These notes are made available to students of the course ECE-541 at the University of New Mexico. They cannot and will not be used for any type of profitable reproduction or sale. Parts of the materials have been extracted from texts and notes by other authors. These include the text of Probability Theory by Y. Chow and H. Teicher, the notes on stochastic analysis by T. G. Kurtz, the text of Real and Complex Analysis by W. Rudin, the notes on measure and integration by A. Beck, the text of Probability and Measure by P. Billingsley, the text of Introduction to Stochastic Processes by P. G. Hoel, S. C. Port, and C. J. Stone, and possibly other sources. In compiling these notes I tried, as much as possible, to make explicit reference to these sources. If any omission is found, it is due to my oversight for which I seek the authors' forgiveness.

I am truly indebted to these outstanding scholars and authors that I mentioned above and whom I obtained most of the materials from; I am truly privileged to have been taught by some of them. I also like to thank Dr. Bradley Ratliff for helping in the tedious task of typing these notes.

This material is intended as a graduate-level treatment of probability and stochastic processes. It requires basic undergraduate knowledge of probability, random variables, probability distributions and density functions, and moments. A course like ECE340 will cover these topics at an undergraduate level. The material also requires some knowledge of elementary analysis; concepts such as limits and continuity, basic set theory, some basic topology, Fourier and Laplace transforms, and elementary linear systems theory.

\textit{Date:} December 17, 2008.
1. Fundamental concepts

1.1. Experiments. The most fundamental component in probability theory is the notion of a physical (or sometimes imaginary) “experiment,” whose “outcome” is revealed when the experiment is completed. Probability theory aims to provide the tools that will enable us to assess the likelihood of an outcome, or more generally, the likelihood of a collection of outcomes. Let us consider the following example:

Example 1. Shooting a dart: Consider shooting a single dart at a target (board) represented by the unit closed disc, $D$, which is centered at the point $(0,0)$. We write $D = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$. Here, $\mathbb{R}$ denotes the set of real numbers (same as $(-\infty, \infty)$), and $\mathbb{R}^2$ is the set of all points in the plane ($\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, where $\times$ denotes the Cartesian product of sets). We read the above description of $D$ as “the set of all points $(x, y)$ in $\mathbb{R}^2$ such that (or with the property) $x^2 + y^2 \leq 1$.”

1.2. Outcomes and the Sample Space. Now we define what we mean by an outcome: An outcome can be “missing the target,” or “missed” for short, in which case the dart misses the board entirely, or it can be its location in the scenario that it hits the board. Note that we have implicitly decided (or chose) that we do not care where the dart lands as whenever it misses the board. (The definition of an outcome is totally arbitrary and therefore it is not unique for any experiment. It depends on whatever makes sense to us.) Mathematically, we form what is called the sample space as the set containing all possible outcomes. If we call this set $\Omega$, then for our dart example, $\Omega = \{\text{missed}\} \cup D$, where the symbol $\cup$ denotes set union (we say $x \in A \cup B$ if and only if $x \in A$ or $x \in B$). We write $\omega \in \Omega$ to denote a specific outcome from the sample space $\Omega$. For example, we may have $\omega = \text{“missed,”}$ $\omega = (0,0)$, and $\omega = (0.1, 0.2)$; however, according to our definition of an outcome, $\omega$ cannot be $(1,1)$.

1.3. Events. An event is a collection of outcomes, that is, a subset in $\Omega$. Such a subset can be associated with a question that we may ask about the outcome of the experiment and whose answer can be determined after the outcome is revealed. For example, the question: “Q1: Did the dart land within 0.5 of the bull’s eye?” can be
associated with the subset of $\Omega$ (or event) given by $E = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1/4\}$. We would like to call the set $E$ an event. Now consider the complement of Q1, that is: Q2: “Did the dart not land within 0.5 of the bull’s eye?” with which we can associate the event $E^c$, were the superscript “c” denotes set complementation (relative to $\Omega$). Namely, $E^c = \Omega \setminus E$, which is the set of outcomes (or points or members) of $\Omega$ that are not already in $E$. The notation “\” represents set difference (or subtraction). In general, for any two sets $A$ and $B$, $A \setminus B$ is the set of all elements that are in $A$ but not in $B$. Note that $E^c = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 > 1/4\} \cup \{\text{missed}\}$. The point here is that if $E$ is an event, then we would want $E^c$ to qualify as an event as well since we would like to be able to ask the logical negative of any question. In addition, we would also like to be able to form a logical “or” of any two questions about the experiment outcome. Thus, if $E_1$ and $E_2$ are events, we would like their union to also be an event.

Here is another illustrative example: for each $n = 1, 2, \ldots$, define the subset $E_n = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1 - 1/n\}$ and let $\bigcup_{n=1}^{\infty} E_n$ be their countable union. (Notation: we say $\omega \in \bigcup_{n=1}^{\infty} E_n$ if and only if $\omega \in E_n$ for some $n$. Similarly, we say $\omega \in \bigcap_{n=1}^{\infty} E_n$ if and only if $\omega \in E_n$ for all $n$.) It is not hard to see (prove it) that $\bigcap_{n=1}^{\infty} E_n = E = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$, which corresponds to the valid question “did the dart land \textit{inside} the board?” Thus, we would want $E$ (which is the countable union of events) to be an event as well.

**Example 2.** For each $n \geq 1$, take $A_n = (-1/n, 1/n)$. Then, $\bigcap_{n=1}^{\infty} A_n = \{0\}$ and $\bigcup_{n=1}^{\infty} A_n = (-1, 1)$. Prove these.

Finally, we should be able to ask whether or not the experiment was conducted or not, that is, we would like to label the sample space $\Omega$ as an event as well. With this (hopefully) motivating introduction, we proceed to formally define what we mean by events.

**Definition 1.** A collection $\mathcal{F}$ of subsets of $\Omega$ is called a $\sigma$-algebra (reads as sigma-algebra) if:

1. $\Omega \in \mathcal{F}$
If $\mathcal{F}$ is a $\sigma$-algebra, then its members are called *events*. Note that $\mathcal{F}$ is a collection of subsets and not a union of subsets; thus, $\mathcal{F}$ itself is not a subset of $\Omega$.

Here are some consequences of the above definition (you should prove all of them):

1. $\emptyset \in \mathcal{F}$.
2. $\bigcap_{n=1}^{\infty} E_n \in \mathcal{F}$ whenever $E_n \in \mathcal{F}$, $n = 1, 2, \ldots$. Here, the countable intersection $\bigcap_{n=1}^{\infty} E_n$ is defined as follows: $\omega \in \bigcap_{n=1}^{\infty} E_n$ if and only if $\omega \in E_n$ for all $n$.
3. $A \setminus B \in \mathcal{F}$ whenever $A, B \in \mathcal{F}$. (First prove that $A \setminus B = A \cap B^c$.)

Generally, members of any $\sigma$-algebra (not necessary corresponding to a random experiment and a sample space $\Omega$) are called *measurable sets*. Measurable sets were first introduced in the branch of mathematics called analysis. They were adopted to probability by a great mathematician named Kolmogorov. Clearly, in probability theory we call measurable subsets of $\Omega$ events.

**Definition 2.** Let $\Omega$ be a sample space and let $\mathcal{F}$ be a $\sigma$-algebra of events. We call the pair $(\Omega, \mathcal{F})$ a *measurable space*.

**Definition 3.** A collection $\mathcal{D}$ of subsets of $\Omega$ is called a sub-$\sigma$-algebra of $\mathcal{D}$ if

1. $\mathcal{D} \subset \mathcal{F}$ (this means that if $A \in \mathcal{D}$, then automatically $A \in \mathcal{F}$).
2. $\mathcal{D}$ is itself a $\sigma$-algebra.

**Example 3.** $\{\emptyset, \Omega\}$ is a sub-$\sigma$-algebra of any other $\sigma$-algebra.

**Example 4.** The power set of $\Omega$, which is the set of all subsets of $\Omega$, is a $\sigma$-algebra. In fact it is a maximal $\sigma$-algebra in a sense that it contains any other $\sigma$-algebra. The power set of a set $\Omega$ is often denoted by the notation $2^{\Omega}$.

**Interpretation:** Once again, we emphasize that it would be meaningful to think of a $\sigma$-algebra as a collection of all valid questions that one may ask about an experiment. The collection has to satisfy certain self-consistency rules, dictated by the
requirements for a \(\sigma\)-algebra, but what we mean by “valid” is really up to us as long as the self-consistency rules defining the collection of events are met.

**Generation of \(\sigma\)-algebras:** Let \(\mathcal{M}\) be a collection of events (not necessarily a \(\sigma\)-algebra); that is, \(\mathcal{M} \subset \mathcal{F}\). This collection could be a collection of certain events of interest. For example, in the dart experiment we may define \(\mathcal{M} = \{\text{missed}, \{(x, y) \in \mathbb{R}^2 : 1/4 \leq x^2 + y^2 \leq 1/2\}\}\). The main question now is the following: Can we construct a minimal (or smallest) \(\sigma\)-algebra that contains \(\mathcal{M}\)? If such a \(\sigma\)-algebra exists, call it \(\mathcal{F}_M\), then it must possess the following property: If \(\mathcal{D}\) is another \(\sigma\)-algebra containing \(\mathcal{M}\), then necessarily \(\mathcal{F}_M \subset \mathcal{D}\). Hence \(\mathcal{F}_M\) is minimal.

The following Theorem states that there is such a minimal \(\sigma\)-algebra.

**Theorem 1.** Let \(\mathcal{M}\) be a collection of events, then there is a minimal \(\sigma\)-algebra containing \(\mathcal{M}\).

Before we prove the Theorem let us look at an example.

**Example 5.** Suppose that \(\Omega = (-\infty, \infty)\), and \(\mathcal{M} = \{(-\infty, 1), (0, \infty)\}\). It is easy to check that the minimal \(\sigma\)-algebra containing \(\mathcal{M}\) is

\[\mathcal{F}_M = \{\emptyset, \Omega, (-\infty, 1), (0, \infty), (0, 1), (-\infty, 0], [1, \infty), (-\infty, 0] \cup [1, \infty)\}\]

*Explain where each member is coming from.*

**Proof of Theorem:** Let \(\mathcal{K}_M\) be the collection of all \(\sigma\)-algebras that contain \(\mathcal{M}\). We observe that such a collection is not empty since at least it contains the power set \(2^\Omega\). Let us label each member of \(\mathcal{K}_M\) by an index \(\alpha\), namely \(D_\alpha\), where \(\alpha \in I\), where \(I\) is some index set. Define \(\mathcal{F}_M = \bigcap_{\alpha \in I} D_\alpha\). We need to show that 1) \(\mathcal{F}_M\) is a \(\sigma\)-algebra containing \(\mathcal{M}\), and 2) that \(\mathcal{F}_M\) is a minimal \(\sigma\)-algebra. Note that each \(D_\alpha\) contains \(\Omega\) (since each \(D_\alpha\) is a \(\sigma\)-algebra); thus, \(\mathcal{F}_M\) contains \(\Omega\). Next, if \(A \in \mathcal{F}_M\), then \(A \in D_\alpha\), for each \(\alpha \in I\). Thus, \(A^c \in D_\alpha\), for each \(\alpha \in I\) (since each \(D_\alpha\) is a \(\sigma\)-algebra) which implies that \(A^c \in \bigcap_{\alpha \in I} D_\alpha\). Next, suppose that \(A_1, A_2, \ldots \in \mathcal{F}_M\). Then, we know that \(A_1, A_2, \ldots \in D_\alpha\), for each \(\alpha \in I\). Moreover, \(\bigcup_{n=1}^{\infty} A_n \in D_\alpha\),
for each $\alpha \in I$ (again, since each $D_\alpha$ is a $\sigma$-algebra) and thus $\bigcup_{n=1}^{\infty} A_n \in \bigcap_{\alpha \in I} D_\alpha$.
This completes proving that $\mathcal{F}_M$ is a $\sigma$-algebra. Now suppose that $\mathcal{F}'_M$ is another $\sigma$-algebra containing $\mathcal{M}$, we will show that $\mathcal{F}_M \subset \mathcal{F}'_M$. First note that since $\mathcal{K}_M$ is the collection of all $\sigma$-algebras containing $\mathcal{M}$, then it must be true that $\mathcal{F}'_M = D_{\alpha^*}$ for some $\alpha^* \in I$. Now if $A \in \mathcal{F}_M$, then necessarily $A \in D_{\alpha^*}$ (since $D_{\alpha^*}$ is one of the members of the countable intersection that defines $\mathcal{F}_M$), and consequently $A \in \mathcal{F}'_M$, since $\mathcal{F}'_M = D_{\alpha^*}$. This establishes that $\mathcal{F}_M \subset \mathcal{F}'_M$ and completes the proof of the theorem. \hfill $\Box$

**Example 6.** Let $U$ be the collection of all open sets in $\mathbb{R}$. Then, according to the above theorem, there exists a minimal $\sigma$-algebra containing $U$. This $\sigma$-algebra is called the Borel $\sigma$-algebra, $\mathcal{B}$, and its elements are called the Borel subsets of $\mathbb{R}$. Note that by virtue of set complementation, union and intersection, $\mathcal{B}$ contains all closed sets, half open intervals, their countable unions, intersections, and so on.

(Reminder: A subset $U$ in $\mathbb{R}$ is called open if for every $x \in U$, there exists an open interval centered at $x$ which lies entirely in $U$. Closed sets are defined as complements of open sets. These definitions extend to $\mathbb{R}^n$ in a straightforward manner.)

**Restrictions of $\sigma$-algebras:** Let $\mathcal{F}$ be a $\sigma$-algebra. For any measurable set $U$, we define $\mathcal{F} \cap U$ as the collection of all intersections between $U$ and the members of $\mathcal{F}$, that is, $\mathcal{F} \cap U = \{V \cap U : V \in \mathcal{F}\}$. It is easy to show that $\mathcal{F} \cap U$ is also a $\sigma$-algebra, which is called the restriction of $\mathcal{F}$ to $U$.

**Example 7.** Back to the dart experiment: What is a reasonable $\sigma$-algebra for this experiment? I would say any such $\sigma$-algebra should contain all the Borel subsets of $D$ and the \{missed\} event. So we can take $\mathcal{M} = \{\{\text{missed}\}, \mathcal{B} \cap D\}$. It is easy to check that in this case

\begin{equation}
\mathcal{F}_M = (\mathcal{B} \cap D) \cup \{\text{missed}\} \cup (\mathcal{B} \cap D),
\end{equation}

where for any $\sigma$-algebra $\mathcal{F}$ and any measurable set $U$, we define $\mathcal{F} \cup U$ as the collection of all unions between $U$ and the members of $\mathcal{F}$, that is, $\mathcal{F} \cup U = \{V \cup U : V \in \mathcal{F}\}$. Note that $\mathcal{F} \cup U$ is not always a $\sigma$-algebra (contrary to $\mathcal{F} \cap U$), but in this example it is because $\{\text{missed}\}$ is the complement of $D$. 

1.4. Random Variables. Motivation: Recall the dart experiment, and define the following transformation on the sample space:

$$X: \Omega \rightarrow \mathbb{R}$$ defined as

$$X(\omega) = \begin{cases} 
1, & \text{if } \omega \in D \\
0, & \text{if } \omega = \text{missed} 
\end{cases}$$

where as before $$D = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$$.

Consider the collection of outcomes that we can identify if we knew that $$X$$ fell in the interval $$(-\infty, r)$$. More precisely, we want to identify $$\{\omega \in \Omega : X(\omega) \in (-\infty, r)\}$$, or equivalently, the set $$X^{-1}((-\infty, r))$$, which is the inverse image of the set $$(-\infty, r)$$ under the transformation (or function) $$X$$. It can be checked that (you must work these out carefully):

- For $$r < 0$$, $$X^{-1}((-\infty, r)) = \emptyset$$,
- for $$0 \leq r < 1$$, $$X^{-1}((-\infty, r)) = \{\text{missed}\}$$,
- and for $$1 \leq r < \infty$$, $$X^{-1}((-\infty, r)) = \Omega$$.

The important thing to note is that in each case, $$X^{-1}((-\infty, r))$$ is an event; that is, $$X^{-1}((-\infty, r)) \in \mathcal{F}_M$$, where $$\mathcal{F}_M$$ was defined earlier for this experiment (1). This is a direct consequence of the way we defined the function $$X$$.

Here is another transformation defined on the outcomes of the dart experiment:
Define

$$(2) \quad Y(\omega) = \begin{cases} 
10, & \text{if } \omega = \text{missed} \\
\sqrt{x^2 + y^2}, & \text{if } \omega \in D.
\end{cases}$$

Let’s consider the collection of outcomes that correspond to $$Y < 1/2$$, which we can write as $$\{\omega \in \Omega : Y(\omega) < 1/2\} = Y^{-1}((-\infty, 1/2))$$. Note that

$$Y^{-1}((-\infty, 1/2)) = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1/4\} \in \mathcal{F}_M.$$ 

Moreover, we can also show that

$$Y^{-1}((-\infty, 2)) = D \in \mathcal{F}_M$$

$$Y^{-1}((-\infty, 11)) = D \cup \{\text{missed}\} = \Omega \in \mathcal{F}_M$$

$$Y^{-1}((-\infty, 0]) = \{0\} \in \mathcal{F}_M$$
We emphasize again that \( Y^{-1}((-\infty, r)) \) is always an event; that is, \( Y^{-1}((-\infty, r)) \in \mathcal{F}_M \), where \( \mathcal{F}_M \) was defined earlier for this experiment (1). Again, this is a direct consequence of the way we defined the function \( Y \).

Motivated by these examples, we proceed to define what we mean by a random variable in general.

**Definition 4.** Let \( (\Omega, \mathcal{F}) \) be a measurable space. A transformation \( X : \Omega \to \mathbb{R} \) is said to be \( \mathcal{F} \)-measurable if for every \( r \in \mathbb{R}, X^{-1}((-\infty, r)) \in \mathcal{F} \). Any measurable \( X \) is called a random variable.

Now let \( X \) be a random variable and consider the collection of events \( \mathcal{M} = \{ \omega \in \Omega : X(\omega) \in (-\infty, r), r \in \mathbb{R} \} \), which can also be written more conveniently as \( \{ X^{-1}((-\infty, r)), r \in \mathbb{R} \} \). As before, let \( \mathcal{F}_M \) be the minimal \( \sigma \)-algebra containing \( \mathcal{M} \). Then, \( \mathcal{F}_M \) is the “information” that the random variable \( X \) conveys about the experiment. We define such a \( \sigma \)-algebra from this point on as \( \sigma(X) \), the \( \sigma \)-algebra generated by the random variable \( X \). In the above example of \( X \), \( \sigma(X) = \{ \emptyset, \{ \text{missed} \}, \Omega, D \} \).

This concept is demonstrated in the next example. (Also see HW#1 for another example of \( \sigma(X) \).)

**Example 8.** Back to the dart example: Let \( \mathcal{M}' \triangleq \{ \emptyset, \{ \text{missed} \}, \Omega \} \), then it is easy to check that \( \mathcal{F}_{M'} = \{ \emptyset, \{ \text{missed} \}, \Omega, D \} \), which also happens to be a subset of \( \mathcal{F}_M \) as defined in (1). However, we observe that in fact \( X^{-1}((-\infty, r)) \in \mathcal{F}_{M'} \) for any \( r \in \mathbb{R} \). Intuitively, \( \mathcal{F}_{M'} \), which also call \( \sigma(X) \), can be identified as the set of all events that the function \( X \) can convey about the experiment. In particular, \( \mathcal{F}_{M'} \) consists of precisely those events whose occurrence or nonoccurrence can be determined through our observation of the value of \( X \). In other words, \( \mathcal{F}_{M'} \) is all the “information” that the mapping \( X \) can provide about the outcome of the experiments. Note that \( \mathcal{F}_{M'} \) is much smaller than the original \( \sigma \)-algebra \( \mathcal{F} \), which was \( ((\mathcal{B} \cap D) \cup \{ \text{missed} \}) \cup (\mathcal{B} \cap D) \). As we have seen before, \( \{ \emptyset, \{ \text{missed} \}, \Omega, D \} \subset ((\mathcal{B} \cap D) \cup \{ \text{missed} \}) \cup (\mathcal{B} \cap D) \). Thus, \( X \) can only partially inform us of the true outcome of the experiment; it can only tell us if we hit the target or missed it, nothing else, which is precisely the information contained in \( \mathcal{F}_{M'} \).
Facts about Measurable Transformations:

**Theorem 2.** Let \((\Omega, \mathcal{F})\) be a measurable space. The following statements are equivalent:

1. \(X\) is a measurable transformation.
2. \(X^{-1}((-\infty, r]) \in \mathcal{F}\) for all \(r \in \mathbb{R}\).
3. \(X^{-1}((r, \infty)) \in \mathcal{F}\) for all \(r \in \mathbb{R}\).
4. \(X^{-1}([r, \infty)) \in \mathcal{F}\) for all \(r \in \mathbb{R}\).
5. \(X^{-1}((a, b)) \in \mathcal{F}\) for all \(a < b\).
6. \(X^{-1}([a, b)) \in \mathcal{F}\) for all \(a < b\).
7. \(X^{-1}([a, b]) \in \mathcal{F}\) for all \(a < b\).
8. \(X^{-1}((a, b]) \in \mathcal{F}\) for all \(a < b\).
9. \(X^{-1}(B) \in \mathcal{F}\) for all \(B \in \mathcal{B}\).

The equivalence amongst (1), (2) and (3) is left as an exercise (see HW1). The equivalence amongst (3) and (4) through (8) can be shown using the same technique used in HW1. To show that (1) implies (9) requires additional work and will not be shown here. In fact, (9) is the most powerful definition of a random variable. Using (9), we can equivalently define \(\sigma(X)\) as \(\{X^{-1}(B) : B \in \mathcal{B}\}\), which can be directly shown to be a \(\sigma\)-algebra (see HW1).
1.5. Probability Measure.

**Definition 5.** Consider the measurable space \((\Omega, \mathcal{F})\), and a function, \(P\), that maps \(\mathcal{F}\) into \(\mathbb{R}\). \(P\) is called a *probability measure* if

1. \(P(E) \geq 0\), for any \(E \in \mathcal{F}\).
2. \(P(\Omega) = 1\).
3. If \(E_1, E_2, \ldots \in \mathcal{F}\) and if \(E_i \cap E_j = \emptyset\) when \(i \neq j\), then \(P(\bigcup_{n=1}^{\infty} E_n) = \sum_{n=1}^{\infty} P(E_n)\).

The following properties follow directly from the above definition.

**Property 1.** \(P(\emptyset) = 0\).

Proof. Put \(E_1 = \Omega, E_2 = \ldots = E_n = \emptyset\) in (3) and use (2) to get \(1 = P(\Omega) = P(\Omega \cup \emptyset \cup \emptyset \cup \ldots) = P(\Omega) + \sum_{n=2}^{\infty} P(\emptyset) = 1 + \sum_{n=2}^{\infty} P(\emptyset)\). Thus, \(\sum_{n=2}^{\infty} P(\emptyset) = 0\), which implies that \(P(\emptyset) = 0\) since \(P(\emptyset) \geq 0\) according to (1).

