $R_{XX}(t) = \begin{cases} 0 & t \neq 0 \\ \sigma^2 & t = 0 \end{cases}$

Theorem 22.1 specialized to $X : WSS$ immediately gives the following result.

Lemma 22.2 *Let X be WSS. Then,*

(a) $R_{XX}(-t) = \overline{R_{XX}(t)}$ *(conjugate-symmetry)*

(b) Let $e = \{t_1, \cdots, t_N\}$ *be any collection of indices. Then,* $\Lambda = [R_{XX}(t_i - t_j)] \geq 0$ *(positive semi-definiteness)*

(c) $R_{XX}(0) \geq 0$

(d) $R_{XX}(0) \geq |R_{XX}(t)|$ *for all* t.

23 Ergodicity

Let $X = \{X_k\}_{-\infty}^{\infty}$ be a SSS random sequence. Then, for any function f,

$$\mathbf{E}[f(X_t)] = \text{ ensemble average of } f = \text{ independent of } t$$

Unfortunately, $\mathbf{E}[f(X_t)]$ requires knowledge of $p_{X_t}(x)$ which is rarely available. Often, however, we can recover $\mathbf{E}[f(X_t)]$ from a *time-average* along a sample path of X.
A random sequence $X = \{X_k\}_{-\infty}^{\infty}$ is called *ergodic* if it is SSS and for all functions f such that $\mathbf{E}[f(X_t)] \leq \infty$, we have

$$\left\{\frac{1}{2n+1}\sum_{-n}^{n} f(X_t)\right\} \xrightarrow{as} \mathbf{E}[f(X_t)]$$

Equivalently, the time-averages of $f(X_t)$ are equal to the ensemble-averages.
Ergodicity is a very desirable property as it allows us to connect the axiomatic formulation of probability to the intuitive notion of time-averaging.

24 Power Spectral Density

From now, we deal exclusively with random sequences $X = \{X_k\}_{-\infty}^{\infty}$ that are WSS. Note that $R_{XX}(t)$ is a two-sided complex sequence.
We will *assume* that $R_{XX} \in \ell_1$, i.e.

$$||R_{XX}||_1 = \sum_{-\infty}^{\infty} |R_{XX}(t)| < \infty$$

As a consequence of $R_{XX} \in \ell_1$, we have that the sample means

$$S_N \stackrel{\triangle}{=} \frac{1}{2N+1} \sum_{-N}^{N} X_k$$

converge in quadratic mean to $E[X_k] = m$.
Proof: Observe that

$$\begin{aligned} \mathsf{E}\left[(S_N - m)^2\right] &= \frac{1}{(2N+1)^2} \mathsf{E}\left[\sum_{-N}^{N}(X_k - m) \sum_{-N}^{N}(X_i - m)\right] \\ &= \frac{1}{(2N+1)^2} \sum_{-N}^{N}(N - |k|)\ R_{XX}(k) \\ &\leq \frac{N}{(2N+1)^2}\ ||R_{XX}||_1 \to 0 \end{aligned}$$

Later, we will address the issue of recovering $R_{XX}(t)$ from sample means as

$$\frac{1}{2N+1} \sum_{-N}^{N}(X_{t+k}m)\overline{(X_k - m)} \stackrel{?}{\longrightarrow} R_{XX}(t)$$

Define the *power-spectral-density* of X as the Fourier Transform of R_{XX}, i.e.

$$S_{XX}(\omega) = \sum_{-\infty}^{\infty} R_{XX}(t) e^{-j\omega t} \tag{43}$$

Note that $S_{XX}(\omega)$ is periodic with period 2π. We can thus restrict our attention to $-\pi \leq \omega \leq \pi$. Also, $S_{XX}(\omega)$ exists for all ω because $R_{XX} \in \ell_1$.

Theorem 24.1 *(a)* $R_{XX}(t) = 1/2\pi \int_{-\pi}^{\pi} S_{XX}(\omega) e^{j\omega t} d\omega$

(b) $S_{XX}(\omega)$ *is* real *for all* $\omega \in [-\pi, \pi]$.

(c) $\int_{-\pi}^{\pi} S_{XX}(\omega) d\omega = 2\pi R_{XX}(0) \geq 0$

(d) $S_{XX}(\omega) \geq 0$ *for all* $\omega \in [-\pi, \pi]$

(e) If X *is a* real *random sequence,* $S_{XX}(\omega)$ *is* even, *i.e.* $S_{XX}(-\omega) = S_{XX}(\omega)$.

Proof:

(a) is just the inverse Fourier Transform formula.

(b) Since $R_{XX}(-t) = \overline{R_{XX}(t)}$, we have

$$\overline{S_{XX}(\omega)} = \sum_{-\infty}^{\infty} \overline{R_{XX}(t)} \tag{44}$$

$$= \sum_{-\infty}^{\infty} \overline{R_{XX}(-t)}\, e^{j\omega t} = S_{XX}(\omega). \tag{45}$$

(c) follows immediately from (a) upon noting that

$$R_{XX}(0) = E[(X_t - m)\overline{(X_t - m)}] \geq 0 \tag{46}$$

(d) will be proved later.

(e) If X is a real random sequence, then $R_{XX}(t)$ is real also, and

$$\begin{aligned} S_{XX}(-\omega) &= \sum_{-\infty}^{\infty} R_{XX}(t)e^{j\omega t} = \sum_{-\infty}^{\infty} R_{XX}(-t)e^{j\omega t} \\ &= \sum_{-\infty}^{\infty} R_{XX}(t)e^{-j\omega t} \\ &= S_{XX}(\omega) \end{aligned}$$

We now prove the following central result:

Theorem 24.2 *Let* $\mathbf{H}$ *be a stable, linear time-invariant system (not necessarily causal) with impulse response* $\{h_k\}_{-\infty}^{\infty}$ *and associated transfer function*

$$\mathbf{H}(z) = \sum_{-\infty}^{\infty} h_k z^{-k} \tag{47}$$

Let $X = \{X_k\}_{-\infty}^{\infty}$ *be WSS and consider the random sequence* $Y = \mathbf{H}(X)$*. Then*

(a) Y *is WSS*

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

Proof: All summations run from $-\infty$ to ∞.

(a) Note first that $Y_t = \sum_k h_k X_{t-k}$. Then,

$$\mathsf{E}\,[Y_t] = \left(\sum_k h_k\right) m_X = \mathbf{H}(1)m_X = m_Y$$

which is independent of t. Also,

$$\begin{aligned} R_{YY}(n+t,t) &= E[(Y_{n+t}-m_Y)\overline{(Y_t-m_Y)}] \\ &= E\left[\sum_i\sum_k h_i\overline{h}_k(X_{n+t-i}-m_X)\overline{(X_{t-k}-m_X)}\right] \\ &= \sum_i\sum_k h_i\overline{h}_k R_{XX}(n+k-i) \end{aligned}$$

which is independent of t, proving that Y is WSS.

(b) We compute

$$\begin{aligned} S_{YY}(\omega) &= \sum_t R_{YY}(t)e^{-j\omega t} \\ &= \sum_t\sum_i\sum_k h_i\overline{h}_k R_{XX}(t+k-i)e^{-j\omega t} \\ &= \sum_i\sum_k\sum_l h_i\overline{h}_k R_{XX}(l)e^{-j\omega l}e^{-j\omega i}e^{-j\omega k} \\ &= |\mathbf{H}(e^{j\omega})|^2 S_{XX}(\omega) \end{aligned}$$

[where did we use stability of H in this argument?]