**Property 2.** If \(E_1, E_2, \ldots, E_n \in \mathcal{F}\) and if \(E_i \cap E_j = \emptyset\) when \(i \neq j\), then \(P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)\).

Proof. Put \(E_{n+1} = E_{n+2} = \ldots = \emptyset\) and the result will follow from 3 since \(P(\emptyset) = 0\) (from Property 2).

**Property 3.** If \(E_1, E_2 \in \mathcal{F}\) and \(E_1 \subset E_2\), then \(P(E_1) \leq P(E_2)\).

Proof. Note that \(E_1 \cup E_2 \setminus E_1 = E_2\) and \(E_1 \cap E_2 \setminus E_1 = \emptyset\). Thus, by Property 2 (use \(n = 2\)), \(P(E_2) = P(E_1) + P(E_2 \setminus E_1) \geq P(E_1)\), since \(P(E_2 \setminus E_1) \geq 0\).

**Property 4.** If \(A_1 \subset A_2 \subset A_3 \ldots \in \mathcal{F}\), then

\[
\lim_{n \to \infty} P(A_n) = P(\bigcup_{n=1}^{\infty} A_n).
\]

Proof. Put \(B_1 = A_1, B_2 = A_2 \setminus A_1, \ldots, B_n = A_n \setminus A_{n-1}, \ldots\). Then, it is easy to check that \(\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n\) and \(\bigcup_{n=1}^{m} A_n = \bigcup_{n=1}^{m} B_n\) for any \(m \geq 1\), and that \(B_i \cap B_j = \emptyset\) when \(i \neq j\). Hence, \(P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(B_i)\) and \(P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(B_i)\).
But, $\bigcup_{i=1}^{n} A_i = A_n$, so that $P(A_n) = \sum_{i=1}^{n} P(B_i)$. Now since $\sum_{i=1}^{n} P(B_i)$ converges to $\sum_{i=1}^{\infty} P(B_i)$, we conclude that $P(A_n)$ converges to $\sum_{i=1}^{\infty} P(B_i)$, which is equal to $P(\bigcup_{i=1}^{\infty} A_i)$.

**Property 5.** If $A_1 \supset A_2 \supset A_3 \ldots$, then $\lim_{n \to \infty} P(A_n) = P(\bigcap_{n=1}^{\infty} A_n)$.

Proof. See HW.

**Property 6.** For any $A \in \mathcal{F}$, $0 \leq P(A) \leq 1$.

The triplet $(\Omega, \mathcal{F}, P)$ is called a probability space.

**Example 9.** Recall the dart experiment. We now define $P$ on $(\Omega, \mathcal{F}_M)$ as follows. Assign $P(\{\text{missed}\}) = 0.5$, and for any $A \in D \cap B$, assign $P(A) = \text{area}(A)/2\pi$. It is easy to check that $P$ defines a probability measure. (For example, $P(\Omega) = P(D \cup \{\text{missed}\}) = P(D) + P(\{\text{missed}\}) = \text{area}(D)/2\pi + 0.5 = 0.5 + 0.5 = 1$. Check the other requirements as an exercise.)

1.6. **Distributions and distribution functions:** Consider a probability space $(\Omega, \mathcal{F}, P)$, and consider a random variable $X$ defined on it. Up to this point, we have developed a formalism that allows us to ask questions of the form “what is the probability that $X$ assumes a value in a Borel set $B$?” Symbolically, this is written as $P(\{\omega \in \Omega : X(\omega) \in B\})$, or for short, $P\{X \in B\}$ with the understanding that the set $\{X \in B\}$ is an event (i.e., member of $\mathcal{F}$). Answering all the questions of the above form is tantamount to assigning a number in the interval $[0,1]$ to every Borel set. Thus, we can think of a mapping from $B$ into $[0,1]$, whose knowledge will provide an answer to all the questions of the form described earlier. We call this mapping the distribution of the random variable $X$, and it is denoted by $\mu_X$. Formally, we have $\mu_X : \mathcal{B} \to [0,1]$ according to the rule $\mu_X(B) = P\{X \in B\}, B \in \mathcal{B}$.

**Proposition 1.** $\mu_X$ defines a probability measure on the measurable space $(\mathbb{R}, \mathcal{B})$.

Proof. See HW.
**Distribution Functions:** Recall from your undergraduate probability that we often associate with each random variable a distribution function, defined as $F_X(x) = \mathbb{P}\{X \leq x\}$. This function can also be obtained from the distribution of $X$, $\mu_X$, by evaluating $\mu_X$ at $B = (-\infty, x]$, which is a Borel set. That is, $F_X(x) \triangleq \mu_X((-\infty, x])$.

Note that for any $x \leq y$, $\mu_X((x, y]) = F_X(y) - F_X(x)$.

**Property 7.** $F_X$ is nondecreasing.

Proof. For $x_1 \leq x_2, (-\infty, x_1] \subset (-\infty, x_2]$ and $\mu_X((-\infty, x_1]) \leq \mu_X((-\infty, x_2])$ since $\mu_X$ is a probability measure (see Property 3 above).

**Property 8.** $\lim_{x \to -\infty} F_X(x) = 1$ and $\lim_{x \to -\infty} F_X(x) = 0$.

Proof. Note that $(-\infty, \infty) = \bigcup_{n=1}^{\infty} (-\infty, n]$ and by Property 4 above, $\lim_{n \to \infty} \mu_X((-\infty, n]) = \mu_X(\bigcup_{n=1}^{\infty} (-\infty, n]) = \mu_X((-\infty, \infty)) = 1$ since $\mu_X$ is a probability measure. Thus we proved that $\lim_{n \to \infty} F_X(n) = 1$. Now the same argument can be repeated if we replace the sequence $n$ by any increasing sequence $x_n$. Thus, $\lim_{x_n \to -\infty} F_X(x_n) = 1$ for any increasing sequence $x_n$, and consequently $\lim_{x \to -\infty} F_X(x) = 1$.

The proof of the second assertion is left as an exercise.

**Property 9.** $F_X$ is right continuous, that is, $\lim_{n \to \infty} F_X(x_n) = F_X(y)$ for any monotonic and convergent sequence $x_n$ for which $x_n \downarrow y$.

Proof. For simplicity, assume that $x_n = y + n^{-1}$. Note that $F_X(y) = \mu_X((-\infty, y])$, and $(-\infty, y] = \bigcap_{n=1}^{\infty} (-\infty, y + n^{-1}]$. So, by Property 5 above, $\lim_{n \to \infty} \mu_X((-\infty, y + n^{-1}]) = \mu_X(\bigcap_{n=1}^{\infty} (-\infty, y + n^{-1}]) = \mu_X((-\infty, y])$. Thus, we proved that $\lim_{n \to \infty} F_X(y + n^{-1}) = F_X(y)$. In the same fashion, we can generalize the result to obtain $\lim_{n \to \infty} F_X(y + x_n) = F_X(y)$ for any sequence for which $x_n \downarrow 0$. This completes the proof.

**Property 10.** $F_X$ has a left limit at every point, that is, $\lim_{n \to \infty} F_X(x_n)$ exists for any monotonic and convergent sequence $x_n$.

Proof. For simplicity, assume that $x_n = y - n^{-1}$ for some $y$. Note that $(-\infty, y) = \bigcup_{n=1}^{\infty} (-\infty, y - n^{-1}]$. So, by Property 4 above, $\lim_{n \to \infty} F_X(y - n^{-1}) = \lim_{n \to \infty} \mu_X((-\infty, y-}$
\[ n^{-1}] = \mu_X(\bigcup_{n=1}^{\infty}(-\infty, y - n^{-1}]) = \mu_X((-\infty, y)) \triangleq F_X(y^-). \]

In the same fashion, we can generalize the result to obtain \( \lim_{n \to \infty} F_X(x_n) = \mu_X((-\infty, y)) \) for any sequence \( x_n \uparrow y \).

**Property 11.** \( F_X \) has at most countably many discontinuities.

Proof. Let \( D \) be the set of discontinuity points of \( F_X \). We first show a simple fact that says that the jumps (or discontinuity intervals) corresponding to distinct discontinuity points are disjoint. More precisely, pick \( \alpha, \beta \in D \), and suppose, without loss of generality, that \( \alpha < \beta \). Let \( I_\alpha = (F_X(\alpha^-), F_X(\alpha]) \) and \( I_\beta = (F_X(\beta^-), F_X(\beta]) \) represent the discontinuities associated with \( \alpha \) and \( \beta \), respectively. Note that \( F_X(\alpha) < F_X(\beta^-) \); this follows from the definition of \( F_X(\beta^-) \), the fact that \( \alpha < \beta \), and the fact that \( F_X \) is nondecreasing. Hence, \( I_\alpha \) and \( I_\beta \) are disjoint. From this we also conclude that the discontinuities (jumps) associated with the points of \( D \) form a collection of disjoint intervals. (*)

Next, note that \( D = \bigcup_{n=1}^{\infty} D_n \), where \( D_n = \{ x \in \mathbb{R} : F_X(x) - F_X(x^-) > n^{-1}\} \).

In words, \( D_n \) is the set of all discontinuity points that have jumps greater than \( n^{-1} \).

Since the discontinuities corresponding to the points of \( D_n \) form a collection of disjoint intervals, the length of the union of all these disjoint intervals should not exceed 1 (why?). Hence, if we denote the cardinality of \( D_n \) by \( D_n^\# \), then we have \( n^{-1}D_n^\# \leq 1 \), or \( D_n^\# \leq n \). Hence, \( D \) is countable since it is a countable union of finite sets (this is a fact from elementary set theory).

Note: We could have finished the proof right after (*) by associating the points of \( D \) with a disjoint collection of intervals, each containing a rational number. In turn, we can associate the points of \( D \) with distinct rational numbers, which proves that \( D \) is countable.

**Discrete random variables:** If the distribution function of a random variable is piecewise constant, then we say that the random variable is *discrete*. Note that in this case the number of discontinuities is at most countably infinite, and the random variable may assume at most countably many values, say \( a_1, a_2, \ldots \).
Absolutely continuous random variables: If there is a Borel function $f_X : \mathbb{R} \to [0, \infty)$, with the property that \( \int_{-\infty}^{\infty} f(t) \, dt = 1 \), such that $F_X(x) = \int_{-\infty}^{x} f_X(t) \, dt$, then we say that $X$ is absolutely continuous. Note that in this event we term $f_X$ the probability density function (pdf) of $X$. Also note that if $F_X$ is differentiable, then such a density exists and it is given by the derivative of $F_X$.

Example 10. Let $X$ be a uniformly-distributed random variable in $[0,2]$. Let the function $g$ be as shown in Fig. 1 below. Compute the distribution function of $Y = g(X)$.

![Graph of $g(x)$ and $F_Y(y)$](image)

**Figure 1.**

**Solution:** If $y < 0$, $F_Y(y) = 0$ since $Y$ is nonnegative. If $0 \leq y \leq 1$, $F_Y(y) = P(\{X \leq y/2\} \cup \{X > 1 - y/2\}) = 0.5y + 0.5$. Finally, if $y > 1$, $F_Y(y) = 1$ since $Y \leq 1$. The graph of $F_Y(y)$ is shown above. Note that $F_Y(y)$ is indeed right continuous everywhere, but it is discontinuous at 0.

In this example the random variable $X$ is absolutely continuous (why?). On the other hand, the random variable $Y$ is *not* absolutely continuous because there is no Borel function $f_Y$ so that $F_Y(x) = \int_{-\infty}^{x} f_Y(t) \, dt$. Observe that $F_Y$ has a jump at 0, and we cannot reproduce this jump by integrating a Borel function over it (we would need a “delta function,” which is not really a function let alone a Borel function). It is not a discrete random variable either since $Y$ can assume any value from an uncountable collection of real numbers in $[0, 1]$. 
1.7. **Independence:** Consider a probability space \((\Omega, \mathcal{F}, P)\), and let \(\mathcal{D}_1\) and \(\mathcal{D}_2\) be two sub \(\sigma\)-algebras of \(\mathcal{F}\). We say that \(\mathcal{D}_1\) and \(\mathcal{D}_2\) are independent if \(P(A \cap B) = P(A)P(B)\) for every \(A \in \mathcal{D}_1\) and \(B \in \mathcal{D}_2\).

For example, if \(X\) and \(Y\) are random variables in \((\Omega, \mathcal{F})\), then we say that they are independent if \(\sigma(X)\) and \(\sigma(Y)\) are independent. Note that in this case we automatically have \(\mu_{XY} = \mu_X \mu_Y\), \(F_{XY}(x, y) = F_X(x)F_Y(y)\), and \(f_{XY}(x, y) = f_X(x)f_Y(y)\) if these densities exist.
2. Expectations

Recall that in an undergraduate probability course one would talk about the expectation, average, or mean of a random variable. This is done by carrying out an integration (in the Riemann sense) with respect to a probability density function (pdf). It turns out that the definition of an expectation does require having a pdf. It is based on a more-or-less intuitive notion of an average. We will follow this general approach here and then connect it to the usual expectation with respect to a pdf whenever the pdf exists. We begin by introducing the expectation of a nonnegative random variable, and will generalize thereafter.

Consider a nonnegative random variable $X$, and for each $n \geq 1$, we define the sum $S_n = \sum_{i=1}^{\infty} \frac{i}{2^n} P\{ \frac{i}{2^n} \leq X \leq \frac{i+1}{2^n} \}$. We claim that $S_n$ is nondecreasing. If this is the case (to be proven shortly), then we know that $S_n$ is either convergent (to a finite number) or $S_n \uparrow \infty$. In any case, we call the limit of $S_n$ the expectation of $X$, and symbolically we denote it as $E[X]$. Thus, we define $E[X] = \lim_{n \to \infty} S_n$. To see the monotonicity of $S_n$, we follow Chow and Teicher [1] and observe that

$$\{ \frac{i}{2^n} \leq X \leq \frac{i+1}{2^n} \} = \{ \frac{2i}{2^n+1} \leq X \leq \frac{2i+2}{2^n+1} \} = \{ \frac{2i}{2^n+1} \leq X \leq \frac{2i+1}{2^n+1} \} \cup \{ \frac{2i+1}{2^n+1} < X \leq \frac{2i+2}{2^n+1} \};$$

thus,

$$S_n = \sum_{i=1}^{\infty} \frac{2i}{2^n+1} P\{ \frac{2i}{2^n+1} < X \leq \frac{2i+1}{2^n+1} \} + P\{ \frac{2i+1}{2^n+1} < X \leq \frac{2i+2}{2^n+1} \} \leq \frac{1}{2^n+1} P\{ \frac{1}{2^n+1} < X \leq \frac{2}{2^n+1} \} + \sum_{i=1}^{\infty} \frac{2i}{2^n+1} P\{ \frac{2i}{2^n+1} < X \leq \frac{2i+1}{2^n+1} \} + P\{ \frac{2i+1}{2^n+1} < X \leq \frac{2i+2}{2^n+1} \}$$

$$\leq \frac{1}{2^n+1} P\{ \frac{1}{2^n+1} < X \leq \frac{2}{2^n+1} \} + \sum_{i=1}^{\infty} \frac{2i}{2^n+1} P\{ \frac{2i}{2^n+1} < X \leq \frac{2i+1}{2^n+1} \} + \sum_{i=1}^{\infty} \frac{2i+1}{2^n+1} P\{ \frac{2i+1}{2^n+1} < X \leq \frac{2i+2}{2^n+1} \}$$

$$= \sum_{i=1, i \text{ odd}}^{\infty} \frac{i}{2^n+1} P\{ \frac{i}{2^n+1} < X \leq \frac{i+1}{2^n+1} \} + \sum_{i=1, i \text{ even}}^{\infty} \frac{i}{2^n+1} P\{ \frac{i}{2^n+1} < X \leq \frac{i+1}{2^n+1} \}$$

$$= \sum_{i=1}^{\infty} \frac{i}{2^n+1} P\{ \frac{i}{2^n+1} < X \leq \frac{i+1}{2^n+1} \} = S_{n+1}$$

If $E[X] < \infty$, we say that $X$ is integrable.

**General Case:** If $X$ is any random variable, then we decompose it into two parts: a positive part and a negative part. More precisely, let $X^+ = \max(0, X)$ and $X^- = \max(0, -X)$, then clearly $X = X^+ - X^-$, and both of $X^+$ and $X^-$ are nonnegative. Hence, we use our definition of expectation for a nonnegative random
variable to define $E[X] = E[X^+] - E[X^-]$ whenever $E[X^+] < \infty$ or $E[X^-] < \infty$, or both. In cases where $E[X^+] < \infty$ and $E[X^-] = \infty$, or $E[X^-] < \infty$ and $E[X^+] = \infty$, we define $E[X] = -\infty$ and $E[X] = \infty$, respectively. Finally, $E[X]$ is not defined whenever $E[X^+] = E[X^-] = \infty$. Also, $E[|X|] = E[X^+ + X^-]$ exists if and only if $E[X]$ does.

**Special Case:** Consider any event $E$ and let $X = I_E$, a binary random variable. ($I_E(\omega) = 1$ if $\omega \in E$ and $I_E(\omega) = 0$ otherwise; it is called the *indicator function* of the event $E$.) Then, $E[X] = P(E)$.

Proof. Note that in this case

$$S_n = \frac{2^n - 1}{2^n} P\left\{ \frac{2^n - 1}{2^n} < X \leq 1 \right\},$$

and the second quantity is simply $P\{E\}$. Thus, $S_n \to P\{E\}$.

When $P(E) = 1$, we say that $X = 1$ with probability one, or almost surely. In general if $X$ is equal to a constant $c$ almost surely, then $E[X] = c$.

**Another Special Case:** A random variable that takes only finitely many values, that is, $X = \sum_{i=1}^{n} a_i I_{A_i}$, where $A_i$, $i = 1, \ldots, n$ are events. It is straightforward (prove it following the approach as in the previous case) that $E[X] = \sum_{i=1}^{n} a_i P(A_i)$.

**Discrete random variables:** It is easy to see that for a discrete random variable $X$, $E[X] = \sum_{i=1}^{\infty} a_i P\{X = a_i\} = \sum_{i=1}^{\infty} a_i \left( F_X(a_i) - F_X(a_i^-) \right)$. (See HW-5).

**Notation and Terminology:** $E[X]$ is also written as $\int_{\Omega} X(\omega) P(d\omega)$, which is called the Lebesgue integral of $X$ with respect to the probability measure $P$. Often, cumbersome notation is avoided by writing $\int_{\Omega} X dP$ or simply $\int_{\Omega} X dP$.

**Linearity of Expectation:** The expectation is linear, that is, $E[aX + bY] = aE[X] + bE[Y]$. This can be seen, for example, by observing that any nonnegative random variable can be approximated from below by functions of the form $\sum_{i=1}^{\infty} x_i I_{\{x_i < X \leq x_{i+1}\}}(\omega)$, where for any event $E$, the random variable $I_E(\omega) = 1$ if $\omega \in E$ and $I_E(\omega) = 0$ otherwise. (Recall that $I_E$ is called the indicator function of the set $E$.) Indeed, we have seen such an approximation through our definition of the
expectation. Namely, if we define

\[ X_n(\omega) = \sum_{i=1}^{\infty} \frac{i}{2^n} I_{\{\frac{i}{2^n} < X \leq \frac{i+1}{2^n}\}}(\omega), \]

then it is easy to check that \( X_n(\omega) \to X(\omega) \), as \( n \to \infty \). In fact, \( E[X] \) was defined as

\[ \lim_{n \to \infty} E[X_n], \]

where \( E[X_n] \) precisely coincides with \( S_n \) described above [recall that \( E[I_E(\omega)] = P(E) \)]. Now to prove the linearity of expectations, we note that if \( X \) and \( Y \) are random variables with defined expectations, then we can approximate them by \( X_n \) and \( Y_n \), respectively. Also, \( X_n + Y_n \) would approximate \( X + Y \). Next, we observe that for nonnegative \( X \) and \( Y \),

\[ E[X_n] + E[Y_n] = \sum_{i=1}^{\infty} \frac{i}{2^n} P\{\frac{i}{2^n} < X \leq \frac{i+1}{2^n}\} + \sum_{i=1}^{\infty} \frac{i}{2^n} P\{\frac{i}{2^n} < Y \leq \frac{i+1}{2^n}\} \]

\[ = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \frac{i}{2^n} P\left(\left\{\frac{i}{2^n} < X \leq \frac{i+1}{2^n}\right\} \cap \left\{\frac{j}{2^n} < Y \leq \frac{j+1}{2^n}\right\}\right) + \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \frac{i}{2^n} P\left(\left\{\frac{i}{2^n} < X \leq \frac{i+1}{2^n}\right\} \cap \left\{\frac{j}{2^n} < Y \leq \frac{j+1}{2^n}\right\}\right) \]

\[ = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \frac{i+j}{2^n} P\left(\left\{\frac{i}{2^n} < X \leq \frac{i+1}{2^n}\right\} \cap \left\{\frac{j}{2^n} < Y \leq \frac{j+1}{2^n}\right\}\right) \]

But \( \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \frac{i+j}{2^n} P\left(\left\{\frac{i}{2^n} < X \leq \frac{i+1}{2^n}\right\} \cap \left\{\frac{j}{2^n} < Y \leq \frac{j+1}{2^n}\right\}\right) = E[X_n + Y_n]. \) (Think about it.)

Thus, we have shown that \( E[X_n] + E[Y_n] = E[X_n + Y_n] \). Now take limits of both sides and use the definition of \( E[X] \), \( E[Y] \) and \( E[X + Y] \) as the limits of \( E[X_n], E[Y_n], \) and \( E[X_n + Y_n] \), respectively, to conclude that \( E[X] + E[Y] = E[X + Y] \). The homogeneity property \( E[aX] = aE[X] \) can be proved similarly.

It is easy to show that if \( X \geq 0 \) then \( E[X] \geq 0 \); in addition, if \( X \leq Y \) almost surely (this means that \( P\{X \leq Y\} = 1 \)), then \( E[X] \leq E[Y] \).