The above theorem provides an interpretation for $S_{XX}(\omega)$ that justifies the nomenclature "power-spectral-density":
Let $\mathbf{H}$ be a narrow-pass (ideal) filter, i.e.

$$|\mathbf{H}(e^{j\omega})| = \begin{cases} 1 & \omega\in[\omega_0-\epsilon,\omega_0+\epsilon] \\ 0 & \text{else} \end{cases}$$

Then, using Theorem 2 (b), we have

$$S_{YY}(\omega) = \begin{cases} S_{XX}(\omega) & \omega\in[\omega_0-\epsilon,\omega_0+\epsilon] \\ 0 & \text{else} \end{cases}$$

Thus, Y has components in the frequency band $[\omega_0-\epsilon,\omega_0+\epsilon]$ only. Indeed, we can now establish Theorem 24.1(d) as:
Theorem 24.1(d) : $S_{XX}(\omega)\geq 0$ for all ω.
Proof: Suppose $S_{XX}(\omega_0)<0$ for some ω_0. Then, $S_{XX}(\omega)<0$ for some neighborhood $[\omega_0-\epsilon,\omega_0+\epsilon]$. Then, with $\mathbf{H}(z)$ being a narrow-pass filter as above, $S_{YY}(\omega)<0$ for $\omega\in[\omega_0-\epsilon,\omega_0+\epsilon]$ and $S_{YY}(\omega)=0$ else. Thus, $\int_{-\pi}^{\pi}S_{YY}(\omega)d\omega<0$ contradicting Theorem 1(c), proving the claim. □

Example 24.3 Let F,Φ be independent random variables with $\Phi\sim U[0,2\pi]$, and $|F|\leq\pi$ and define a random sequence $X=\{X_t\}_{-\infty}^{\infty}$ by

$$X_t = a\cos(Ft+\Phi)$$

We first show that X is WSS. To this end note that

$$\mathbf{E}[X_t] = a\mathbf{E}[\cos(Ft+\Phi)] = a\mathbf{E}[\cos(Ft)\cos(\Phi) - \sin(Ft)\sin(\Phi)] = 0$$

The last step above follows because F, Φ are independent & because $\Phi \sim U[0, 2\pi]$. Thus, the means $\mathbf{E}[X_t]$ are independent of t. Next,

$$\begin{aligned}
\mathbf{E}\left[X_t\overline{X}_s\right] &= a^2\mathbf{E}[\cos(Ft+\Phi)\cos(Fs+\Phi)] \\
&= \frac{a^2}{2}\mathbf{E}[\cos(F(t-s)) + \cos(Ft+2\Phi+Fs)] \\
&= \frac{a^2}{2}\mathbf{E}[\cos(F(t-s))] = \text{ function of } (t-s) \text{ only}
\end{aligned}$$

Note that the second term above can be seen to have zero expectation by expanding it out as $\cos(F(t+s))\cos(2\Phi) - \sin(F(t+s))\sin(2\Phi)$ and then using the fact that F, Φ are independent *and* $\Phi \sim U[0, 2\pi]$. Thus, X is WSS.
Now *assume* that $p_{_F}(f)$ is even. Then, since $|F| \leq \pi$,

$$\begin{aligned}
R_{_{XX}}(t) &= \frac{a^2}{2}\mathbf{E}[\cos(Ft)] = \frac{a^2}{2}\int_{-\pi}^{\pi}\cos(ft)p_{_F}(f)df \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi}a^2\pi p_{_F}(f)df
\end{aligned}$$

Recognizing the above expression as an inverse Fourier Transform, we conclude that the Fourier Transform of $R_{_{XX}}(t)$ is

$$S_{_{XX}}(\omega) = a^2\pi p_{_F}(\omega)$$

Using the above example, we can establish the following characterization of power spectral densities:

Theorem 24.4 *Consider any function $S(\omega)$, defined on $(-\pi, \pi]$ and such that $S(\omega) \geq 0$ for all ω. Then, there exists a random sequence $X = \{X_k\}_{-\infty}^{\infty}$ whose power spectral density is $S_{_{XX}}(\omega) = S(\omega)$*

Proof: We offer a proof in the special case where $S(\omega)$ is even (i.e. the underlying random sequence is real). Define X as in the previous example with

$$p_{_F}(f) = \frac{S(f)}{\int_{-\pi}^{\pi}S(\lambda)d\lambda} = \frac{S(f)}{c}$$

Note that this qualifies as a density function for F because it is nonnegative and has unit mass. Then, as shown in the example,

$$S_{_{XX}}(\omega) = a^2\pi p_{_F}(\omega) = \frac{a^2\pi}{c}S(\omega)$$

and the constant a can be chosen appropriately to prove the result. □