**Expectations in the Context of Distributions:** Recall that for a nonnegative random variable \( X \), \( E[X] = \lim_{n \to \infty} \sum_{i=1}^{\infty} \frac{i}{2^n} P\{\frac{i}{2^n} < X \leq \frac{i+1}{2^n}\} \). But we had seen earlier that \( P\{\frac{i}{2^n} < X \leq \frac{i+1}{2^n}\} = \mu_X\left(\left(\frac{i}{2^n}, \frac{i+1}{2^n}\right]\right) \). So we can write \( \sum_{i=1}^{\infty} \frac{i}{2^n} P\{\frac{i}{2^n} < X \leq \frac{i+1}{2^n}\} \) as \( \sum_{i=1}^{\infty} \frac{i}{2^n} \mu_X\left(\left(\frac{i}{2^n}, \frac{i+1}{2^n}\right]\right) \). We denote the limit of the latter by \( \int_{\Omega} xd\mu_X \), which is read as the Lebesgue integral of \( x \) with respect to the probability measure (or distribution) \( \mu_X \). In summary, we have \( E[X] = \int_{\Omega} XdP = \int_{\Omega} xd\mu_X \). Of course we can extend this notion to a general \( X \) in the usual way (i.e., splitting \( X \) to its positive and negative parts).
A pair of random variables Suppose that $X$ and $Y$ are rv’s defined on the measurable space $(\Omega, \mathcal{F})$. For any $B_1, B_2 \in \mathcal{B}$, we use the convenient notation $\{X \in B_1, Y \in B_2\}$ for the event $\{X \in B_1\} \cap \{Y \in B_2\}$. Moreover, we can also think of the pair $X$ and $Y$ as a vector $(X, Y)$ in $\mathbb{R}^2$.

Now consider any $B \in \mathcal{B}_2$ (the set of all Borel sets on the plane, i.e., the smallest $\sigma$-algebra containing all open sets in the plane) of the form $B_1 \times B_2$, where $B_1, B_2$ are open sets in $\mathbb{R}$ (i.e., rectangles with open sides, or simply open rectangles). We note that $\{(X, Y) \in B\}$ is always an event because $\{(X, Y) \in B\} = \{X \in B_1\} \cap \{Y \in B_2\}$. In fact, $\{(X, Y) \in B\}$ is an event for any $B \in \mathcal{B}_2$, not just those that are open rectangles. Here is the proof. Let $\mathcal{D}$ be the collection of all Borel sets in $\mathcal{B}_2$ for which $\{(X, Y) \in B\} \in \mathcal{F}$. As we had already seen, $\mathcal{D}$ contains all open rectangles (i.e., Open sets of the form $B_1 \times B_2$). Further, it is easy to show that $\mathcal{D}$ qualifies to be called a Dynkin class of sets (definition to follow below). Also, the collection $\mathcal{S}$ of all Borel sets in the plane that are open rectangles is closed under finite intersection (this is because $(B_1 \times B_2) \cap (C_1 \times C_2) = (B_1 \cap C_1) \times (B_2 \cap C_2)$). To summarize, the collection $\mathcal{D}$ is a Dynkin class and it contains the collection $\mathcal{S}$, that is closed under finite intersection and contains all open rectangles in the plane. Now by the Dynkin class Theorem (see below), $\mathcal{D}$ contains the $\sigma$-algebra generated by $\mathcal{S}$, which is just $\mathcal{B}_2$ (why?). Hence, $\{(X, Y) \in B\} \in \mathcal{F}$ for any $B \in \mathcal{B}_2$.

Thus, for any Borel set $B$ in the plane, we can define the joint distribution of $X$ and $Y$ as $\mu_{XY}(B) \triangleq P\{(X, Y) \in B\}$. This can also be generalized in the obvious way to define a joint distribution of multiple (say $n$) random variables over $\mathcal{B}_n$, the Borel subsets of $\mathbb{R}^n$.

Now back to the definition of a Dynkin class and the Dynkin class Theorem. Any collection $\mathcal{D}$ in any set $\Omega$ is called a Dynkin class if (1) $\Omega \in \mathcal{D}$, (2) $E_2 \setminus E_1 \in \mathcal{D}$ whenever $E_1 \subset E_2$ and $E_1, E_2 \in \mathcal{D}$, and (3) if $E_1 \subset E_2 \subset \ldots$, then $\bigcup_{n=1}^\infty E_n \in \mathcal{D}$.

In the same way that proved that for any collection of sets $\mathcal{M}$ there is a minimal $\sigma$-algebra containing $\mathcal{M}$, we can prove that there is a minimal Dynkin class, $\mathcal{D}_\mathcal{M}$,
that contains $\mathcal{M}$. It is easily shown that any $\sigma$-algebra is automatically a Dynkin class (prove this).

Any collection $\mathcal{S}$ in any set $\Omega$ is called a $\pi$-class if $A \cap B \in \mathcal{S}$ whenever $A, B \in \mathcal{S}$. The Dynkin Class Theorem (see Chow and Teicher, for example) states:

**Theorem 3.** If $\mathcal{S} \subset \mathcal{D}$, where $\mathcal{S}$ is a $\pi$-class and $\mathcal{D}$ is a Dynkin class, then $\mathcal{D}$ contains $\sigma(\mathcal{S})$, the $\sigma$-algebra generated by $\mathcal{S}$. As a special case, $\mathcal{D} \mathcal{S} = \sigma(\mathcal{S})$.

Before we prove the Theorem, we prove the following Lemma.

**Lemma 1.** If $\mathcal{D}$ is a Dynkin class as well as a $\pi$-class then it is automatically $\sigma$-algebra.

**Proof.** For any $E \in \mathcal{D}$, $\Omega \setminus E = E^c \in \mathcal{D}$ since $\mathcal{D}$ is a Dynkin class. Now suppose that $A_1, A_2, \ldots, \in \mathcal{D}$. Put $B_1 = A_1, B_2 = A_1 \cup A_2, \ldots, B_n = \cup_{i=1}^n A_i, \ldots$. Note that $B_n = \cap_{i=1}^n A_i^c \in \mathcal{D}$ since $A_i^c \in \mathcal{D}, i \geq 1,$ and $\mathcal{D}$ is a $\pi$-class; hence, $B_n \in \mathcal{D}, n \geq 1$. Note that $B_1 \subset B_2 \subset \ldots$, and $\bigcup_{i=1}^\infty A_i = \bigcup_{i=1}^\infty B_i \in \mathcal{D}$. Hence, $\mathcal{D}$ is a $\sigma$-algebra.

Proof of the theorem (from Chow & Teicher). We will prove $\mathcal{D} \supset \sigma(\mathcal{S})$ by showing that $\mathcal{D}$ is a $\sigma$-algebra, which can be established, in turn, by using the above Lemma if we can show that $\mathcal{D}$ is a $\pi$-class. In fact, we will show that $\mathcal{D} \mathcal{S} \mathcal{S}$ is a $\sigma$-algebra and use this fact to conclude that $\mathcal{D} \supset \sigma(\mathcal{S})$ (since $\mathcal{D} \mathcal{S} \mathcal{S} \subset \mathcal{D}$). To this end, define $\mathcal{U} = \{A \subset \Omega : A \cap S \in \mathcal{D} \mathcal{S}, \text{for every } S \in \mathcal{S}\}$. Note that $\mathcal{S} \subset \mathcal{U}$ (because of our earlier observation that for any $A \in \mathcal{D} \mathcal{S}$ and any $S \in \mathcal{S}$, $A \cap S \in \mathcal{D} \mathcal{S}$). Now define $\mathcal{U}' = \{A \subset \Omega : A \cap S \in \mathcal{D} \mathcal{S}, \text{for every } S \in \mathcal{D} \mathcal{S}\}$. Note that $\mathcal{U}' \supset \mathcal{S}$ (because of our earlier observation that for any $A \in \mathcal{D} \mathcal{S}$ and any $S \in \mathcal{S}$, $A \cap S \in \mathcal{D} \mathcal{S}$). We can also show (in a similar way to that for $\mathcal{U}$) that $\mathcal{U}'$ is a Dynkin class. Hence, $\mathcal{U}' \supset \mathcal{D} \mathcal{S}$, which implies $A \cap B \in \mathcal{D} \mathcal{S}$ whenever $A \in \mathcal{D} \mathcal{S}$ and $B \in \mathcal{D} \mathcal{S}$. 
Hence, $\mathcal{D}_S$ is a $\pi$-class, which together with the above Lemma establish the fact that $\mathcal{D}_S$ is a $\sigma$-algebra containing $S$. Hence, $\mathcal{D} \supset \mathcal{D}_S \supset \sigma(S)$.

To prove the last conclusion of the Theorem, recall that since $\sigma(S)$ is also a Dynkin class (because it is a $\sigma$-algebra) containing $S$, $\sigma(S) \supset \mathcal{D}_S$ since $\mathcal{D}_S$ is the minimal Dynkin class containing $S$. However, if we apply the above theorem to $\mathcal{D} = \mathcal{D}_S$, we conclude $\mathcal{D}_S \supset \sigma(S)$. Thus, $\mathcal{D}_S = \sigma(S)$. \[ \square \]

**Expectation in the context of distribution function:** Note that we can use the definition of a distribution function to write (for a non-negative rv $X$) $E[X] = \lim_{n \to \infty} \sum_{i=1}^{\infty} \frac{i}{2^n} \left( F_X((\frac{i+1}{2^n}) - F_X((\frac{i}{2^n})) \right)$. Now without being too fussy, we can imagine that if $F_X$ is differentiable with derivative $f_X$, then the above limit is the Reimann integral $\int_{-\infty}^{\infty} x f_X(x) \, dx$, which is the usual formula for expectation for a random variable that has a probability density function.

**Expectation of a function of a random variable.** Now that we have $E[X]$, we can define $E[g(X)]$ in the obvious way, where is $g$ is a Borel-measurable transformation from $\mathbb{R}$ to $\mathbb{R}$. (Recall that this means that $g^{-1}(B) \in \mathcal{B}$ for any $B \in \mathcal{B}$.) In particular, $E[g(X)] = \int_{\Omega} g(X) \, d\mathbb{P} = \int_{\mathbb{R}} g(x) \, d\mu_X$. Also, in the event that $X$ has a density function, we obtain the usual expression $\int_{-\infty}^{\infty} g(x) f_X(x) \, dx$.

**Markov Inequality:** If $\phi$ is a non-decreasing function on $[0, \infty)$, then $\Pr\{|X| > \epsilon\} \leq \frac{1}{\phi(\epsilon)} E[\phi(|X|)]$. In particular, if we take $\phi(t) = t^2$ and consider $X - \bar{X}$ in place of $X$, where $\bar{X} \triangleq E[X]$, then we have $\Pr\{|X - \bar{X}| > \epsilon\} \leq \frac{E[(X - \bar{X})^2]}{\epsilon^2}$. Note that the numerator is simply the variance of $X$.

Also, if we take $\phi(t) = t$, then we have $\Pr\{|X| > \epsilon\} \leq \frac{E[|X|]}{\epsilon}$.

Proof. Since $\phi$ is nondecreasing, $\{\phi(|X|) > \epsilon\} = \{\phi(|X|) > \phi(\epsilon)\}$. Further, note that $1 \geq I_{\{|X| > \epsilon\}}(\omega)$, for any $\omega \in \Omega$ (since any indicator function is either 0 or 1). Now note that $\phi(|X|) \geq I_{\{|X| > \epsilon\}}(\omega) \phi(\epsilon)$. Now take expectations of both ends of the inequalities to obtain
\[ \mathbb{E}[\phi(|X|)] \geq \phi(\epsilon)\mathbb{E}[I_{\{|X|>\epsilon\}}] = \phi(\epsilon)\mathbb{P}\{|X| > \epsilon\}, \] and the desired result follows. (See the first special case after the definition of expectation.)

**Chernoff Bound:** This is an application of the Markov inequality and it is a useful tool in upperbounding error probabilities in digital communications. Let \( X \) be a non-negative random variable and let \( \theta \) be nonnegative. Then by taking \( \phi(t) = e^{\theta t} \), we obtain

\[ \mathbb{P}\{X > x\} \leq e^{-\theta x}\mathbb{E}[e^{\theta X}]. \]

The right hand side can be minimized over \( \theta > 0 \) to yield what is called the *Chernoff bound* for a random variable \( X \).
3. **Elementary Hilbert Space Theory**

Most of the material in this section is extracted from the excellent book by W. Rudin, *Real & Complex Analysis* [2]. A complex vector space $H$ is called an inner product space if $\forall x, y \in H$, we have a complex-valued scalar, $\langle x, y \rangle$, read the “inner product between vectors $x$ and $y$,” such that the following properties are satisfied:

1. $\langle x, y \rangle = \overline{\langle y, x \rangle}$, $\forall x, y \in H$
2. $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$, $\forall x, y, z \in H$
3. $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$, $\forall x, y \in H$, $\alpha \in \mathbb{C}$
4. $\langle x, x \rangle \geq 0$, $\forall x \in H$ and $\langle x, x \rangle = 0$ only if $x = 0$

Note that (3) and (1) together imply $\langle x, \alpha y \rangle = \overline{\alpha} \langle x, y \rangle$. Also, (3) and (4) together imply $\langle x, x \rangle = 0$ if and only if $x = 0$. It would be convenient to write $\langle x, x \rangle$ as $\|x\|^2$, which we later use to introduce a norm on $H$.

3.1. **Schwarz Inequality.** It follows from properties (1)-(4) that $|\langle x, y \rangle| \leq \|x\| \|y\|$, $\forall x, y \in H$.

Proof: Put $A = \|x\|^2$, $C = \|y\|^2$ and $B = |\langle x, y \rangle|$. By (4), $\langle x - \alpha y, x - \alpha y \rangle \geq 0$, for any choice of $\alpha \in \mathbb{C}$ and all $r \in \mathbb{R}$, which can be further written as (using (2)-(3)):

$$\|x\|^2 - r\alpha < y, x > - r\overline{\alpha} < x, y > + r^2\|y\|^2 |\alpha|^2 \geq 0.$$

Now choose $\alpha$ so that $\alpha < y, x > = |\langle x, y \rangle|$. Thus, (6) can be cast as $\|x\|^2 - 2r |\langle x, y \rangle| < x, y > | + r^2\|y\|^2 \geq 0$, or $A - 2rB + r^2C \geq 0$, which is true for any $r \in \mathbb{R}$. Let $r_{1,2} = \frac{2B \pm \sqrt{B^2 - 4AC}}{2C} = \frac{B \pm \sqrt{B^2 - AC}}{C}$ denote the roots of the equation $A - 2rB + r^2C = 0$. Since $A - 2rB + r^2C \geq 0$, it must be true that $B^2 - AC \leq 0$ (since the roots cannot be real unless they are the same), which implies $|\langle x, y \rangle|^2 \leq \|x\|^2\|y\|^2$, or $|\langle x, y \rangle| \leq \|x\| \|y\|$.

Note that $|\langle x, y \rangle| = \|x\| \|y\|$ whenever $x = \beta y$, $\beta \in \mathbb{C}$. Also, if $|\langle x, y \rangle| = \|x\| \|y\|$, then it is easy to verify that $x - \frac{|\langle x, y \rangle| \|x\|}{\langle y, x \rangle \|y\|} y$, $x - \frac{|\langle x, y \rangle| \|x\|}{\langle y, x \rangle \|y\|} y = 0$, which implies (using (4)) $x - \frac{|\langle x, y \rangle| \|x\|}{\langle y, x \rangle \|y\|} y = 0$. Thus, $|\langle x, y \rangle| = \|x\| \|y\|$ if and only if $x$ is proportional to $y$. 

3.2. Triangle Inequality. \( \|x + y\| \leq \|x\| + \|y\|, \forall x, y \in H \). Proof.

This follows from the Schwarz inequality. Note that \( \|x + y\|^2 \geq 0 \) and expand to obtain \( \|x\|^2 + \|y\|^2 + 2 < x, y > \leq \|x\|^2 + \|y\|^2 + 2| < x, y > | \). Now by Schwarz inequality, \( \|x\|^2 + \|y\|^2 + 2| < x, y > | \leq \|x\|^2 + \|y\|^2 + 2\|x\||y\| = (\|x\| + \|y\|)^2 \), from which the desired result follows. Note that \( \|x + y\|^2 = \|x\|^2 + \|y\|^2 \) if \( < x, y > = 0 \) (why?).

This is a generalization of the customary triangle inequality for complex numbers, which states that \( \|x - z\| \leq \|x - y\| + \|y - z\|, x, y, z \in \mathbb{C} \).

3.3. Norm. We say \( \|x\| \) is a norm on \( H \) if:

1. \( \|x\| \geq 0 \).
2. \( \|x\| = 0 \) only if \( x = 0 \).
3. \( \|x + y\| \leq \|x\| + \|y\| \).
4. \( \|\alpha x\| = |\alpha| \|x\|, \alpha \in D \).

With the triangular inequality at hand, we can define \( \| \cdot \| \) on members of \( H \) as follows: \( \|x\| = \sqrt{<x, x>} \). You should check that this actually defines a norm. This yields a “yardstick for distance” between the two vectors \( x, y \in H \), defined as \( \|x - y\| \).

We can now say that \( H \) is normed space.

3.4. Convergence. We can then talk about convergence: a sequence \( x_n \in H \) is said to be convergent to \( x \), written as \( x_n \to x \), or \( \lim_{n \to \infty} x_n = x \), if for every \( \epsilon > 0 \), there exists \( N \in \mathbb{N} \) such that \( \|x_n - x\| < \epsilon \) whenever \( n > N \).

3.5. Completeness. An inner-product space \( H \) is complete if any Cauchy sequence in \( H \) converges to a point in \( H \). A sequence \( \{y_n\}_{n=1}^{\infty} \) in \( H \) is called a Cauchy if for any \( \epsilon > 0, \exists N \in \mathbb{N} \) such that \( \|y_n - y_m\| < \epsilon \) for all \( n, m > N \).

Now, if \( H \) is complete, then \( H \) is called a Hilbert space.

Fact: If \( \mathcal{H} \) is complete, then it is closed. This is because any convergent sequence is automatically a Cauchy sequence.

3.6. Convex Sets. A set \( E \) in a vector space \( V \) is said to be a convex set if for any \( x, y \in E \), and \( t \in (0, 1) \), the following point \( Z_t = tx + (1 - t)y \in E \). In other words,
the line segment between \(x\) and \(y\) lies in \(E\). Note that if \(E\) is a convex set, then the translation of \(E\), \(E + x \triangleq \{y + x : y \in E\}\), is also a convex set.

### 3.7. Parallelogram Law.
For any \(x\) and \(y\) in an inner-product space, \(\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2\). This can be simply verified using the properties of an inner product. See the schematic below.

![Figure 2](image)

### 3.8. Orthogonality.
If \(\langle x, y \rangle = 0\), then we say that \(x\) and \(y\) are orthogonal. We write \(x \bot y\). Note that the relation \(\bot\) is a symmetric relation; that is, \(x \bot y \Leftrightarrow y \bot x\).

Pick a vector \(x \in H\), then find all vectors in \(H\) that are orthogonal to \(x\). We write it as: \(x^\perp = \{y \in H : \langle x, y \rangle = 0\}\) and \(x^\perp\) is a closed subspace. To see that it is a subspace, we must show that \(x^\perp\) is closed under addition and scalar multiplication; these are immediate from the definition of an inner product (see Assignment 4). To see that \(x^\perp\) is closed, we must show that the limit of every convergent sequence in \(x^\perp\) is also in \(x^\perp\). Let \(x_n\) be a sequence in \(x^\perp\), and assume that \(x_n \to x_0\). We need to show that \(x_0 \in x^\perp\). To see this, note that \(|\langle x_0, x \rangle| = |\langle x_0 - x_n + x_n, x \rangle| = |\langle x_0 - x_n, x \rangle + \langle x_n, x \rangle| = |\langle x_0 - x_n, x \rangle| \leq \|x_0 - x_n\|\|x\|\), by Schwarz inequality; but, the term on the right converges to zero, so \(|\langle x_0, x \rangle| = 0\), which implies \(x_0 \perp x\) and \(x_0 \in x^\perp\).

Let \(M \subset H\) be a subspace of \(H\). We define \(M^\perp \triangleq \bigcap_{x \in M} x^\perp\), \(\forall x \in M\). It is easy to see that \(M^\perp\) is actually a subspace of \(H\) and that \(M^\perp \cap M = \{0\}\). It is also true that \(M^\perp\) is closed, which is simply because it is an intersection of closed sets (recall that \(x^\perp\) is closed).
Theorem 4. Any non-empty, closed and convex set \( E \) in a Hilbert space \( H \) contains a unique element with smallest norm. That is, there exists \( x_o \in E \) such that \( \| x_o \| < \| x \| , x \in E \).

Proof (see also Fig. 3): Existence: Let \( \delta = \inf \{ \| x \| : x \in E \} \).

In the parallelogram law, replace \( x \) by \( x/2 \) and \( y \) by \( y/2 \) to obtain
\[
\| x - y \|^2 = 2 \| x \|^2 + 2 \| y \|^2 - 4 \| \frac{x+y}{2} \|^2.
\]
Since \( \frac{x+y}{2} \in E \) (because \( E \) is convex), \( \| \frac{x+y}{2} \|^2 \geq \delta^2 \). Hence,
\[
(7) \quad \| x - y \|^2 \leq 2 \| x \|^2 + 2 \| y \|^2 - 4 \delta^2.
\]

By definition of \( \delta \), there exists a sequence \( \{ y_n \}_{n=1}^{\infty} \in H \) such that \( \lim_{n \to \infty} \| y_n \| = \delta \).

Now replace \( x \) by \( y_n \) and \( y \) by \( y_m \) in (7) to obtain
\[
(8) \quad \| y_n - y_m \|^2 \leq 2 \| y_n \|^2 + 2 \| y_m \|^2 - 4 \delta^2.
\]

Hence,
\[
\lim_{n,m \to \infty} \| y_n - y_m \|^2 \leq 2 \lim_{n \to \infty} \| y_n \|^2 + 2 \lim_{m \to \infty} \| y_m \|^2 - 4 \delta^2 = 2 \delta^2 + 2 \delta^2 - 4 \delta^2 = 0
\]
and we conclude that \( \{ y_n \}_{n=1}^{\infty} \) is a Cauchy sequence in \( H \). Since \( H \) is complete, there exists \( x_o \in H \) such that \( y_n \to x_o \). We next show that \( \| x_o \| = \delta \), which would complete the existence of a minimal-norm member in \( E \).

Note that by the triangle inequality,
\[
\left| \| x_o \| - \| y_n \| \right| \leq \| x_o - y_n \|.
\]

Now take limits of both sides to conclude that \( \lim_{n \to \infty} \| y_n \| = \| x_o \| \). But we already know that \( \lim_{n \to \infty} \| y_n \| = \delta \), which implies that \( \| x_o \| = \delta \).

Uniqueness: suppose that \( x' \neq x, x, x' \in E \), and \( \| x \| = \| x' \| = \delta \). Replace \( y \) in (7) by \( x' \) to obtain
\[
\| x - x' \|^2 \leq 2 \delta^2 + 2 \delta^2 - 4 \delta^2 = 0,
\]
which implies that \( x = x' \). Hence the minimal-norm element in \( E \) is unique. \( \square \)

Example 11. For any fixed \( n \), the set \( \mathbb{C}^n \) of all \( n \)-tuples of complex numbers, \( x = (x_1, x_2, \ldots, x_n), x_i \in \mathbb{C} \), is a Hilbert Space, where with \( y = (y_1, y_2, \ldots, y_n) \) we define
\[
\langle x, y \rangle \triangleq \sum_{i=1}^{n} x_i \overline{y}_i.
\]
Example 12. \( L_2[a, b] = \left\{ f : \int_a^b |f(x)|^2 dx < \infty \right\} \) is a Hilbert space, with \( \langle f, g \rangle \triangleq \int_a^b f(x)\bar{g}(x)dx \). \( \|f\| = \sqrt{\langle f, f \rangle} = [\int_a^b |f|^2 dx]^{1/2} \triangleq \|f\|_2 \). Actually, we need Schwarz inequality for expectations before we can see that we have an inner product in this case. More on this to follow.

Theorem 5. Projection Theorem. If \( M \subset H \) is a closed subspace of a Hilbert space \( H \), then for every \( x \in H \) there exists a unique decomposition \( x = Px + Qx \), where \( Px \in M \), and \( Qx \in M^\perp \). Moreover,

1. \( \|x\|^2 = \|Px\|^2 + \|Qx\|^2 \).
2. If we think of \( Px \) and \( Qx \) as mappings from \( H \) to \( M \) and \( M^\perp \), respectively, then the mappings \( P \) and \( Q \) are linear.
(3) \( Px \) is the nearest point in \( M \) to \( x \), and \( Qx \) is the nearest point in \( M^\perp \) to \( x \). \( Px \) and \( Qx \) are called the orthogonal projections of \( x \) into \( M \) and \( M^\perp \), respectively.

Proof.

Existence: Consider \( M + x \); we claim that \( M + x \) is closed. Recall that \( M \) is closed, which means that if \( x_n \in M \) and \( x_n \to x_o \), then \( x_o \in M \). Pick a convergent sequence in \( x + M \), call it \( z_n \). Now \( z_n = x + y_n \), for some \( y_n \in M \). Since \( z_n \) is convergent, so is \( y_n \), but the limit of \( y_n \) is in \( M \), so \( x + \lim_{n \to \infty} y_n \in x + M \).

We next show that \( x + M \) is convex. Pick \( x_1 \) and \( x_2 \in x + M \). We need to show that for any \( 0 < \alpha < 1 \), \( \alpha x_1 + (1 - \alpha)x_2 \in x + M \). But \( x_1 = x + y_1 \), \( y_1 \in M \), and \( x_2 = x + y_2 \), \( y_2 \in M \). So \( \alpha x_1 + (1 - \alpha)x_2 = x + \alpha y_1 + (1 - \alpha)y_2 \in x + M \) since \( y_1 + (1 - \alpha)y_2 \in M \).

By the minimal-norm theorem, there exists a member in \( x + M \) of smallest norm. Call it \( Qx \). Let \( Px = x - Qx \). Note that \( Px \in M \). We need to show that \( Qx \in M^\perp \).

Namely, \( < Qx, y > = 0 \), \( \forall y \in M \). Call \( Qx = z \), and note that \( \|z\| \leq \|\tilde{y}\| \), \( \forall \tilde{y} \in M + x \).

Pick \( \tilde{y} = z - \alpha y \), where \( y \in M \), \( \|y\| = 1 \). \( \|z\|^2 \leq \|z - \alpha y\|^2 = < z - \alpha y, z - \alpha y > \), or \( 0 \leq -\alpha < y, z > -\tilde{\alpha} < z, y > + \|\alpha\|^2 \). Pick \( \alpha = < z, y > \). We obtain \( 0 \leq -| < z, y > |^2 \). This can hold only if \( < z, y > = 0 \), i.e., \( z \) is orthogonal to every \( y \in M \); therefore, \( Qx \in M^\perp \).

Uniqueness: Suppose that \( x = Px + Qx = (Px)' + (Qx)' \), where \( Px, (Px)' \in M \) and \( Qx, (Qx)' \in M^\perp \). Then, \( Px - (Px)' = (Qx)' - Qx \), where the left side belongs to \( M \) while the right side belongs to \( M^\perp \). Hence, each side can only be the zero vector (why?), and we conclude that \( Px = (Px)' \) and \( (Qx)' = Qx \).

Minimum Distance Properties: To show that \( Px \) is the nearest point in \( M \) to \( x \), pick any \( y \in M \) and observe that \( \|x - y\|^2 = \|Px + Qx - y\|^2 = \|Qx + (Px - y)\|^2 = \|Qx\|^2 + \|Px - y\|^2 \) (since \( Px - y \in M \)). The right-hand side is minimized when \( \|Px - y\| = 0 \), which happens if and only if \( y = Px \). The fact that \( Qx \) is the nearest point in \( M^\perp \) to \( x \) can be shown similarly.

Linearity: Take \( x, y \in H \). Then, we have \( x = Px + Qx \) and \( y = Py + Qy \). Now \( ax + by = aPx + aQx + bPy + bQy \). On the other hand, \( ax + by = P(ax + by) + Q(ax + by) \).
Thus, we can write $P(ax + by) - aPx - bPy = -Q(ax + by) + aQx + bQy$ and observe that the left side is in $M$ while the right side is in $M^\perp$. Hence, each side can only be the zero vector. Therefore, $P(ax + by) = aPx + bPy$ and $Q(ax + by) = aQx + bQy$. \qed
4. Conditional expectations for \( L_2 \) random variables

4.1. Holder inequality for expectations. Consider the probability space \((\Omega, \mathcal{F}, P)\), and let \(X\) and \(Y\) be random variables. The next result is called Holder inequality for expectations.

**Theorem 6.** If \( p > 1, q > 1 \) for which \( p^{-1} + q^{-1} = 1 \), then \( \mathbb{E}[|XY|] \leq \mathbb{E}[|X|^p]^{1/p} \mathbb{E}[|Y|^q]^{1/q} \).

Proof: (See for example [1] p. 105).

Note that if \( \mathbb{E}[|X|^p]^{1/p} = 0 \), then \( X = 0 \) a.s., and hence \( \mathbb{E}[|XY|] = 0 \) and (6) holds.

If \( \mathbb{E}[|X|^p]^{1/p} > 0 \) and \( \mathbb{E}[|Y|^q]^{1/q} = \infty \), then (6) also holds trivially. Next, consider the case when \( 0 < \mathbb{E}[|X|^p]^{1/p} < \infty \) and \( 0 < \mathbb{E}[|Y|^q]^{1/q} < \infty \). Let \( U = |X|/\mathbb{E}[|X|^p]^{1/p} \) and \( V = |Y|/\mathbb{E}[|Y|^q]^{1/q} \), and note that \( \mathbb{E}[|U|^p] = \mathbb{E}[|V|^q] = 1 \). Using the convexity of the logarithm function, it is easy to see that for any \( a, b > 0 \),

\[
\log\left(\frac{a^p}{p} + \frac{b^q}{q}\right) \geq \frac{1}{p} \log a^p + \frac{1}{q} \log b^q,
\]

and the last term is simply \( \log ab \). From this it follows that

\[
ab \leq \frac{a^p}{p} + \frac{b^q}{q}.
\]

Thus,

\[
\mathbb{E}[UV] \leq \frac{1}{p} \mathbb{E}[U^p] + \frac{1}{q} \mathbb{E}[V^q] = \frac{1}{p} + \frac{1}{q} = 1,
\]

from which the desired follows. \(\square\)

When \( p = q = 2 \), we have \( \mathbb{E}[|XY|] \leq \mathbb{E}[X^2]^{1/2} \mathbb{E}[Y^2]^{1/2} \), and this result is called Schwarz inequality for expectations.

Let \( H \) be the collection of square-integrable random variables. Now \( \mathbb{E}[(aX+bY)^2] = a^2 \mathbb{E}[X^2] + b^2 \mathbb{E}[Y^2] + 2ab \mathbb{E}[XY] \leq a^2 \mathbb{E}[X^2] + b^2 \mathbb{E}[Y^2] + 2ab \mathbb{E}[X^2]^{1/2} \mathbb{E}[Y^2]^{1/2} < \infty \), where we have used Schwarz inequality for expectations in the last step. Hence, \( H \) is a vector space. Next, Schwarz inequality for expectations also tells us that if \( X \) and \( Y \) are square integrable, then \( \mathbb{E}[XY] \) is defined (i.e., finite). We can therefore define \( \langle X, Y \rangle \triangleq \mathbb{E}[XY] \) and see that it defines an inner product. (Technically we need to consider the equivalence class of random variables that are equal a.s.; otherwise,
we may not have an inner product. Can you see why?) We can also define the norm \( \|X\|_2 \triangleq \langle X, X \rangle^{1/2} \). This is called the \( L_2 \)-norm associated with \((\Omega, \mathcal{F}, \mathbb{P})\). Hence, we can recast \( H \) as the vector space of all random variables that have finite \( L_2 \)-norm. This collection is often written as \( L_2(\Omega, \mathcal{F}, \mathbb{P}) \). It can be shown that \( L_2(\Omega, \mathcal{F}, \mathbb{P}) \) is complete (see [2], pp. 67). Hence, \( L_2(\Omega, \mathcal{D}, \mathbb{P}) \) is a Hilbert space. In fact, \( L_2(\Omega, \mathcal{D}, \mathbb{P}) \) is a Hilbert space for any \( \sigma \)-algebra \( \mathcal{D} \).

4.2. The \( L_2 \) conditional expectation. Let \( X, Y \in L_2(\Omega, \mathcal{F}, \mathbb{P}) \). Clearly, \( L_2(\Omega, \sigma(X), \mathbb{P}) \) is a closed (why?) subspace of \( L_2(\Omega, \mathcal{F}, \mathbb{P}) \). We can now apply the projection theorem to \( Y \), with \( L_2(\Omega, \sigma(X), \mathbb{P}) \) being our closed subspace, and obtain the decomposition \( Y = PY + QY \), where \( PY \in L_2(\Omega, \sigma(X), \mathbb{P}) \) and \( QY \in L_2(\Omega, \sigma(X), \mathbb{P})^\perp \). We also have the property that \( \|Y - PY\|_2 \leq \|Y - Y'\|_2 \) \( \forall \) \( Y' \in L_2(\Omega, \sigma(X), \mathbb{P}) \). We call \( PY \) the conditional expectation of \( Y \) given \( \sigma(X) \), which we will write from this point on as \( \mathbb{E}[Y|\sigma(X)] \).

Exercise 1. (1) Show that if \( X \) is a \( \mathcal{D} \)-measurable r.v. then so is \( aX, a \in \mathbb{R} \). (2) Show that if \( X \) and \( Y \) are \( \mathcal{D} \)-measurable, so is \( X + Y \). [See Homework]

The above exercise shows that the collection of square-integrable \( \mathcal{D} \)-measurable random variables is indeed a vector space.

Theorem 7. Consider \((\Omega, \mathcal{F}, \mathbb{P})\), and let \( X \) be a random variable. A random variable \( Z \) is a \( \sigma(X) \)-measurable random variable if and only if \( Z = h(X) \) for some Borel function \( h \). \( \Box \) (The proof is based on a corollary to the Dynkin class Theorem. We omit the proof, which can be found in [1].)

Since \( \mathbb{E}[X|\sigma(Y)] \) is \( \sigma(Y) \)-measurable by definition, we can write it explicitly as a Borel function of \( Y \). For this reason, we often time write \( \mathbb{E}[X|Y] \) in place of \( \mathbb{E}[X|\sigma(Y)] \).

4.3. Properties of Conditional Expectation. Our definition of conditional expectation lends itself to many powerful properties that are useful in practice. One of the properties also leads to an equivalent definition of the conditional expectation, which is actually the way commonly done—see the comments after Property 4.3.3.
However, proving the existence of the conditional expectation according to the alternative definition requires knowledge of a major theorem in measure theory called the Radon-Nikodym Theorem, which is beyond the scope of this course. That is why in this course we took a path (for defining the conditional expectation) that is based on the projection theorem, which is an important theorem to signal processing for other reasons as well. Here are some key properties of conditional expectations.

4.3.1. Property 1. For any constant $a$, $\mathbb{E}[a|\mathcal{D}] = a$. This follows trivially from the observation that $Pa = a$ (why?).

4.3.2. Property 2 (Linearity). For any constants $a, b$, $\mathbb{E}[aX + bY|\mathcal{D}] = a\mathbb{E}[X|\mathcal{D}] + b\mathbb{E}[Y|\mathcal{D}]$. This follows trivially from the linearity of the projection operator $P$ (see the Projection Theorem).

4.3.3. Property 3. Let $Z = \mathbb{E}[X|\mathcal{D}]$, then $\mathbb{E}[XY] = \mathbb{E}[ZY]$, for all $Y \in L_2(\Omega, \mathcal{D}, \mathbb{P})$. Interpretation: $Z$ contains all the information that $X$ contains that is relevant to any $\mathcal{D}$-measurable random variable $Y$.

Proof: Note that $\mathbb{E}[XY] = \mathbb{E}[(PX + QX)Y] = \mathbb{E}[(PX)Y] + \mathbb{E}[(QX)Y] = \mathbb{E}[(PX)Y] = \mathbb{E}[ZY]$. The last equality follows from the definition of $Z$.

Conversely, if a random variable $Z \in L_2(\Omega, \mathcal{D}, \mathbb{P})$ has the property that $\mathbb{E}[ZY] = \mathbb{E}[XY], \forall Y \in L_2(\Omega, \mathcal{D}, \mathbb{P})$, then $Z = \mathbb{E}[X|\mathcal{D}]$. To see this, we need to show that $Z = PX$ almost surely. Note that by assumption $\mathbb{E}[Y(X - Z)] = 0$ for any $Y \in L_2(\Omega, \mathcal{D}, \mathbb{P})$. Therefore, $\mathbb{E}[Y(PX + QX - Z)] = 0$, or $\mathbb{E}[YPX] + \mathbb{E}[YQX] - \mathbb{E}[YZ] = 0$, or $\mathbb{E}[Y(Z - PX)] = 0, \forall Y \in L_2(\Omega, \mathcal{D}, \mathbb{P})$ (since $\mathbb{E}[YQX] = 0$). In particular, if we take $Y = Z - PX$ we will conclude that $\mathbb{E}[(Z - PX)^2] = 0$, which implies $Z = PX$ almost surely.

Thus, we arrive at an alternative definition for the $L_2$ conditional expectation.
**Definition 6.** We define $Z \triangleq \mathbb{E}[X|\mathcal{D}]$ if (1) $Z \in L_2(\Omega, \mathcal{D}, \mathbb{P})$ and (2) $\mathbb{E}[ZY] = \mathbb{E}[XY]$ for all $Y \in L_2(\Omega, \mathcal{D}, \mathbb{P})$.

We will use this new definition frequently in the remainder of this chapter.

4.3.4. **Property 4 (Smoothing Property).** For any $X$, $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|\mathcal{D}]]$. To see this, note that if $Z = \mathbb{E}[X|\mathcal{D}]$, then $\mathbb{E}[ZY] = \mathbb{E}[XY]$ for all $Y \in L_2(\Omega, \mathcal{D}, \mathbb{P})$. Now take $Y = 1$ to conclude that $\mathbb{E}[X] = \mathbb{E}[Z]$.

4.3.5. **Property 5.** If $Y \in L_2(\Omega, \mathcal{D}, \mathbb{P})$, then $\mathbb{E}[XY|\mathcal{D}] = YE[X|\mathcal{D}]$. To show this, we check the second definition for conditional expectations. Note that $Y \mathbb{E}[X|\mathcal{D}] \in L_2(\Omega, \mathcal{D}, \mathbb{P})$, so all we need to show is that $\mathbb{E}[(XY)W] = \mathbb{E}[(Y\mathbb{E}[X|\mathcal{D}])W]$ for any $W \in L_2(\Omega, \mathcal{D}, \mathbb{P})$. But we already know from the definition of $\mathbb{E}[X|\mathcal{D}]$ that $\mathbb{E}[(WY)\mathbb{E}[X|\mathcal{D}]] = \mathbb{E}[(WY)X]$, since $WY \in L_2(\Omega, \mathcal{D}, \mathbb{P})$.

As a special case, we have $\mathbb{E}[XY] = \mathbb{E}[\mathbb{E}[XY]|Y] = \mathbb{E}[YE[X|Y]]$.

4.3.6. **Property 6.** For any Borel function $g : \mathbb{R} \to \mathbb{R}$, $\mathbb{E}[g(Y)|\mathcal{D}] = g(Y)$ whenever $Y \in L_2(\Omega, \mathcal{D}, \mathbb{P})$. (For example, we have $\mathbb{E}[g(Y)|Y] = g(Y)$.) To prove this, all we need to do is to observe that $g(Y) \in L_2(\Omega, \mathcal{D}, \mathbb{P})$ and then apply Property 4.3.5 with $X = 1$.

4.3.7. **Property 7 (Iterated Conditioning).** Let $\mathcal{D}_1 \subset \mathcal{D}_2 \subset \mathcal{F}$. Then, $\mathbb{E}[X|\mathcal{D}_1] = \mathbb{E}[\mathbb{E}[X|\mathcal{D}_2]|\mathcal{D}_1]$.

Proof: Let $Z = \mathbb{E}[\mathbb{E}[X|\mathcal{D}_2]|\mathcal{D}_1]$. First note that $Z \in L_2(\Omega, \mathcal{D}_1, \mathbb{P})$, so we only need to show that $\mathbb{E}[ZY] = \mathbb{E}[XY]$, for all $Y \in L_2(\Omega, \mathcal{D}_1, \mathbb{P})$. Now $\mathbb{E}[ZY] = \mathbb{E}[YE[\mathbb{E}[X|\mathcal{D}_2]|\mathcal{D}_1]]$. Since $Y$ is $\mathcal{D}_1$-measurable, it is also $\mathcal{D}_2$-measurable; thus, we have $\mathbb{E}[YE[\mathbb{E}[X|\mathcal{D}_2]|\mathcal{D}_1]] = \mathbb{E}[E[YE[X|\mathcal{D}_2]|\mathcal{D}_1]]$ for all $Y \in L_2(\Omega, \mathcal{D}_1, \mathbb{P})$. But by property 4.3.4, $\mathbb{E}[E[YE[X|\mathcal{D}_2]|\mathcal{D}_1]] = \mathbb{E}[(E[YX|\mathcal{D}_2]|[\mathcal{D}_1]) = \mathbb{E}[YX]$, which completes the proof.
Also note that $E[X|\mathcal{D}_1] = E[E[X|\mathcal{D}_1]|\mathcal{D}_2]$. This is much easier to prove. (Prove it.)

The next property is very important in applications.

4.3.8. Property 8. If $X$ and $Y$ are independent r.v.’s, and if $g : \mathbb{R}^2 \to \mathbb{R}$ is Borel measurable, then $E[g(X,Y)|\sigma(Y)] = h(Y)$, where for any scalar $t$, $h(t) \triangleq E[g(X,t)]$. As a consequence, $E[g(X,Y)] = E[h(Y)]$.

Proof: We prove this in the case $g(x,y) = g_1(x)g_2(y)$, where $g_1$ and $g_2$ are bounded, real-valued Borel functions. (The general result follows through an application of a corollary to the Dynkin class theorem.) First note that $E[g_1(X)g_2(Y)|\sigma(Y)] = g_2(Y)E[g_1(X)|\sigma(Y)]$. Also note that $h(t) = g_2(t)E[g_1(X)]$, so $h(Y) = g_2(Y)E[g_1(X)]$.

We need to show that $h(Y) = E[g_1(X)g_2(Y) | \sigma(Y)]$. To do so, we need to show that $E[h(Y)Z] = E[g_1(X)g_2(Y)Z]$ for every $Z \in L_2(\Omega, \sigma(Y), \mathbb{P})$. But $E[h(Y)Z] = E[g_2(Y)E[g_1(X)]Z] = E[g_1(X)]E[g_2(Y)Z]$, and by the independence of $X$ and $Y$ the latter is equal to $E[g_1(X)g_2(Y)Z]$.

The stated consequence follows from Property 4.3.4. Also, the above result can be easily generalized to random vectors.

4.4. Some applications of Conditional Expectations.

Example 13 (Photon counting).

![Photon counting diagram](image)

**Figure 5.**
It is known that the energy of a photon is \( h\nu \), where \( h \) is Planck’s constant and \( \nu \) is the optical frequency (color) of the photon in Hertz. We are interested in detecting and counting the number of photons in a certain time interval. We assume that we have a detector, and upon the detection of the \( i \)th photon, it generates an instantaneous pulse (a delta function) whose area is random non-negative integer, \( G_i \). We then integrate the sum of all delta functions (i.e., responses) in a given interval and call that the photon count \( M \). We will also assume that the \( G_i \)'s are mutually independent, and they are also independent of the number of photons \( N \) that impinge on the detector in the given time interval. Mathematically, we can write

\[
M = \sum_{i=1}^{N} G_i.
\]

Let us try to calculate \( E[M] \), the average number of detected photons in a given time interval. Assume that we know \( P\{M = k\} (k = 0, 1, 2, 3\ldots) \) and that \( P\{G_i = k\} \) is also known \( (k = 0, 1, 2, 3\ldots) \).

By using Property 4.3.4, we know that \( E[M] = E[E[M|N]] \). To find \( E[M|N] \), we must appeal to Property 4.3.8. We think of the random sequence \( \{G_i\}_{i=1}^{\infty} \) as \( X \), and the random variable \( N \) as \( Y \). With these, the function \( h \) becomes

\[
h(t) = E[\sum_{i=1}^{t} G_i] = \sum_{i=1}^{t} E[G_i].
\]

Let us assume further that the \( G_i \)'s are identically distributed, so that \( E[G_i] \equiv E[G_1] \). With this, \( h(t) = tE[G_1] \) and \( h(N) = E[G_1]N \). Therefore, \( E[M] = E[h(N)] = E[N]E[G_1] \).

What is the variance of \( M \)? Let us first calculate \( E[M^2] \). By Property 4.3.4, \( E[M^2] = E[E[M^2|N]] \), and by Property 4.3.8, \( E[M^2|N] \) is equal to \( h_2(N) \), where

\[
h_2(t) \triangleq E\left[\left(\sum_{i=1}^{t} G_i\right)^2\right] = E\left[\sum_{i=1}^{t} G_i G_j\right] = E\left[\sum_{i=1}^{t} \sum_{j=1}^{t} G_i G_j\right] + E\left[\sum_{i=1}^{t} \sum_{j=1}^{t} G_i^2\right] = (t^2 - t)E[G_1]^2 + t(E[G_1]^2 + \sigma_G^2)
\]

and \( \sigma_G^2 \triangleq E[G_1^2] - E[G_1]^2 \) is the variance of \( G_1 \), which is common to all the other \( G_i \)'s due to independence.
Next, \( E[M^2] = E[h_2(N)] = E[(N^2 - N)]E[G_1]^2 + (E[G_1]^2 + \sigma_G^2)E[N] = E[G_1]^2(\sigma_N^2 + \bar{N}^2 - \bar{N}) + (E[G_1]^2 + \sigma_G^2)\bar{N} \), where \( \bar{N} \triangleq E[N] \) and \( \sigma_N^2 \triangleq E[N^2] - \bar{N}^2 \) is the variance of \( N \). Finally, if we assume that \( N \) is a Poisson random variable (as in coherent light) the variance of \( M \) becomes

\[
\sigma_M^2 = E[G_1]^2\bar{N}^2 + (E[G_1]^2 + \sigma_G^2)\bar{N}.
\]

**Exercise:** For an integer-valued random variable \( X \), we define the *generating function* of \( X \) as \( \psi_X(s) \triangleq E[s^X], s \in \mathbb{C}, |s| \leq 1 \). We can think of the generating function as the \( z \)-transform of the probability mass function associated with \( X \).

For the photon-counting example described above, show that \( \psi_M(s) = \psi_N(\psi_G(s)) \).

**Example 14** (First Occurrence of Successive Hits).

Consider the problem of flipping a coin successively. Suppose \( \mathbb{P}\{H\} = p \) and \( \mathbb{P}\{T\} = q = 1 - p \). In this case, we define \( \Omega \triangleq \bigotimes_{i=1}^{\infty} \Omega_i \), where for each \( i \), \( \Omega_i = \{H, T\} \).

This means that members of \( \Omega \) are simply of the form \( \omega = (\omega_1, \omega_2, \ldots) \), where \( \omega_i \in \Omega_i \). For each \( \Omega_i \), we define \( \mathcal{F}_i = \{\emptyset, \{H, T\}, \{T\}, \{H\}\} \). We take \( \mathcal{F} \) associated with \( \Omega \) as the minimal \( \sigma \)-algebra containing all cylinders with measurable bases in \( \Omega \). A an event \( E \) in \( \bigotimes_{i=1}^{\infty} \Omega_i \) is called a *cylinder with a measurable base* if \( E \) is of the form \( \{(\omega_1, \omega_2, \ldots) \in \Omega : \omega_{i_1} \in B_1, \ldots, \omega_{i_k} \in B_k, k \text{ finite}, B_i \in \mathcal{F}_i \} \). In words, in a cylinder, only finitely many of the coordinates are specified. (For example, think of a vertical cylinder in \( \mathbb{R}^3 \), where we specify only two coordinates out of three, now extend this to \( \mathbb{R}^\infty \).) Let \( X_i = I_{(\omega_i = H)} \) be a \( \{0, 1\} \)-valued random variable, and define \( X \) on \( \bigotimes_{i=1}^{\infty} \Omega_i \) as \( X \triangleq (X_1, X_2, \ldots) \). Finally, we define \( \mathbb{P} \) on cylinders in \( \Omega \) by forming the product of the probabilities of the coordinates (enforcing independence). For each \( (\Omega_i, \mathcal{F}_i) \), define \( \mathbb{P}_i\{H\} = p \) and \( q = 1 - p \).

We would like to define a random variable that tells us when a run of \( n \) successive heads appears for the first time. More precisely, define the random variable \( T_1 \) as a function of \( X \) as follows: \( T_1 \triangleq \min\{i \geq 1 : X_i = 1\} \). For example, if \( X = (0, 0, 1, 0, 1, 0, 1, 1, 1, \ldots) \), then \( T_1 = 3 \). More generally, we define \( T_k = \min\{i : X_{i-k+1} = X_{i-k+2} = \ldots = X_i = 1\} \). For each \( k \), \( T_k \) is a r.v. on \( \Omega \). For example, if \( X = (0, 0, 1, 0, 1, 1, 0, 1, 1, 1, \ldots) \), then \( T_2 = 6 \), and \( T_3 = 10 \).
Let \( y_k = E[T_k] \). In what follows we will characterize \( y_k \).

**Special case: \( k=1 \)**

It is easy to see that 
\[
P\{T_1 = 1\} = p,
\]
\[
P\{T_1 = 2\} = qp,
\]
\[
P\{T_1 = 3\} = q^2 p,
\]

\[\vdots\]
\[
P\{T_1 = n\} = q^{n-1} p.
\]

Recall from undergraduate probability that above probability law is called the geometric law, and \( T_1 \) is called a geometric random variable. Also, 
\[
y_1 = \sum_{i=1}^{\infty} ipq^{i-1} = \frac{1}{p}.
\]

**General case: \( k > 1 \)**

We begin by observing that \( T_k \) is actually an explicit function, \( f \), say, of 
\( T_{k-1}, X_{T_{k-1}+1}, X_{T_{k-1}+2}, \ldots \). Note, however, that \( T_{k-1} \) and \( \{X_{T_{k-1}+1}, X_{T_{k-1}+2}, \ldots \} \) are independent. Therefore, by Property 4.3.8, 
\[
E[T_k] = E[f(T_{k-1}, X_{T_{k-1}+1}, X_{T_{k-1}+2}, \ldots)] = E[h(T_{k-1})],
\]
where \( h(t) = E[f(t, X_{t+1}, X_{t+2}, \ldots)] \).

It is easy to check (using the equivalent definition of a conditional expectation) that 
\[
E[f(t, X_{t+1}, \ldots)] = E[f(t, X_{t+1}, \ldots) | X_{t+1}] = (t + 1)I_{\{X_{t+1}=1\}} + (t + 1 + y_k)I_{\{X_{t+1}=0\}}.
\]

This essentially says that if it took us \( t \) flips to see \( k-1 \) consecutive heads for the first time, then if the \((t + 1)\)st flip is a head, then we have achieved \( k \) successive heads at time \( t + 1 \); alternatively, if the \((t + 1)\)st flip is a tail, then we have to start all over again (start afresh) while we have already waisted \( t + 1 \) units of time.

Now, 
\[
E[f(t, X_{t+1}, \ldots)] = E[E[f(t, X_{t+1}, \ldots) | X_{t+1}]] = E[(t + 1)I_{\{X_{t+1}=1\}} + (t + 1 + y_k)I_{\{X_{t+1}=0\}}]
\]
\[= (t + 1)p + (t + 1 + y_k)q. \]

Thus, 
\[
h(t) = (t + 1)p + (t + 1 + y_k)q \quad \text{and} \quad y_k = E[h(T_{k-1})] = E[(T_{k-1} + 1)p + (T_{k-1} + 1 + y_k)q] = p + py_{k-1} + qy_k + qy_{k-1} + q.
\]

Finally, we obtain 
\[
y_k = p^{-1}y_{k-1} + p^{-1}.
\]

We now invoke the initial condition \( y_1 = \frac{1}{p} \), which completes the characterization of \( y_k \).

For example, if \( p = \frac{1}{2} \), then \( y_2 = 2y_1 + 2 = 6 \), and so on.
Example 15. Importance of the Independence Hypothesis in Property 4.3.8.

Let $Y = XZ$ where $Z = 1 + X$. Now, $E[Y] = E[XZ] = E[X(1 + X)] = \bar{X} + \bar{X}^2$. However, if we erroneously attempt to use Property 4.3.8 (by ignoring the dependence of $Z$ on $X$), we will have $E[X^t] = E[X]^t$, and $E[E[X]Z] = E[X]E[Z] = E[X](1 + E[X]) = \bar{X} + \bar{X}^2$. Note that the two are different since $\bar{X}^2 \neq \bar{X}^2$ in general.

4.5. The $L_1$ conditional expectation. Consider a probability space $(\Omega, \mathcal{F}, P)$ and let $X$ be an integrable random variable, that is, $E[|X|] < \infty$. We write $X \in L_1(\Omega, \mathcal{F}, P)$. Let $\mathcal{D}$ be a sub $\sigma$-algebra. For each $M \geq 1$, we define $X \wedge M$ as $X$ when $|X| \leq M$ and $M$ otherwise. Note that $X \wedge M \in L_2(\Omega, \mathcal{F}, P)$ (since it is bounded), so we can talk about $E[(X \wedge M)|\mathcal{D}]$, which we call $Z_M$. The question is can we somehow use $Z_M$ to define $E[X|\mathcal{D}]$? Intuitively, we would want to think of $E[X|\mathcal{D}]$ as some kind of limit of $Z_M$, as $M \to \infty$. So one approach for the construction of $E[X|\mathcal{D}]$ is to take a “limit” of $Z_M$ and then show that the limit, $Z$, say, satisfies Definition 6 with $L_2$ replaced by $L_1$. It turns out that such an approach would lead to the following definition, which we will adopt from now on.

Definition 7. [L_1 Conditional Expectation] If $X$ is an integrable random variable, then we define $Z = E[X|\mathcal{D}]$ if $Z$ has the following two properties: (1) $Z$ is $\mathcal{D}$ measurable, (2) $E[ZY] = E[XY]$ for any bounded $\mathcal{D}$ measurable random variable $Y$.

All the properties of the $L_2$ conditional expectation will carry on to the $L_1$ conditional expectation.

We make the final remark that if a random variable $X$ is square integrable, then it is integrable. The proof is easy. Suppose that $X$ is square integrable. Then, $E[|X|] = E[|X|I_{\{|X|\leq 1\}}] + E[|X|I_{\{|X|>1\}}] \leq E[I_{\{|X|\leq 1\}}] + E[X^2I_{\{|X|>1\}}] \leq 1 + E[X^2] < \infty$.

4.6. Conditional probabilities. Now we will connect conditional expectations to the familiar conditional probabilities. In particular, we recall from undergraduate probability that if $A$ and $B$ are events, with $P(B) > 0$, then we define the conditional probability of $A$ given that the event $B$ has occurred as $P(A|B) \triangleq P(A \cap B)/P(B)$.
What is the connection between this definition and our notion of a conditional expectation?

Consider the $\sigma$-algebra $\mathcal{D} = \{\emptyset, \Omega, B, B^c\}$, and consider $E[I_A|\mathcal{D}]$. Because of the special form of this $\mathcal{D}$, and because we know that this conditional expectation is $\mathcal{D}$-measurable, we infer that $E[I_A|\mathcal{D}]$ can assume only two values: one value on $B$ and another value on $B^c$. That is, we can write $E[I_A|\mathcal{D}] = aI_B + bI_{B^c}$, where $a$ and $b$ are constants. We claim that $a = P(A \cap B)/P(B)$ and $b = P(A \cap B^c)/P(B^c)$. Note that because $P(B) > 0$, $1 - P(B) > 0$, so we are not dividing by zero. As seen from a homework problem, we can prove that $(P(A \cap B)/P(B))I_B + (P(A \cap B^c)/P(B^c))I_{B^c}$ is actually the conditional expectation $E[I_A|\mathcal{D}]$ by showing that $(P(A \cap B)/P(B))I_B + (P(A \cap B^c)/P(B^c))I_{B^c}$ satisfies the two defining properties listed in Definition 6 or Definition 7. Thus, $E[I_A|\mathcal{D}]$ encompasses both $P(A|B)$ and $P(A|B^c)$; that is, $P(A|B)$ and $P(A|B^c)$ are simply the values of $E[I_A|\mathcal{D}]$ on $B$ and $B^c$, respectively. Also note that $P(\cdot|B)$ is actually a probability measure for each $B$ as long as $P(B) > 0$.

4.7. Joint densities and marginal densities. Let $X$ and $Y$ be random variables on $(\Omega, \mathcal{F}, P)$ with a distribution $\mu_{XY}(B)$ defined on all Borel subsets $B$ of the plane. Suppose that $X$ and $Y$ have a joint density. That is, there exists an integrable function $f_{XY}(\cdot, \cdot) : \mathbb{R}^2 \to \mathbb{R}$, such that

$$
\mu_{XY}(B) = \int_B f_{XY}(x, y) \, dx \, dy, \quad \text{for any } B \in \mathcal{B}_2.
$$

4.7.1. Marginal densities. Consider $\mu_{XY}(A \times \mathbb{R})$, where $A \in \mathcal{B}$. Notice that $\mu_{XY}(\cdot \times \mathbb{R})$ is a probability measure for any $A \in \mathcal{B}$. Also, by the integral representation shown in (10),

$$
\mu_{XY}(A \times \mathbb{R}) = \int_{A \times \mathbb{R}} f_{XY}(x, y) \, dx \, dy = \int_A \left( \int_{\mathbb{R}} f_{XY}(x, y) \, dy \right) \, dx \equiv \int_A f_X(x) \, dx,
$$

where $f_X(\cdot)$ is called the marginal density of $X$. Note that $f_X(\cdot)$ qualifies as the pdf of $X$. Thus, we arrive at the familiar result that the pdf of $X$ can be obtained from the joint pdf of $X$ and $Y$ through integration. Similarly, $f_Y(y) = \int_{\mathbb{R}} f_{XY}(x, y) \, dx.$
4.7.2. Conditional densities. Suppose that $X$ and $Y$ have a joint density $f_{XY}$. Then,

\begin{align}
\mathbb{E}[Y|X] &= \frac{\int_{\mathbb{R}} y f_{XY}(X, y) \, dy}{\int_{\mathbb{R}} f_{XY}(X, y) \, dy}, \\
\text{and in particular,} \\
\mathbb{E}[Y|X = x] &= \frac{\int_{\mathbb{R}} y f_{XY}(x, y) \, dy}{\int_{\mathbb{R}} f_{XY}(x, y) \, dy} \equiv g(x).
\end{align}

Proof

We will verify the definition of a conditional expectation given in Definition 7. First, we must show that $g(X)$ is $\sigma(X)$ measurable. This follows from the fact that $g$ is Borel measurable. (This is because $g$ is obtained from integrating a Borel-measurable function in the plane over one of the variables. This is proved in a course on integration. Let’s not worry about the details now.)

Next, we must show that $\mathbb{E}[g(X)W] = \mathbb{E}[YW]$ for every bounded $\sigma(X)$-measurable $W$. Without loss of generality, assume that $W = h(X)$ for some Borel function $h$. To see that $\mathbb{E}[g(X)h(X)] = \mathbb{E}[Yh(X)]$, write

\begin{align*}
\mathbb{E}[g(X)h(X)] &= \mathbb{E} \left[ \frac{\int_{\mathbb{R}} y f_{XY}(X, y) \, dy}{\int_{\mathbb{R}} f_{XY}(X, y) \, dy} h(X) \right] \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \left[ \frac{\int_{\mathbb{R}} t f_{XY}(x, t) \, dt}{f_X(x)} h(x) f_{XY}(x, y) \right] dx \, dy \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \left[ \frac{\int_{\mathbb{R}} t f_{XY}(x, t) \, dt}{f_X(x)} h(x) f_{XY}(x, y) \right] dx \, dy \\
&= \int_{\mathbb{R}} \left[ \frac{\int_{\mathbb{R}} t f_{XY}(x, t) \, dt}{f_X(x)} h(x) \int_{\mathbb{R}} f_{XY}(x, y) \, dy \right] dx \\
&= \int_{\mathbb{R}} \left[ \frac{\int_{\mathbb{R}} t f_{XY}(x, t) \, dt}{f_X(x)} h(x) f_X(x) \right] dx \\
&= \int_{\mathbb{R}} \left[ \left( \int_{\mathbb{R}} t f_{XY}(x, t) \, dt \right) h(x) \right] dx \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \left( \int_{\mathbb{R}} t f_{XY}(x, t) \, dt \right) h(x) \, dx \\
&= \mathbb{E}[Yh(X)].
\end{align*}
In the above, we have used two facts about integrations: (1) that we can write a double integral as an iterated integral, and (2) that we can exchange the order of the integration in the iterated integrals. These two properties are always true as long as the integrand is non-negative. This is a consequence of what is called Fubini’s Theorem in analysis [2, Chapter 8]. To complete the proof, we need to address the concern over the points over which $f_X(\cdot) = 0$ (in which case we will not be able to cancel $f_X$’s in the numerator and the denominator in the above development). However, it is straightforward to show that $\mathbb{P}\{f_X(X) = 0\} = 0$, which guarantees that we can exclude these problematic point from the integration in the $x$-direction without changing the integral. □

With the above results, we can think of

$$\frac{f_{XY}(x, y)}{\int_{\mathbb{R}} f_{XY}(x, y) \, dy}$$

as a conditional pdf. In particular, if we define

$$f_{Y|X}(y|x) \triangleq \frac{f_{XY}(x, y)}{\int_{\mathbb{R}} f_{XY}(x, y) \, dy} = \frac{f_{XY}(x, y)}{f_X(x)},$$

then we can calculate the conditional expectation $\mathbb{E}[Y|X]$ using the formula $\int_{\mathbb{R}} y f_{Y|X}(y|X) \, dy$, which is the familiar result that we know from undergraduate probability.
5. Convergence of random sequences

Consider a sequence of random variables $X_1, X_2, \ldots$, defined on the product space $(\Omega, \mathcal{F}, P)$, where as before, $\Omega = \bigotimes_{i=1}^{\infty} \Omega_i$ is the infinite-dimensional Cartesian product space and $\mathcal{F}$ is the smallest $\sigma$-algebra containing all cylinders. Since each $X_i$ is a random variable, when we talk about convergence of random variables we must do so in the context of functions.

5.1. Pointwise convergence. Let us take $\Omega_i = [0, 1]$, $\mathcal{F} = \mathcal{B} \cap [0, 1]$, and take $P$ as Lebesgue measure (generalized length) in $[0, 1]$. Consider the sequence of random variables $f_n(\omega)$ defined as follows:

\begin{equation}
X_n(\omega) = \begin{cases} 
  n^2 \omega, & 0 \leq \omega \leq n^{-1}/2 \\
  n-n^2\omega, & n^{-1}/2 < \omega \leq n^{-1} \\
  0, & \text{otherwise.}
\end{cases}
\end{equation}

It is straightforward to see that for any $\omega$, the sequence of random variables $X_n(\omega)$ converges to the constant function 0. We say that $X_n(\omega) \to 0$ pointwise, everywhere, or surely. Generally, a sequence of random variables $X_n$ converges to a random variable $X$ if for every $\omega \in \Omega$, $X_n(\omega) \to X(\omega)$. To be more precise, for every $\epsilon > 0$ and every $\omega \in \Omega$, there exists $n_0 \in \mathbb{N}$ such that $|X_n(\omega) - X(\omega)| < \epsilon$ whenever $n \geq n_0$. Note that in this definition $n_0$ not only depends upon the choice of $\epsilon$ but also on the choice of $\omega$. Note that in the example above we cannot drop the dependence of $n_0$ on $\epsilon$. To see that, observe that $\sup_{\omega} |X_n(\omega) - 0| = n/2$; thus, there is no single $n_0$ that will work for every $\omega$.

5.2. Uniform convergence. The strongest type of convergence is what’s called uniform convergence, in which case $n_0$ will be independent of the choice of $\omega$. More precisely, a sequence of random variables $X_n$ converges to a random variable $X$ uniformly if for every $\epsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that $|X_n(\omega) - X(\omega)| < \epsilon$ for any $\omega \in \Omega$ and whenever $n \geq n_0$. This is equivalent to saying that for every $\epsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that $\sup_{\omega \in \Omega} |X_n(\omega) - X(\omega)| < \epsilon$ whenever $n \geq n_0$. Can you make a small change in the example above to make the convergence uniform?
5.3. Almost-sure convergence (or almost-everywhere convergence). Now consider a slight variation of the example in (15).

\[
X_n(\omega) = \begin{cases} 
    n^2\omega, & 0 < \omega \leq n^{-1}/2 \\
    n - n^2\omega, & n^{-1}/2 < \omega \leq n^{-1} \\
    1, & \omega \in (n^{-1}, 1] \cap \mathbb{Q} \\
    0, & \omega \in (n^{-1}, 1] \cap \mathbb{Q}^c
\end{cases}
\]  

Note that in this case \(X_n(\omega) \to 0\) for \(\omega \in [0,1) \cap \mathbb{Q}^c\). Note that the Lebesgue measure (or generalize length) of the set of points for which \(X_n(\omega)\) does not converge to the function 0 is zero. (The latter is because the measure of the set of rational numbers is zero. To see that, let \(r_1, r_2, \ldots\) be an enumeration of the rational numbers. Pick any \(\epsilon > 0\), and define the intervals \(J_n = (r_n - 2^{-n-1}\epsilon, r_n + 2^{-n-1}\epsilon)\). Now, if we sum up the lengths of all of these intervals, we obtain \(\epsilon\). However, since \(\mathbb{Q} \subset \bigcup_{n=1}^{\infty} J_n\), the “length” of the set \(\mathbb{Q}\) cannot exceed \(\epsilon\). Since \(\epsilon\) can be selected arbitrarily small, we conclude that the Lebesgue measure of \(\mathbb{Q}\) must be zero.) In this case, we say that the sequence of functions \(X_n\) converges to the constant random variable 0 almost everywhere.

Generally, we say that a sequence of random variables \(X_n\) converges to a random variable \(X\) almost surely if there exists \(A \in \mathcal{F}\), with the property \(P(A) = 1\), such that for every \(\omega \in A\), \(X_n(\omega) \to X(\omega)\). This is the strongest convergence statement that we can make in a probabilistic sense (because we don’t care about the points that belong to a set that has probability zero). Note that we can write this type of convergence by saying that \(X_n \to X\) almost surely (or a.s.) if \(P\{\omega \in \Omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)\} = 1\), or simply when \(P\{\lim_{n \to \infty} X_n = X\} = 1\).

5.4. Convergence in probability (or convergence in measure). We say that the sequence \(X_n\) converges to a random variable \(X\) in probability (also called in measure) when for every \(\epsilon > 0\), \(\lim_{n \to \infty} P\{|X_n - X| > \epsilon\} = 0\). This a weaker type of convergence than almost sure convergence, as seen next.

**Theorem 8.** Almost sure convergence implies convergence in probability.

**Proof**

Let \(\epsilon > 0\) be given. For each \(N \geq 1\), define \(E_N = \{|X_n - X| > \epsilon, \text{ for some } n \geq N\}\).
If we define $E = \bigcap_{N=1}^{\infty} E_N$, we'll observe that $\omega \in E$ implies that $X_n(\omega)$ does not converge to $X$. (This is because if $\omega \in E$, then no matter how large we pick $N$, $N_0$, say, there will be an $n \geq N_0$ such that $|X_n - X| > \epsilon$.) Hence, since we know that $X_n$ converges to $X$ a.s., it must be true that $P(E) = 0$. Now observe that since $E_1 \supset E_2 \supset \ldots$, $P(E) = \lim_{N \to \infty} P(E_N)$, and therefore $\lim_{N \to \infty} P(E_N) = 0$. Observe that if $\omega \in \{|X_n - X| > \epsilon\}$ and $n \geq N$, then $\omega \in \{|X_n - X| > \epsilon\}$ for some $n \geq N$. Thus, for $n \geq N$, $\omega \in \{|X_n - X| > \epsilon\} \subset E_N$ and $P\{|X_n - X| > \epsilon\} \leq P(E_N)$. Hence, $\lim_{n \to \infty} P\{|X_n - X| > \epsilon\} = 0$. \square

This theorem is not true, however, in infinite measure spaces. For example, take $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}$, and the Lebesgue measure $m$. Let $X_n(\omega) = \omega/n$. Then clearly, $X_n \to 0$ everywhere, but $m(\{|X_n - 0| > \epsilon\}) = \infty$ for any $\epsilon > 0$. Where did we use the fact that $P$ is a finite measure in the proof of the Theorem 8?

5.5. **Convergence in the mean-square sense (or in $L_2$).** We say that the sequence $X_n$ converges to a random variable $X$ in the mean-square sense (also in $L_2$) if $\lim_{n \to \infty} \mathbb{E}[|X_n - X|^2] = 0$. Recall that this is precisely the type of convergence we defined in the Hilbert-space context, which we used to introduce the conditional expectation. This a stronger type of convergence than convergence in probability, as seen next.

**Theorem 9.** Convergence in the mean-square sense implies convergence in probability.

**Proof**

This is an easy consequence of Markov inequality when taking $\psi(\cdot) = (\cdot)^2$, which yields $P\{|X_n - X| > \epsilon\} \leq \mathbb{E}[|X_n - X|^2]/\epsilon^2$. \square

Convergence in probability does not imply almost-sure convergence, as seen from the following example.
Example 16 (The marching-band function). Consider \( \Omega = [0, 1] \), take \( \mathbf{P} \) to be Lebesgue measure on \([0, 1]\), and define \( X_n \) as follows:

\[
X_n = \begin{cases} 
I_{[2^{-1}(n-1),2^{-1}n]}(\omega), & n \in \{1, 2\} \\
I_{[2^{-2}(n-1-2^i),2^{-2}(n-2^i)]}(\omega), & n \in \{1 + 2^1, \ldots, 2^1 + 2^2\} \\
\vdots \\
I_{[2^{-k}(n-1-2^i,\ldots,2^{k-1},2^i(n-2^i,\ldots,2^{k-1})]}(\omega), & n \in \{1 + 2^1 + \ldots + 2^{k-1}, 2^1 + \ldots + 2^{k-1} + 2^k\} \\
\vdots 
\end{cases}
\]

Note that for any \( \omega \in [0, 1] \), \( X_n(\omega) = 1 \) for infinitely many \( n \)'s; thus, \( X_n \) does not converge to 0 anywhere. At the same time, if \( n \in \{1 + 2^1 + \ldots + 2^{k-1}, 2^1 + \ldots + 2^{k-1} + 2^k\} \), then \( \mathbf{P}\{|X_n - 0| > \epsilon\} = 2^{-k} \); hence, \( \lim_{n \to \infty} \mathbf{P}\{|X_n - 0| > \epsilon\} = 0 \) and \( X_n \) converges to 0 in probability.

Note that in the above example, \( \lim_{n \to \infty} \mathbb{E}[|X_n - 0|^2] = 0 \), so \( X_n \) also converges to 0 in \( L_2 \).

Later we shall prove that any sequence of random variables that converges in probability will have a subsequence that converges almost surely to the same limit.

Recall the example given in (15) and let’s modify it slightly as follows.

\[
X_n(\omega) = \begin{cases} 
n^{3/2}\omega, & 0 \leq \omega \leq n^{-1}/2 \\
n^{1/2} - n^{3/2}\omega, & n^{-1}/2 < \omega \leq n^{-1} \\
0, & \text{otherwise.} 
\end{cases}
\]

In this case, \( X_n \) continues to converge to 0 at every point; nonetheless, \( \mathbb{E}[|X_n - 0|^2] = 1/12 \) for every \( n \). Thus, \( X_n \) does not converge to 0 in \( L_2 \).

5.6. Convergence in distribution. A sequence \( X_n \) is said to converge to a random variable \( X \) in distribution if the distribution functions, \( F_{X_n}(x) \), converge to the distribution function of \( X \), \( F_X(x) \), at every point of continuity of \( F_X(x) \).

Theorem 10. Convergence in probability implies convergence in distribution.
Proof (Adapted from T. G. Kurtz)

Pick $\epsilon > 0$ and note that

$$P\{X \leq x + \epsilon\} = P\{X \leq x + \epsilon, X_n > x\} + P\{X \leq x + \epsilon, X_n \leq x\}$$

and

$$P\{X_n \leq x\} = P\{X_n \leq x, X > x + \epsilon\} + P\{X_n \leq x, X \leq x + \epsilon\}.$$ 

Thus,

$$P\{X_n \leq x\} - P\{X \leq x + \epsilon\} = P\{X_n \leq x, X > x + \epsilon\} - P\{X \leq x + \epsilon, X_n > x\}.$$ 

Note that since $\{X_n \leq x, X > x + \epsilon\} \subset \{|X_n - X| > \epsilon\}$,

$$P\{X_n \leq x\} - P\{X \leq x + \epsilon\} \leq P\{|X_n - X| > \epsilon\}.$$ 

By taking the limit superior of both sides and noting that $\lim_{n \to \infty} P\{|X_n - X| > \epsilon\} = 0$, we obtain $\overline{\lim}_{n \to \infty} P\{X_n \leq x\} \leq P\{X \leq x + \epsilon\}$.

Now replace every occurrence of $\epsilon$ in (19) with $-\epsilon$, multiply both side by -1 and obtain

$$-P\{X_n \leq x\} + P\{X \leq x - \epsilon\} = -P\{X_n \leq x, X > x - \epsilon\} + P\{X \leq x - \epsilon, X_n > x\}.$$ 

Since $\{X \leq x - \epsilon, X_n > x\} \subset \{|X_n - X| > \epsilon\}$, we have

$$-P\{X_n \leq x\} + P\{X \leq x - \epsilon\} \leq P\{|X_n - X| > \epsilon\}.$$ 

By taking the limit inferior of both sides, we obtain $\underline{\lim}_{n \to \infty} P\{X_n \leq x\} \geq P\{X \leq x - \epsilon\}$.

Combining, we obtain $P\{X \leq x - \epsilon\} \leq \overline{\lim}_{n \to \infty} P\{X_n \leq x\} \leq \overline{\lim}_{n \to \infty} P\{X_n \leq x\} \leq P\{X \leq x + \epsilon\}$. Thus, by letting $\epsilon \to 0$ we conclude that $\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$ at every point of continuity of $F_X$. □

Theorem 11 (Bounded Convergence Theorem). Suppose that $X_n$ converges to $X$ in probability, and that $|X_n| \leq M$, a.s., $\forall n$, for some fixed $M$. Then $E[X_n] \longrightarrow E[X]$. 


This example shows that the boundedness requirement is not superfluous. Take \( \Omega = [0, 1] \), \( \mathcal{F} = \mathcal{B} \cap [0, 1] \), and \( m \) as Lebesgue measure (i.e., \( m(A) = \int_A \, dx \)). Consider the functions \( f_n(t) \) shown below in the figure. Note that \( f_n(t) \to 0, \forall t \in \Omega \). But, \( E[f_n(t) - 0] = 1 \), for any \( n \). Therefore, \( E[f_n] \not\to E[0] = 0 \).

Proof of Theorem (Adapted from A. Beck).

Let \( E_n = \{|X_n| > M\} \) and note that \( P(E_n) = 0 \) by the boundedness assumption. Define \( E_0 = \bigcup_{n=1}^{\infty} E_n \) and note that \( P(E_0) \leq \sum_{n=1}^{\infty} P(E_n) = P(E_1) + P(E_2) + \ldots = 0 \). We will first show that \( |X| \leq M \) a.s. Consider \( W_{n,k} \triangleq \{|X_n - X| > 1/k\}, k = 1, 2, \ldots \). Since \( X_n \) converges to \( X \) in probability, \( P(W_{n,k}) \xrightarrow{n \to \infty} 0 \). Let \( F_k = \{|X| > M + 1/k\} \). Note that off \( E_0 \), \( F_k \subset W_{n,k} \) a.s., regardless of \( n \), and \( P(F_k) \leq P(W_{n,k}) \). Then, \( P(F_k) \leq \lim_{n \to \infty} P(W_{n,k}) = 0 \), or \( P(F_k) = 0 \). Therefore, \( P(\bigcup_{i=1}^{\infty} F_k) = 0 \). Since \( \bigcup_{i=1}^{\infty} F_k = \{|X| > M\} \), \( |X| \leq M \) a.s.

Let \( \epsilon > 0 \) be given. Let \( F_0 \triangleq \bigcup_{i=1}^{\infty} F_k \) and \( G_n = \{|X_n - X| > \epsilon\} \). Note that \( \Omega \) can be written as the disjoint union \((F_0 \cup E_0) \cup (G_n \cup F_0 \cup E_0)^c \cup (G_n \setminus (F_0 \cup E_0))\). Thus, \(|E[X_n - X]| \leq E[|X_n - X|] = E[|X_n - X|I_{F_0 \cup E_0}] + E[|X_n - X|I_{G_n \setminus (F_0 \cup E_0)}] + E[|X_n - X|I_{(G_n \cup F_0 \cup E_0)^c}] \leq 0 + 2M E[I_{G_n}] + \epsilon = 2ME[I_{G_n}] + \epsilon \).

Since \( P(G_n) \to 0 \), there exist \( n_0 \) such that \( n \geq n_0 \) implies \( P(G_n) \leq \frac{\epsilon}{2M} \). Hence, for \( n \geq n_0 \), \(|E[X_n] - E[X]| \leq \frac{2M\epsilon}{2M} + \epsilon = 2\epsilon \), which establishes \( E[X_n] \to E[X] \) as \( n \to \infty \). \( \square \)
Lemma 2 (Adapted from T. G. Kurtz). Suppose that $X \geq 0 \ a.s.$ Then, $\lim_{M \to \infty} E[X \wedge M] = E[X]$.

Proof
First note that the statement is true when $X_n$ is a discrete random variable. Then recall that we can always approximate a random variable $X$ monotonically from below by a discrete random variable. More precisely, define $X_n \leq X$ as in (5) and note that $X_n \uparrow X$ and $E[X_n] \uparrow E[X]$. Now since $X_n \wedge M \leq X \wedge M \leq X$, we have

$$E[X_n] = \lim_{M \to \infty} E[X_n \wedge M] \leq \lim_{M \to \infty} E[X \wedge M] \leq \lim_{M \to \infty} E[X \wedge M] \leq E[X].$$

Take the limit as $n \to \infty$ to obtain

$$E[X] \leq \lim_{M \to \infty} E[X \wedge M] \leq \lim_{M \to \infty} E[X \wedge M] \leq E[X],$$

from which the Lemma follows. \(\square\)

Theorem 12 (Monotone Convergence Theorem, Adapted from T. Kurtz). Suppose that $0 \leq X_n \leq X$ and $X_n$ converges to $X$ in probability. Then, $\lim_{n \to \infty} E[X_n] = E[X]$.

Proof
For $M > 0$, $X_n \wedge M \leq X_n \leq X$. Thus,

$$E[X_n \wedge M] \leq E[X_n] \leq E[X].$$

By the bounded convergence theorem, $\lim_{n \to \infty} E[X_n \wedge M] = E[X \wedge M]$. Hence,

$$E[X \wedge M] \leq \lim_{n \to \infty} E[X_n] \leq \lim_{n \to \infty} E[X_n] \leq E[X].$$
Now by Lemma 2, \( \lim_{M \to \infty} E[X \wedge M] = E[X] \), and we finally obtain
\[
E[X] \leq \lim_{n \to \infty} E[X_n] \leq \lim_{n \to \infty} E[X_n] \leq E[X],
\]
and the theorem is proven. \( \square \)

**Lemma 3** (Fatou’s Lemma, Adapted from T. G. Kurtz). *Suppose that \( X_n \geq 0 \) a.s. and \( X_n \) converges to \( X \) in probability. Then, \( \lim_{n \to \infty} E[X_n] \geq E[X] \).*

**Proof**
For \( M > 0 \), \( X_n \wedge M \leq X_n \). Thus,
\[
E[X \wedge M] = \lim_{n \to \infty} E[X_n \wedge M] \leq \lim_{n \to \infty} E[X_n],
\]
where the left equality is due to the bounded convergence theorem. Now by Lemma 2, \( \lim_{M \to \infty} E[X \wedge M] = E[X] \), and we obtain \( E[X] \leq \lim_{n \to \infty} E[X_n] \). \( \square \)

**Theorem 13** (Dominated Convergence Theorem). *Suppose that \( X_n \to X \) in probability, \( |X_n| \leq Y \), and \( E[Y] < \infty \). Then, \( E[X_n] \to E[X] \).*

**Proof:**
Since \( X_n + Y \geq 0 \) and \( Y + X_n \to Y + X \) in probability, we use Fatou’s lemma to write \( \lim_{n \to \infty} E[Y + X_n] \geq E[Y + X] \). This implies \( E[Y] + \lim_{n \to \infty} E[X_n] \geq E[Y] + E[X] \), or \( \lim_{n \to \infty} E[X_n] \geq E[X] \). Similarly, \( \lim_{n \to \infty} E[Y - X_n] \geq E[Y - X] \), which implies \( E[Y] + \lim_{n \to \infty} (-E[X_n]) \geq E[Y] - E[X] \), or \( \lim_{n \to \infty} (-E[X_n]) \geq -E[X] \). But \( \lim_{n \to \infty} (-x_n) = -\lim_{n \to \infty} x_n \), and therefore \( \lim_{n \to \infty} E[X_n] \leq E[X] \). In summary, we have \( \lim_{n \to \infty} E[X_n] \leq E[X] \leq \lim_{n \to \infty} E[X_n] \), which implies \( \lim_{n \to \infty} E[X_n] = E[X] \). \( \square \)

### 5.7. The \( L_1 \) Conditional expectation, revisited.
Consider a probability space \( (\Omega, \mathcal{F}, \mathbb{P}) \) and let \( X \in L_1(\Omega, \mathcal{F}, \mathbb{P}) \). Let \( \mathcal{D} \) be a sub \( \sigma \)-algebra. For each \( n \geq 1 \), we define \( X_n \overset{\Delta}{=} X \wedge n \). Note that \( X_n \in L_2(\Omega, \mathcal{F}, \mathbb{P}) \) (since it is bounded), so we can talk about \( E[(X_n) | \mathcal{D}] \) as a projection of \( X_n \) onto \( L_2(\Omega, \mathcal{D}, \mathbb{P}) \), which we call \( Z_n \). We will now show that \( Z_n \) is a Cauchy sequence in \( L_1 \), which means \( \lim_{m,n \to \infty} E[|Z_n - Z_m|] = 0 \).
We already know that \( E[(Z_n - Z_m)Y] = E[(X_n - X_m)Y] \) for any \( Y \in L_2(\Omega, \mathcal{D}, \mathbb{P}) \). In particular, pick \( Y = I_{\{Z_n - Z_m > 0\}} - I_{\{Z_n - Z_m \leq 0\}} \). In this case, \( (Z_n - Z_m)Y = |Z_n - Z_m| \). Thus, we conclude that \( E[|Z_n - Z_m|] = E[(X_n - X_m)(I_{\{Z_n - Z_m > 0\}} - I_{\{Z_n - Z_m \leq 0\}})] \).
But the right-hand side is no greater than $E[|X_n - X_m|]$ (why?), and we obtain 
$E[|Z_n - Z_m|] \leq E[|X_n - X_m|]$. However, $E[|X_n - X_m|] \leq E[|X|I_{\{|X| \geq \min(m,n)\}}] \to 0$ by 
the dominated convergence theorem (verify this), and we conclude that $\lim_{m,n \to \infty} E[|Z_n - 
Z_m|] = 0$.

From a key theorem in analysis (e.g., see Rudin, Chapter 3), we know that any 
$L_1$ space is complete. Thus, there exists $Z \in L_1(\Omega, \mathcal{D}, \mathbb{P})$ such that 
$\lim_{n \to \infty} E[|Z_n - Z|] = 0$. Further, $Z_n$ has a subsequence, $Z_{n_k}$, that converges 
almost surely to $Z$. We take this $Z$ as a candidate for $E[X|D]$. But first, we must show that $Z$ satisfies 
$E[ZY] = E[XY]$ for any bounded, $\mathcal{D}$-measurable $Y$. This is easy to show. Suppose 
that $|Y| < M$, for some $M$. We already know that $E[Z_nY] = E[X_nY]$, and by 
the dominated convergence theorem, $E[X_nY] \to E[XY]$ (since $|X_n| \leq |X|$). Also, 
This leads to the conclusion that $E[ZY] = E[XY]$ for any bounded, $\mathcal{D}$-measurable $Y$.

In summary, we have found a $Z \in L_1(\Omega, \mathcal{D}, \mathbb{P})$ such that $E[ZY] = E[XY]$ for any 
bounded, $\mathcal{D}$-measurable $Y$. We define this $Z$ as $E[X|D]$.

5.8. Central Limit Theorems. To establish this theorem, we need the concept of 
a characteristic function. Let $Y$ be a r.v. We define the characteristic function of 
$Y$, $\phi_Y(u)$, as $E[e^{iyY}]$, where $E[e^{iyY}]$ exists if $E[\text{Re}(e^{iyY})] < \infty$ and $E[\text{Im}(e^{iyY})] < \infty$. 
Note that $\text{Re}(e^{iyY}) \leq |e^{iyY}| = 1$ and $\text{Im}(e^{iyY}) \leq 1$. Therefore, $E[|\text{Re}(e^{iyY})|] \leq 1$ and 
$E[|\text{Im}(e^{iyY})|] \leq 1$. Hence, $E[e^{iyY}]$ is well defined for any $u \in \mathbb{R}$. If $Y$ has a pdf $f_Y(y)$, 
then $E[e^{iyY}] = \int_{-\infty}^{\infty} e^{iyu} f_Y(y)dy$, which is the Fourier transform of $f_Y$ evaluated at 
$-u$.

Example 17. $Y \in \{0, 1\}$ and $P\{Y = 1\} = p$. In this case $\phi_Y(u) = E[e^{iyY}] = 
pe^{iu} + (1-p)e^{i0} = pe^{iu} + (1-p)$.

Example 18. $Y \sim N(0, 1)$. Then, $E[e^{iyY}] = e^{-\frac{y^2}{2}}$ (see justification below).

We also need this Theorem that relates pdf’s to characteristic functions.

5.9. Levy’s Inversion Lemma.
Lemma 4. If $\phi_X(u)$ is the characteristic function of $X$, then
\[
\lim_{c \to \infty} \frac{1}{2\pi} \int_{-c}^{c} e^{iuX} \phi_X(u) du = P\{a < X < b\} + \frac{P\{X=a\} + P\{X=b\}}{2}.
\]

See Chow and Teicher for proof.

As a special case, if $X$ is an absolutely continuous random variable with pdf $f_X$, then
\[
\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iuX} \phi_X(u) du = f_X(x).
\]

Theorem 14 (Central limit theorem). Suppose \( \{X_k\} \) is a sequence of i.i.d. random variables with $E[X_k] = \mu$ and $\text{Var}X_k = \sigma^2$. Let $S_n = \sum_{k=1}^{n} X_k$ and $Y_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$. Then, $Y_n$ converges to $Z$ in distribution, where $Z \sim N(0,1)$ (zero-mean Gaussian random variable with unit variance).

Sketch of the proof

Observe that
\[
\Phi_{Y_n}(\omega) = E[e^{i\omega Y_n}]
\]
\[
= E\left[e^{i\omega \sum_{k=1}^{n} X_k - \mu} \right]
\]
\[
= \prod_{k=1}^{n} E\left[e^{i\omega \sqrt{n}\sigma(X_k - \mu)} \right]
\]
\[
= \prod_{k=1}^{n} \Phi_{X_k - \mu}(\frac{\omega}{\sqrt{n}\sigma})
\]
\[
= (\Phi_{X_k - \mu}(\frac{\omega}{\sqrt{n}\sigma}))^n
\]

We now expand $\Phi_{X_k - \mu}(\omega)$ as
\[
\Phi_{X_k - \mu}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iu(X_k - \mu)} du
\]
\[
= E[1 + j\omega(X_k - \mu) + \frac{(j\omega)^2}{2!}(X_k - \mu)^2 + \frac{(j\omega)^3}{3!}(X_k - \mu)^3 + \ldots]
\]
\[
= 1 + j\omega E[X_k - \mu] - \frac{\omega^2}{2} E[(X_k - \mu)^2] + \frac{(j\omega)^3}{3!} E[(X_k - \mu)^3] + \ldots
\]
\[
= 1 - \frac{\omega^2}{2} E[(X_k - \mu)^2] + \frac{(j\omega)^3}{3!} E[(X_k - \mu)^3] + \ldots
\]
Then we have
\[
\Phi_{(X_k-\mu)}\left(\frac{\omega}{\sqrt{n}\sigma}\right) = 1 - \frac{(\frac{\omega}{\sqrt{n}\sigma})^2}{2} E[(X_k-\mu)^2] + \frac{(\frac{\omega}{\sqrt{n}\sigma})^3}{3!} E[(X_k-\mu)^3] + \ldots \\
= 1 - \frac{\omega^2}{2n\sigma^2} E[(X_k-\mu)^2] - \frac{j\omega^3}{3!\sqrt{n}\sigma^3} E[(X_k-\mu)^3] + \ldots \\
\approx 1 - \frac{\omega^2}{2n\sigma^2} E[(X_k-\mu)^2].
\]

Then \( \Phi_{Y_n}(\omega) \approx (1 - \frac{\omega^2}{2n})^n \) and \( \lim_{n \to \infty} \Phi_{Y_n}(\omega) = \lim_{n \to \infty} (1 - \frac{\omega^2}{n})^n = e^{-\omega^2/2}. \) On the other hand, the characteristic function of \( Z \) is \( \Phi_Z(\omega) = E[e^{j\omega z}]. \) Then
\[
\Phi_Z(\omega) = E[e^{j\omega z}]
\]
\[
= \int_{-\infty}^{\infty} e^{j\omega z} f_Z(z) dz
\]
\[
= \int_{-\infty}^{\infty} e^{j\omega z} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz
\]
\[
= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z^2-2j\omega z)} dz
\]
\[
= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}((z-j\omega)^2+\omega^2)} dz
\]
\[
= \frac{1}{\sqrt{2\pi}} e^{-\omega^2/2} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-j\omega)^2} dz
\]
\[
= e^{-\omega^2/2}
\]

Therefore, \( \lim_{n \to \infty} \Phi_{Y_n}(\omega) = \Phi_Z(\omega) \) and \( F_{Y_n} \xrightarrow{n \to \infty} F_Z \) by the inversion lemma. 

5.10. Central limit theorem for non-identical random variables. Suppose \( \{X_k\} \) is a sequence of random variables that are independent but non-identical. \( E[X_k] = \mu_k \) and \( \text{Var}\{X_k\} = \sigma_k^2. \) Let \( S_n = \sum_{k=1}^{n} X_k. \) Then \( E[S_n] = \sum_{k=1}^{n} E[X_k] = \sum_{k=1}^{n} \mu_k = \eta_n \) and \( \text{Var}\{S_n\} = \sum_{k=1}^{n} \sigma_k^2 = \beta_n^2. \)

Let \( Y_n = \frac{S_n-\eta_n}{\beta_n} \) and suppose \( \{X_k\} \) satisfies the following two conditions:

a) \( \lim_{n \to \infty} \beta_n^2 = \lim_{n \to \infty} \sum_{k=1}^{n} \sigma_k^2 = \infty; \)

b) There exists \( \alpha > 2 \) and \( c < \infty \) such that \( E[|X_k|^\alpha] < c \forall k.\)
Then, $Y_n$ converges in distribution to a zero-mean unit variance Gaussian random variable. (For proof see, for example, Chow and Teicher.). Equivalent conditions to the above two conditions are:

a’) There is a constant $\epsilon > 0$ such that $\sigma_k^2 > \epsilon \forall k$.

b’) All densities $f_k(x)$ (if they exist) are zero outside a finite interval $(-c, c)$ for some $c$.

5.11. **Strong Law of Large Numbers.** Suppose that $X_1, X_2, \ldots$, are i.i.d. random variables, $E[X_1] = \mu < \infty$, and $E[(X_1 - \mu)^2] = \sigma^2 < \infty$. Let $S_n \triangleq n^{-1} \sum_{i=1}^{n} X_i$. Note that $E[S_n] = \mu < \infty$ and $E[(S_n - \mu)^2] = n^{-1}\sigma^2$. By Chebychev’s inequality,$$
P\{|S_n - \mu| > \epsilon\} \leq \frac{\sigma^2}{n\epsilon},$$and therefore we conclude that $S_n$ converges to $\mu$ in probability. This is called the **weak law of large numbers**.

As it turns out, this convergence occurs almost surely, yielding what is called the **strong law of large numbers**. To prove one version of the strong law, one needs the notion of a sequence of events occurring infinitely often. Roughly speaking, if $A_1, A_2, \ldots$, is a sequence of events (think for example of the sequence $\{|S_n - \mu| > \epsilon\}$), then one can look for the collection of all outcomes $\omega$, each of which belonging to infinitely many of the events $A_n$. For example, if $\omega_o$ is such outcome, and if we take $A_n = \{|S_n - \mu| > \epsilon\}$, where $\epsilon > 0$, then we would know that $S_n(\omega_o)$ cannot converge to $\mu$.

More generally, we define the event $\{A_n$ occurs infinitely often $\}$, or for short $\{A_n \text{ i.o.}\}$, as

$$\{A_n \text{ i.o.}\} \triangleq \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k.$$The above event is also referred to as the limit superior of the sequence $A_n$, that is, $\{A_n \text{ i.o.}\} = \lim_{n \to \infty} A_n$. The terminology of limit superior makes sense since the $\bigcup_{k=n}^{\infty} A_k$ is a decreasing sequence and the intersection yields the infimum of the decreasing sequence.

Before we prove a key result on the probability of events such as $\{A_n \text{ i.o.}\}$, we will give an example showing their use.
Let $Y_n$ be any sequence, and let $Y$ be a random variable (possibly being a candidate limit of the sequence provided that the sequence is convergent). Consider the event $\{\lim_{n \to \infty} Y_n = Y\}$, the collection of all outcomes $\omega$ for which $Y_n(\omega) \to Y(\omega)$. It is easy to see that

\[\{\lim_{n \to \infty} Y_n = Y\}^c = \cup_{\epsilon > 0, \epsilon \in \mathbb{Q}} \{|Y_n - Y| > \epsilon\ \text{i.o.}\}.\]

This is because if $\omega_o \in \{|Y_n - Y| > \epsilon\ \text{i.o.}\}$ for some $\epsilon > 0$, then we are guaranteed that for any $n \geq 1$, $|Y_k - Y| > \epsilon$ for infinitely many $k \geq n$; thus, $Y_n(\omega_o)$ does not converge to $Y(\omega_o)$.

**Lemma 5** (Borel-Cantelli Lemma). Let $A_1, A_2, \ldots$, be any sequence of events. If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P\{A_n \text{ i.o.}\} = 0$.

**Proof**

Note that $\sum_{n=1}^{\infty} P(A_n) < \infty$ implies $\sum_{k=n}^{\infty} P(A_k) \to 0$ as $n \to \infty$. Since $P(\cup_{k=n}^{\infty} A_k) \leq \sum_{k=n}^{\infty} P(A_k)$, $P(\cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k) = \lim_{n \to \infty} P(\cup_{k=n}^{\infty} A_k) = 0$ (recall that $\cup_{k=n}^{\infty} A_k$ is a decreasing sequence of events). Hence, $P\{A_n \text{ i.o.}\} = 0$. □

The next Theorem is an application of the Borel-Cantelli Lemma.

**Theorem 15** (A strong law of large numbers). Let $X_1, X_2, \ldots$, be an i.i.d. sequence, $E[X_1] = \mu < \infty$, $E[X_1^2] = \mu_2 < \infty$, and $E[X_1^4] = \mu_4 < \infty$. Then, $S_n \overset{\Delta}{=} n^{-1} \sum_{i=1}^{n} X_i \to \mu$ almost surely.

**Proof**

Without loss of generality, we will assume that $\mu = 0$. From (20), we know that

\[\{\lim_{n \to \infty} S_n = 0\}^c = \cup_{\epsilon > 0, \epsilon \in \mathbb{Q}} \{|S_n| > \epsilon\ \text{i.o.}\}.\]

By Markov inequality (with $\psi(\cdot) = (\cdot)^4$),

$$P\{|S_n| > \epsilon\} \leq \frac{E[S_n^4]}{\epsilon^4} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \sum_{m=1}^{n} E[X_i X_j X_l X_m]}{n^4 \epsilon^4}.$$  

Considering the term in the numerator, we observe that all the terms in the summation that are of the form $E[X_i^3] E[X_j]$, $E[X_i^3] E[X_j] E[X_l]$, $E[X_i] E[X_j] E[X_l] E[X_m]$
\((i \neq j \neq \ell \neq m)\) are zero. Hence,
\[
\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{\ell=1}^{n} \sum_{m=1}^{n} E[X_iX_jX_\ell X_m] = nE[X_1^4] + 3n(n-1)E[X_1^2]E[X_1^2],
\]
and therefore
\[
P\{|S_n| > \epsilon\} \leq \frac{n\mu_4 + 3n(n-1)(2)}{n^4\epsilon^4}.
\]
Hence, we conclude that
\[
\sum_{n=1}^{\infty} P\{|S_n| > \epsilon\} < \infty.
\]
By the Borel-Cantelli Lemma,
\[
P\{|S_n| > \epsilon \text{ i.o.}\} = 0.
\]
Hence, by (21)
\[
P\left(\{ \lim_{n \to \infty} S_n = 0 \}^c\right) = 0
\]
and the theorem is proved. \(\square\)

Actually, the frequent version of the law of large numbers does not require any assumptions on second and fourth moment. We will not prove this version here, but it basically states that if \(E[X_1] < \infty\), then \(n^{-1} \sum_{i=1}^{n} X_i \to \mu\) almost surely. We next prove a corollary that shows that the law of large numbers holds even when \(E[X_1] = \infty\).

**Corollary 1.** Let \(X_1, X_2, \ldots\), be an i.i.d. sequence, and \(E[X_1] = \infty\). Then, \(n^{-1} \sum_{i=1}^{n} X_i \to \infty\) almost surely.

**Proof**

Without loss of generality, assume that \(X_n \geq 0\). Let \(M\) be an integer; note that
\[
n^{-1} \sum_{i=1}^{n} X_i \geq n^{-1} \sum_{i=1}^{\infty} X_i \wedge M.
\]

Therefore,
\[
\lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} X_i \geq \lim_{n \to \infty} n^{-1} \sum_{i=1}^{\infty} X_i \wedge M = E[X_1 \wedge M],
\]
where the last equality follows from the law of large numbers. Hence, it follows that

\[ \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} X_i \geq \lim_{M \to \infty} E[X_1 \wedge M] = \infty, \]

where the last limit follows from Lemma 2.

The next theorem uses the Borel-Cantelli Lemma to characterize convergence of probability.

**Theorem 16** (From Billingsley). A necessary and sufficient condition for \( X_n \to X \) in probability is that each subsequence \( \{X_{n_k}\} \) contains a further subsequence \( \{X_{n_{k(i)}}\} \) such that \( X_{n_{k(i)}} \to X \) a.s. as \( i \to \infty \).

**Proof.** If \( X_n \to X \) in probability, then given \( \{n_k\} \), choose a subsequence \( \{n_{k(i)}\} \) so that \( k \geq k(i) \) implies that \( P\{|X_{n_k} - X| \geq i^{-1}\} < 2^{-i} \); therefore, \( \sum_{i=1}^{\infty} P\{|X_{n_{k(i)}} - X| \geq i^{-1}\} < \infty \). By the Borel-Cantelli lemma, \( P\{|X_{n_{k(i)}} - X| < i^{-1}\} = 1 \) for all but finitely many \( i \). Therefore, \( \lim_{i \to \infty} X_{n_{k(i)}} = X \) a.s.

Conversely, if \( X_n \) does not converge to \( X \) in probability, there is some \( \epsilon > 0 \) and a subsequence \( \{n_k\} \) for which \( P\{|X_{n_k} - X| \geq \epsilon\} > \epsilon \) for all \( k \). No subsequence of \( \{X_{n_k}\} \) can converge to \( X \) in probability; hence, none can converge to \( X \) almost surely. \( \square \)

**Corollary 2.** If \( g : \mathbb{R} \to \mathbb{R} \) is continuous and \( X_n \to X \) in probability, then \( g(X_n) \to g(X) \) in probability.

**Proof:** Left as an exercise.
6. Markov Chains

The majority of the materials of this section are extracted from the book by Billingsley (Probability and Measure) and the book by Hoel et al. (Introduction to Stochastic Processes.)

Up to this point we have seen many examples and properties (e.g., convergence) of iid sequences. However, many sequences have correlation among their members. Markov chains (MCs) are a class of sequences that exhibit a simple type of correlation in the sequence of random variables. In words, as far as calculating probabilities pertaining future values of the sequence conditional on the present and the past, it is sufficient to know the present. In other words, as far as the future is concerned, knowledge of the present contains all the knowledge of the past. We next define discrete-space MCs formally.

Consider a sequence of discrete random variables $X_0, X_1, X_2, \ldots$, defined on some probability space $(\Omega, \mathcal{F}, P)$. Let $S$ be a finite or countable set representing the set of all possible values that each $X_n$ may assume. Without loss of generality, assume that $S = \{0, 1, 2, \ldots\}$. The sequence is a Markov chain if

$$P\{X_{n+1} = j | X_0 = i_0, \ldots, X_n = i_n\} = P\{X_{n+1} = j | X_n = i_n\}.$$

For each pair $i$ and $j$ in $S$, we define $p_{ij} \triangleq P\{X_{n+1} = j | X_n = i\}$. Note that we have implicitly assumed (for now) that $P\{X_{n+1} = j | X_n = i\}$ is not dependent upon $n$, in which case the chain is called homogeneous. Also note that the definition of $p_{ij}$ implies that $\sum_{j \in S} p_{ij} = 1, i \in S$. The numbers $p_{ij}$ are called the transition probabilities of the MC. The initial probabilities are defined as $\alpha_i \triangleq P\{X_0 = i\}$. Note that the only restriction on the $\alpha_i$s is that they are nonnegative and they must add up to 1. We denote the matrix whose $(i, j)$th entry is $p_{ij}$ by $\mathbf{P}$, which is termed the transition matrix of the MC. Note that if $S$ is infinite, then $\mathbf{P}$ is a matrix with infinitely many rows and columns. The transition matrix $\mathbf{P}$ is called a stochastic matrix, as each row sums up to one and all its elements are nonnegative. Figure 6 shows a state diagram illustrating the transition probabilities for a two-state MC.
6.1. Higher-Order Transitions. Consider $P\{X_0 = i_0, X_1 = i_1, X_2 = i_2\}$. By applying Bayes rule repeatedly (i.e., $P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C)$), we obtain

$$P\{X_0 = i_0, X_1 = i_1, X_2 = i_2\} = P\{X_0 = i_0\}P\{X_1 = i_1|X_0 = i_0\}P\{X_2 = i_2|X_0 = i_0, X_1 = i_1\} = \alpha_{i_0}p_{i_0i_1}p_{i_1i_2}.$$ 

More generally, it is easy to verify that

$$P\{X_0 = i_0, X_1 = i_1, \ldots, X_m = i_m\} = \alpha_{i_0}p_{i_0i_1} \cdots p_{i_{m-1}i_m}$$

for any sequence $i_0, i_1, \ldots, i_m$ of states.

Further, it is also easy to see that

$$P\{X_3 = i_3, X_2 = i_2|X_1 = i_1, X_0 = i_0\} = p_{i_1i_2}p_{i_2i_3}.$$ 

More generally,

$$(22) \quad P\{X_{m+\ell} = j_\ell, 1 \leq \ell \leq n|X_s = i_s, 0 \leq s \leq m\} = p_{i_mj_1}p_{j_1j_2} \cdots p_{j_{n-1}j_n}.$$ 

With these preliminaries, we can define the $n$-step transition probability as

$$p^{(n)}_{ij} = P\{X_{m+n} = j|X_m = i\}.$$ 

If we now observe that

$$\{X_{m+n} = j\} = \cup_{j_0, \ldots, j_{n-1}} \{X_{m+n} = j, X_{m+n-1} = j_0, \ldots, X_{m+1} = j_{n-1}\},$$

we conclude

$$p^{(n)}_{ij} = \sum_{k_1 \cdots k_{n-1}} p_{i}k_{1}p_{k_{1}k_{2}} \cdots p_{k_{n-1}j}.$$
From matrix theory, we recognize \( p_{ij}^{(n)} \) as the \((i, j)\)th entry of \( \mathbb{P}^n \), the \( n \)th power of the transition matrix \( \mathbb{P} \). It is convenient to use the convention \( p_{ij}^{(0)} = \delta_{ij} \), consistent with the fact that \( \mathbb{P}^0 \) is the identity matrix \( I \).

Finally, it is straightforward to verify that

\[
p_{ij}^{(m+n)} = \sum_{v \in S} p_{iv}^{(m)} p_{vj}^{(n)},
\]

and that \( \sum_{j \in S} p_{ij}^{(n)} = 1 \). This means \( \mathbb{P}^n = \mathbb{P}^k \mathbb{P}^\ell \), whenever \( k + \ell = n \), which is consistent with taking powers of a matrix.

For convenience, we now define the conditional probabilities \( P_i(A) \triangleq P\{A|X_0 = i\} \), \( A \in \mathcal{F} \). With this notation and as an immediate consequence of (22), we have

\[
P_i\{X_1 = i_1, \ldots, X_m = j, X_{m+1} = j_{m+1}, \ldots, X_{m+n} = j_{m+n}\} = P_i\{X_1 = i_1, \ldots, X_m = j\} P_j\{X_1 = j_{m+1}, \ldots, X_n = j_{m+n}\}.
\]

Example 19. A gambling problem (see related homework problem)

Initial fortune of a gambler is \( X_0 = x_0 \). After playing each hand, he/she increases (decreases) the fortune by one dollar with probability \( p \) (\( q \)). The gambler quits if either her/his fortune becomes 0 (bankruptcy) or reaching a goal of \( L \) dollars. Let \( X_n \) denote the fortune at time \( n \). Note that

\[
P\{X_n = j | X_1 = i_1 \ldots X_{n-1} = i_{n-1}\} = P\{X_n = j | X_{n-1} = i_{n-1}\}.
\]

Hence, the sequence \( X_n \) is a Markov chain with state space \( S = \{0, 1, \ldots, L\} \). Also, the time-independent transition probabilities are given by

\[
p_{ij} = P\{X_{n+1} = j | X_n = i\} = \begin{cases} p, & \text{if } j = i + 1, \ i \neq L \\ q, & \text{if } j = i - 1, \ i \neq 0 \\ 1, & \text{if } i = j = L \text{ or if } i = j = 0 \end{cases}
\]
The \((L + 1) \times (L + 1)\) probability transition matrix, \(\mathbf{IP} \triangleq ((p_{ij}))\) is therefore

\[
\mathbf{IP} = \begin{bmatrix}
1 & 0 & 0 & \ldots & \ldots & \ldots & \ldots & \ldots & 0 \\
q & 0 & p & 0 & \ldots & \ldots & \ldots & \ldots & \vdots \\
0 & q & 0 & p & 0 & \ldots & \ldots & \ldots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \ldots & \ldots & \ldots & q & 0 & p \\
0 & 0 & \ldots & \ldots & \ldots & 0 & 0 & \ldots & 1
\end{bmatrix}.
\]

Note that the sum of any row of \(\mathbf{IP}\) is 1, a characteristic of a stochastic matrix.

**Exercise 2.** Show that \(\lambda = 1\) is always an eigenvalue for any stochastic matrix \(\mathbf{IP}\).

Let \(P(x)\) be the probability of achieving the goal of \(L\) dollars, where \(x\) is the initial fortune. In a homework problem we have shown that

\[
P(x) = pP(x + 1) + qP(x - 1), x = 1, 2, \ldots, L - 1,
\]

with boundary conditions \(P(0) = 0\) and \(P(L) = 1\). Similarly, define \(Q(x)\) as the probability of going bankrupt. Then,

\[
Q(x) = pQ(x + 1) + qQ(x - 1), x = 1, 2, \ldots, L - 1,
\]

with boundary conditions \(Q(0) = 1\) and \(Q(L) = 0\).

For example, if \(p = q = \frac{1}{2}\), then (see homework solutions)

\[
P(x) = \frac{x}{L},
\]

\[
Q(x) = 1 - \frac{x}{L}.
\]

Thus, \(\lim_{L \to \infty} P(x) = 0\). (what is the implication of this?).

**Exercise 3.** Show that

\[
\lim_{L \to \infty} P(x) = 0
\]

when \(q \neq p\).
Back to Gambler’s ruin problem: take $L = 4$, $p = 0.6$. Then,

$$\mathbf{P} = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0.4 & 0 & 0.6 & 0 & 0 \\
0 & 0.4 & 0 & 0.6 & 0 \\
0 & 0 & 0.4 & 0 & 0.6 \\
0 & 0 & 0 & 0 & 1
\end{bmatrix}. \tag{31}$$

Also, straightforward calculation yields

$$\mathbf{P}^2 = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0.4 & 0.24 & 0 & 0.36 & 0 \\
0.16 & 0.48 & 0 & 0.36 & 0 \\
0 & 0.16 & 0 & 0.24 & 0.6 \\
0 & 0 & 0 & 0 & 1
\end{bmatrix}, \text{ and} \tag{32}$$

$$\lim_{n \to \infty} \mathbf{P}^n = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0.5846 & 0 & 0 & 0 & 0.4154 \\
0.3077 & 0 & 0 & 0 & 0.6923 \\
0.1231 & 0 & 0 & 0 & 0.8769 \\
0 & 0 & 0 & 0 & 1
\end{bmatrix}. \tag{33}$$

(This can be obtained, for example, using the Cayley-Hamilton Theorem). Note that the last column is

$$\mathbf{Q} = \begin{bmatrix}
Q(0) \\
\vdots \\
Q(4)
\end{bmatrix}, \tag{34}$$

and the first column is

$$\mathbf{P} = \begin{bmatrix}
P(0) \\
\vdots \\
P(4)
\end{bmatrix}. \tag{35}$$
6.2. Transience and Recurrence. Suppose that $X_0 = i$ and define

$$E_{ij} \triangleq \bigcup_{n=1}^{\infty} \{X_0 = i, X_n = j\}$$

as the event that the chain eventually visits state $j$ provided that $X_0 = i$. A state $i$ is said to be recurrent if $P_i(E_{ii}) = 1$; otherwise, it is said to be transient.

Note that we can write $E_{ij}$ as the disjoint union

$$E_{ij} = \bigcup_{k=1}^{\infty} \{X_0 = i, X_k = j, X_{k-1} \neq j, \ldots, X_1 \neq j\}.$$

In the above, we are saying that the chain can eventually visit state $j$ precisely in the following mutually exclusive ways: visit $j$ for the first time in one step, or visit $j$ for the first time in two steps, and so on. If we define

$$f^{(n)}_{ij} \triangleq P_i\{X_1 \neq j, \ldots, X_{n-1} \neq j, X_n = j\}$$

as the probability of a first visit to state $j$ at time $n$ provided that $X_0 = i$, then the quantity

$$f_{ij} \triangleq \sum_{n=1}^{\infty} f^{(n)}_{ij}$$

is precisely $P_i(E_{ij})$. Hence, a state $i$ is recurrent precisely when $f_{ii} = 1$ and it is transient when $f_{ii} < 1$.

6.2.1. Visiting a state infinitely often. Suppose that $X_0 = i$ and consider the event $A_{jk}$ defined as the event that the chain visits state $j$ at least $k$ times. Note that if $A_{jk}$ occurs, then it must be true that the chain must have made a visit from $i$ to $j$ and then revisited $j$ at least $k - 1$ times. Realizing the $P_i$ probability that the chain visits state $j$ from state $i$ is $f_{ij}$ and the $P_i$ probability that the chain visits state $j$ from state $j$ ($k - 1$) times is $f_{jj}^{k-1}$, we conclude that $P_i(A_{jk}) \leq f_{ij} f_{jj}^{k-1}$. On the other hand, if the chain happens to visit $j$ from $i$ and also revisit $j$ at least $k - 1$ times, then we would know that the event $A_{jk}$ has occurred. Hence, $P_i(A_{jk}) = f_{ij} f_{jj}^{k-1}$.

Note that $A_{jk}$ is a decreasing sequence of events in $k$. Consequently,

$$P_i(\cap_k A_{jk}) = \lim_{k \to \infty} P_i(A_{jk}) = \begin{cases} 0, & \text{if } f_{jj} < 1 \\ f_{ij}, & \text{if } f_{jj} = 1 \end{cases}$$
However, $\bigcap_k A_{jk}$ is precisely the event $\{X_n = j \text{ i.o.}\}$. Hence, we arrive at the conclusion

$$
\text{(36)} \quad P_i\{X_n = j \text{ i.o.}\} = \begin{cases} 0, & \text{if } f_{jj} < 1 \\ f_{ij}, & \text{if } f_{jj} = 1 \end{cases}
$$

Taking $i = j$ yields

$$
\text{(37)} \quad P_i\{X_n = i \text{ i.o.}\} = \begin{cases} 0, & \text{if } f_{ii} < 1 \\ 1, & \text{if } f_{ii} = 1 \end{cases}
$$

Thus, $P_i\{X_n = i \text{ i.o.}\}$ is either 0 or 1!

We have therefore proved the following result.

**Theorem 17** (Adopted from Billingsley). Recurrence of $i$ is equivalent to $P_i\{X_n = i \text{ i.o.}\} = 1$ and transience of $i$ is equivalent to $P_i\{X_n = i \text{ i.o.}\} = 0$.

The next theorem further characterizes recurrence and transience.

**Theorem 18** (Adopted from Billingsley). Recurrence of $i$ is equivalent to $\sum_n p_{ii}^{(n)} = \infty$; transience of $i$ equivalent to $\sum_n p_{ii}^{(n)} < \infty$.

**Proof.**

Since $p_{ii}^{(n)} = P_i\{X_n = i\}$, by the Borel-Cantelli lemma $\sum_n p_{ii}^{(n)} < \infty$ implies $P_i\{X_n = i \text{ i.o.}\} = 0$. According to (37), this implies $f_{ii} < 1$. Hence, we have shown that $\sum_n p_{ii}^{(n)} < \infty$ implies transience. Consequently, recurrence implies that $\sum_n p_{ii}^{(n)} = \infty$.

We next prove that $f_{ii} < 1$ (or transience) implies $\sum_n p_{ii}^{(n)} < \infty$, which, in turn, implies “$\sum_n p_{ii}^{(n)} = \infty$ implies recurrence,” and the entire theorem will therefore be proved. Now back to proving “$f_{ii} < 1$ implies $\sum_n p_{ii}^{(n)} < \infty$.”

Observe that the chain visits $j$ from $i$ in $n$ steps precisely in the following mutually exclusive ways: visit $j$ from $i$ for the first time $n$ steps, visit $j$ from $i$ for the first time in $n - 1$ steps and then revisit $j$ in one step, visit $j$ from $i$ for the first time in $n - 2$
steps and then revisit \( j \) in two steps, and so on. More precisely, we write

\[
P^{(n)}_{ij} = \mathbb{P}_i \{ X_n = j \} = \sum_{s=0}^{n-1} \mathbb{P}_i \{ X_1 \neq j, \ldots, X_{n-s-1} \neq j, X_{n-s} = j, X_n = j \}
\]

\[
= \sum_{s=0}^{n-1} \mathbb{P}_i \{ X_1 \neq j, \ldots, X_{n-s-1} \neq j, X_{n-s} = j \} \mathbb{P}_j \{ X_s = j \}
\]

\[
= \sum_{s=0}^{n-1} f^{(n-s)}_{ij} p^{(s)}_{jj}.
\]

Put \( j = i \) and summing up over \( n = 1 \) to \( n = \ell \) we obtain,

\[
\sum_{n=1}^{\ell} p^{(n)}_{ii} = \sum_{n=1}^{\ell} \sum_{s=0}^{n-1} f^{(n-s)}_{ii} p^{(s)}_{ii}
\]

\[
= \sum_{s=0}^{\ell-1} p^{(s)}_{ii} \sum_{n=s+1}^{\ell} f^{(n-s)}_{ii} \leq f_{ii} \sum_{s=0}^{\ell} p^{(s)}_{ii}
\]

The last inequality comes from the fact that \( \sum_{n=s+1}^{\ell} f^{(n-s)}_{ii} = \sum_{u=1}^{\ell-s} f^{(u)}_{ii} \leq f_{ii} \). Realizing that \( p^{(0)}_{ii} = 1 \), we conclude \( (1 - f_{ii}) \sum_{n=1}^{\ell} p^{(n)}_{ii} \leq f_{ii} \). Hence, if \( f_{ii} < 1 \), then \( \sum_{n=1}^{\ell} p^{(n)}_{ii} \leq f_{ii} / (1 - f_{ii}) \), and the series \( \sum_{n=1}^{\infty} p^{(n)}_{ii} \) will therefore be convergent since the partial sum is increasing and bounded. \( \square \)

**Exercise 4.** Show that in the gambler’s problem discussed earlier, states 0 and \( L \) are recurrent, but states 1, \ldots, \( L - 1 \) are transient.

6.3. Irreducible chains. A MC is said to be irreducible if for every \( i, j \in S \), there exists an \( n \) (that is generally dependent up on \( i \) and \( j \)) such that \( p^{(n)}_{ij} > 0 \). In words, if the chain is irreducible then we can always go from any state to any other state in finite time with a nonzero probability. The next theorem shows that for irreducible chains, either all states are recurrent or all states are transient, there is no option in between. Thus, we can say that an irreducible MC is either recurrent or transient.

**Theorem 19** (Adopted from Billingsley). If the Markov chain is irreducible, then one of the following two alternatives holds.

(i) All states are recurrent, \( \mathbb{P}_i(\cap_j \{ X_n = j \ i.o. \}) = 1 \) for all \( i \), and \( \sum_n p^{(n)}_{ij} = \infty \) for all \( i \) and \( j \).
(ii) **All states are transient**, \( P_i(\cup_j \{X_n = j \text{ i.o.}\}) = 0 \) for all \( i \), and \( \sum_n p_{ij}^{(n)} < \infty \) for all \( i \) and \( j \).

*Proof.*

By irreducibility, for each \( i \) and \( j \) there exist integers \( r \) and \( s \) so that \( p_{ij}^{(r)} > 0 \) and \( p_{ji}^{(s)} > 0 \). Note that if a chain with \( X_0 = i \) visits \( j \) from \( i \) in \( r \) steps, then visits \( j \) from \( j \) in \( n \) steps, and then visits \( i \) from \( j \) in \( s \) steps, then it has visited \( i \) from \( i \) in \( r + n + s \) steps. Therefore,

\[
p_{ii}^{(r+s+n)} \geq p_{ij}^{(r)} p_{jj}^{(s)} p_{ji}^{(s)}.
\]

Now if we sum up both sides of the above inequality over \( n \) and use the fact that \( p_{ij}^{(r)} p_{ji}^{(s)} > 0 \), we conclude that \( \sum_n p_{ii}^{(n)} < \infty \) implies \( \sum_n p_{jj}^{(n)} < \infty \) (why do we need the fact that \( p_{ij}^{(r)} p_{ji}^{(s)} > 0 \) to come up with this conclusion?) In particular, if one state is transient, then all other states are also transient (i.e., \( f_{ij} < 1 \) for all \( j \)). In this case, (36) tells us that \( P_i(\{X_n = j \text{ i.o.}\}) = 0 \) for all \( i \) and \( j \). Therefore,

\[
P_i(\cup_j \{X_n = j \text{ i.o.}\}) \leq \sum_j P_i(\{X_n = j \text{ i.o.}\}) = 0 \text{ for all } i.
\]

Next, note that \( \sum_{n=1}^{\infty} p_{ij}^{(n)} = \sum_{n=1}^{\infty} \sum_{v=1}^{n} f_{ij}^{(v)} p_{jj}^{(n-v)} = \sum_{v=1}^{\infty} f_{ij}^{(v)} \sum_{m=0}^{\infty} p_{jj}^{(m)} \leq \sum_{m=0}^{\infty} p_{jj}^{(m)}, \)

where the inequality follows from the fact that \( \sum_{v=1}^{\infty} f_{ij}^{(v)} = f_{ij} \leq 1 \). Hence, if \( j \) is transient (in which case \( \sum_n p_{jj}^{(n)} < \infty \) according to Theorem 18), then \( p_{ij}^{(n)} < \infty \) for all \( i \). We have therefore established the second alternative of the theorem.

If any state is not transient, then the only possibility is that *all* states are recurrent. In this case \( P_j(\{X_n = j \text{ i.o.}\}) = 1 \) by Theorem 17 and \( \{X_n = j \text{ i.o.}\} \) is an almost certain event for all \( j \) with respect to \( P_j \). Namely, \( P_j(A \cap \{X_n = j \text{ i.o.}\}) = P_j(A) \) for any \( A \in \mathcal{F} \). Hence, we can write

\[
p_{ji}^{(m)} = P_j(\{X_m = i\} \cap \{X_n = j \text{ i.o.}\}) \leq \sum_{n>m} P_j(\{X_m = i, X_{m+1} \neq j, \ldots, X_{n-1} \neq j, X_n = j\}) = \sum_{n>m} p_{ji}^{(m)} f_{ij}^{(n-m)} = p_{ji}^{(m)} f_{ij}.
\]

By irreducibility, there is an \( m_0 \) for which \( p_{ji}^{(m_0)} > 0 \); hence, \( f_{ij} = 1 \). Now by (37), \( P_i(\{X_n = j \text{ i.o.}\}) = 1 \). The only item left to prove is that \( \sum_n p_{ij}^{(n)} \) under the second
alternative. Note that if \( \sum_n p_{ij}^{(n)} < \infty \) for some \( i \) and \( j \), then the Borel-Cantelli lemma would dictate that \( P_i\{X_n = j \text{ i.o.}\} = 0 \), and the chain would not be recurrent, which is a contradiction. Hence, under the second alternative, \( \sum_n p_{ij}^{(n)} = \infty \) for all \( i \) and \( j \).

\[ \square \]

**Special Case:** Suppose that \( \sum_n p_{ij}^{(n)} < \infty \) and \( S \) is a finite set. Then \( \sum_{j \in S} \sum_n p_{ij}^{(n)} = \sum_n \sum_{j \in S} p_{ij}^{(n)} \). But since \( \sum_{j \in S} p_{ij}^{(n)} = 1 \), we obtain \( \sum_{j \in S} \sum_n p_{ij}^{(n)} = \sum_n 1 = \infty \), which is a contradiction. Hence, in a finite irreducible MC, alternative (ii) is impossible; hence, *every finite state-space irreducible MC is recurrent*. Consequently, if the transition matrix of a finite state-space MC has all positive entries (making the chain irreducible), then the chain is recurrent.

**Random walk example** (from Billingsley).

Pólya’s theorem. For the symmetric \( k \)-dimensional random walk, all states are recurrent if \( k = 1 \) or \( k = 2 \), and all states are transient if \( k \geq 3 \). We will only give the proof for the case \( k = 1 \). Note first that \( p_{ii}^{(n)} \) is the same for all states \( i \). Clearly, \( p_{ii}^{(2n+1)} = 0 \). Since return in \( 2n \) steps means \( n \) steps right and \( n \) steps left,

\[
p_{ii}^{(2n)} = \binom{2n}{n} \frac{1}{2^{2n}}.
\]

By Stirling’s formula, \( p_{ii}^{(2n)} \sim (\pi n)^{-1/2} \). Therefore, \( \sum_n p_{ii}^{(n)} = \infty \), and all states are recurrent by Theorem 19.

**6.4. Birth and death Markov chains.** The material here is extracted from the book by Hoel *et al.* Birth and death chains are common examples of Markov chains; they include random walks. Consider a Markov chain \( X_n \) on the set of infinite or finite nonnegative integers \( S = \{0, \ldots, d\} \) (in the case of an infinite set we take \( d = \infty \)). The transition function is of the form

\[
p_{ij} = \begin{cases} 
q_i, & j = i - 1, \\
r_i, & j = i, \\
p_i, & j = i + 1
\end{cases}
\]
Note that \( p_i + q_i + r_i = 1 \), \( q_0 = 0 \), and \( p_d = 0 \) if \( d < \infty \). We also assume that \( p_i \) and \( q_i \) are positive for \( 0 < i < d \). With these assumptions, the chain can be shown to be irreducible. Hence, the question is whether such a birth or death chain is transient or recurrent. We will address this question next.

For any \( i \in \mathcal{S} \), we define \( T_i \triangleq \min\{ n : X_n = i \} \) as the time to the first entrance (or first hitting time) to state \( i \), and assume that \( X_0 = 1 \). If state 0 is recurrent (i.e., \( f_{00} = 1 \)), then according to (36), \( P_1(X_n = 0 \text{ i.o.}) = f_{10} \). But according to Theorem 19, \( P_1(X_n = 0 \text{ i.o.}) = 1 \), so \( f_{10} = 1 \). Therefore, since \( f_{10} = P_1(T_0 < \infty) \), we conclude that if state 0 is recurrent then \( P_1(T_0 < \infty) = 1 \). We will next show precisely when \( P_1(T_0 < \infty) = 1 \).

Following Hoel et al., for \( a \) and \( b \) in \( \mathcal{S} \) such that \( a < b \), define the probabilities

\[
u(i) = P_i(T_a < T_b), \quad a < i < b,
\]

with \( \nu(a) = 1 \) and \( \nu(b) = 0 \). Recall the transition probabilities \( q_j, r_j, \) or \( p_j \) and note that we can use conditional expectations to write

\[
u(j) = q_j \nu(j - 1) + r_j \nu(j) + p_j \nu(j + 1), \quad a < j < b.
\]

Since \( r_j = 1 - p_j - q_j \), we can rewrite the above equation as

\[
u(j + 1) - \nu(j) = \frac{q_j}{p_j} (\nu(j) - \nu(j - 1)), \quad a < j < b.
\]

For convenience, define \( \gamma_0 = 1 \) and

\[
\gamma_j = \frac{q_1 \cdots q_j}{p_1 \cdots p_j}, \quad 0 < j < d
\]

and use this to rewrite the above recursion as

\[
u(j) - \nu(j + 1) = \gamma_j \gamma_a (\nu(a) - \nu(a + 1)), \quad a \leq j < b.
\]

Now sum up the above equation over \( j = a, \ldots, b - 1 \), use \( \nu(a) = 1 \) and \( \nu(b) = 0 \), and obtain

\[
\frac{\nu(a) - \nu(a + 1)}{\gamma_a} = \frac{1}{\sum_{j=a}^{b-1} \gamma_j}.
\]
Thus, (38) becomes

\[ u(j) - u(j + 1) = \frac{\gamma_j}{\sum_{j=a}^{b-1} \gamma_j}, \quad a \leq j < b. \]

Let us now sum up this equation over \( j = i, \ldots, b - 1 \) and use the fact that \( u(b) = 0 \) to obtain

\[ u(i) = \frac{\sum_{j=i}^{b-1} \gamma_j}{\sum_{j=a}^{b-1} \gamma_j}, \quad a < i < b. \]

It now follows from the definition of \( u(i) \) that

\[ P_i(T_a < T_b) = \frac{\sum_{j=i}^{b-1} \gamma_j}{\sum_{j=a}^{b-1} \gamma_j}, \quad a < i < b. \]

or

\[ P_i(T_b < T_a) = \frac{\sum_{j=a}^{i-1} \gamma_j}{\sum_{j=a}^{b-1} \gamma_j}, \quad a < i < b. \]

As a special case of (39),

\[ P_1(T_0 < T_n) = 1 - \frac{1}{\sum_{j=0}^{n-1} \gamma_j}, \quad n > 1. \]

Note that

\[ 1 \leq T_2 < T_3 < \ldots, \]

which implies \( \{T_0 < T_n\}, \ n > 1, \) are nondecreasing events. Hence,

\[ \lim_{n \to \infty} P_1(T_0 < T_n) = P_1(\bigcup_{n=1}^{\infty} T_0 < T_n). \]

Equation (41) implies \( T_n \geq n, \) and thus \( T_n \to \infty \) as \( n \to \infty. \) Hence,

\[ \bigcup_{n=1}^{\infty} \{T_0 < T_n\} = \{T_0 < \infty\}. \]

We can therefore rewrite (42) as

\[ \lim_{n \to \infty} P_1(T_0 < T_n) = P_1(T_0 < \infty). \]

Combining this with (40) yields

\[ P_1(T_0 < \infty) = 1 - \frac{1}{\sum_{j=0}^{\infty} \gamma_j}. \]
We now show that the birth and death chain is recurrent if and only if
\[ \sum_{j=0}^{\infty} \gamma_j = \infty. \]

From our earlier discussion, we know that if the irreducible birth and death chain is recurrent, then \( P_1(T_0 < \infty) = 1 \), in which case (43) would imply \( \sum_{j=0}^{\infty} \gamma_j = \infty \). On the other hand, suppose that \( \sum_{j=0}^{\infty} \gamma_j = \infty \), in which case (43) implies \( P_1(T_0 < \infty) = 1 \).

Note that we can write (using conditional expectations)
\[ f_{00} = P_0(T_0 < \infty) = p_{00} + p_{01}P_1(T_0 < \infty) = p_{00} + p_{01} = 1, \]
and “0” is therefore a recurrent state.

6.5. **Stationary distributions.** A stationary distribution is a vector \( \pi = [\pi_0, \pi_1, \ldots, \pi_n] \) that is a solution to the equation \( \pi P = \pi \) (or equivalently \( \sum_{i} \pi_i p_{ij} = \pi_j \)) such that \( \pi_i \geq 0 \) and \( \sum_{i} \pi_i = 1 \). Note that if \( \pi \) is a stationary distribution, then \( \pi P^n = \pi \).

That is, the distribution of \( X_n \) remains as \( \pi \) so long as we select the distribution of \( X_0 \) to be \( \pi \). This is why we call such \( \pi \) a stationary; the distribution becomes invariant in time. The main result here is from Theorem 8.6 in Billingsley, which we reiterate without proof.

**Theorem 20** (From Billingsley). Suppose that for an irreducible, aperiodic chain that there exists a stationary distribution \( \pi \). Then the chain is recurrent,
\[ \lim_{n \to \infty} P_{ij}^{(n)} = \pi_j \]
for all \( i \) and \( j \), the \( \pi_i \) are all positive, and the stationary distribution is unique.

The main point of the conclusion is that the effect of the initial state wears off. Also, \( \lim_{n \to \infty} P^n \) converges to a matrix whose rows are identical to \( \pi \).

6.5.1. **Stationary distributions for birth and death chains.** Consider a birth and death chain on the nonnegative integers. We further assume that the chain is irreducible, i.e., \( p_i > 0 \) for \( i \geq 0 \) and \( q_i > 0 \) for \( i \geq 1 \).

The system of equations \( \sum_{i} \pi_i p_{ij} = \pi_j \) for the stationary distribution yields
\[ \pi_0 r_0 + \pi_1 q_1 = \pi_0, \]
\[ \pi_{j-1} p_{j-1} + \pi_j r_j + \pi_{j+1} q_{j+1} = \pi_j, \quad j \geq 1. \]

Using \( p_j + q_j + r_j = 1 \), we obtain

\[ q_1 \pi_1 - p_0 \pi_0 = 0, \]
\[ q_{j+1} \pi_{j+1} - p_j \pi_j = q_j \pi_j - p_{j-1} \pi_{j-1}, \quad j \geq 1. \]

By induction, we obtain

\[ q_{j+1} \pi_{j+1} - p_j \pi_j = 0, \quad j \geq 0, \]

and hence

\[ \pi_{j+1} = \frac{p_j}{q_{j+1}} \pi_j, \quad j \geq 0. \]

Consequently, we obtain

\[ \pi_i = \frac{p_0 \cdots p_{i-1}}{q_1 \cdots q_i} \pi_0, \quad i \geq 1. \]

If we define

\[ \tilde{\pi}_i \triangleq \begin{cases} 1, & i = 0, \\ \frac{p_0 \cdots p_{i-1}}{q_1 \cdots q_i}, & i \geq 1 \end{cases} \]

we can recast (44) as

\[ \pi_i = \tilde{\pi}_i \pi_0, \quad i \geq 0. \]

Realizing that \( \sum_i \pi_i < \infty \) if and only if

\[ \sum_{i=1}^{\infty} \frac{p_0 \cdots p_{i-1}}{q_1 \cdots q_i} < \infty, \]

we can say that if \( \sum_{i=1}^{\infty} \frac{p_0 \cdots p_{i-1}}{q_1 \cdots q_i} < \infty \), then the birth and death chain has a unique (why is it unique?) stationary distribution, given by

\[ \pi_i = \frac{\tilde{\pi}_i}{\sum_{j=0}^{\infty} \tilde{\pi}_j}, \quad i \geq 0. \]

On the other hand, if \( \sum \tilde{\pi}_i = \infty \), no stationary distribution exists (why?). If the irreducible birth and death chain is finite, then the above formula for the stationary distribution still holds with \( \infty \) replaced by \( d \).
7. Properties of Covariance Matrices and Data Whitening

7.1. Preliminaries. For a random vector \( \mathbf{X} = [X_1 \ldots X_n]^T \), the covariance matrix is \( \mathbf{C}_X \) is an \( n \times n \) matrix = \( \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T] \). This matrix has the following properties:

1. \( \mathbf{C}_X \) is symmetric, which is clear from the definition.
2. There is a matrix \( \Phi \) such that \( \Phi^T \mathbf{C}_X \Phi = \Lambda \), where \( \Lambda \) is a diagonal matrix whose diagonal elements are the eigenvalues, \( \lambda_1 \ldots \lambda_n \), of the matrix \( \mathbf{C}_X \) (assume that they are distinct). The corresponding eigenvectors, \( \mathbf{v}_1 \ldots \mathbf{v}_n \), form the columns of the matrix \( \Phi \): \( \Phi = [\mathbf{v}_1 | \mathbf{v}_2 | \ldots | \mathbf{v}_n] \). This fact comes from elementary linear algebra. Because \( \mathbf{C}_X \) is symmetric, the eigenvectors corresponding to distinct eigenvalues are orthogonal (prove this). Moreover, \( \Phi \) can be selected so that the eigenvectors are orthonormal; that is, \( \mathbf{v}_i^T \mathbf{v}_j = \delta_{i,j} \).
3. The fact that the \( \mathbf{v}_i \)'s are orthonormal implies that they span \( \mathbb{R}^n \) and we also obtain the following representation for \( \mathbf{X} \):
   \[
   \mathbf{X} = \sum_{i=1}^{n} c_i \mathbf{v}_i,
   \]
   with \( c_i = \mathbf{X}^T \mathbf{v}_i \).
4. \( \mathbf{C}_X \) is positive semi-definite, which means \( \mathbf{x}^T \mathbf{C}_X \mathbf{x} \geq 0 \) for any \( \mathbf{x} \in \mathbb{R}^n \). To see this property, note that \( \mathbf{x}^T \mathbf{C}_X \mathbf{x} = \mathbf{x}^T \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T] \mathbf{x} = \mathbb{E}[(\mathbf{x}^T(\mathbf{X} - \mathbb{E}[\mathbf{X}]))^2] \geq 0 \). This calculation also shows as long as no component of \( \mathbf{X} \) can be written as a linear combination of other components, then \( \mathbf{x}^T \mathbf{C}_X \mathbf{x} > 0 \), which is the defining property of positive definite matrices. (Recall that for any random variable \( Y \), \( \mathbb{E}[Y^2] = 0 \) if and only if \( Y = 0 \) almost surely.)

In the homework, we looked at an estimate of \( \mathbf{C}_X \) from \( K \) samples of the random vector \( \mathbf{X} \) as follows:

\[
\hat{\mathbf{C}}_X = \frac{1}{K-1} \sum_{i=1}^{K} \mathbf{X}^{(i)} \mathbf{X}^{(i)T}
\]

(assuming, without loss of generality, that \( \mathbb{E} [\mathbf{X}] = 0 \)), which has rank at most \( k \) (see homework problems). Therefore, if \( n > K \), then \( \mathbf{C}_X \) is not full rank and hence \( \hat{\mathbf{C}}_X \) is not positive definite.
7.2. Whitening of Correlated Data. We say that the components $X_1 \ldots X_n$ are \textit{white} if $\mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] = \delta_{ij}$ for all $i, j$. There is an $n \times n$ matrix, $A$, such that if $Y = AX$, then $C_Y = I$, the identity matrix.

To see this, put $B = \Phi \Lambda^{1/2}$, where

\begin{equation}
\Lambda = \begin{bmatrix}
\lambda_1 & \ldots & 0 \\
\vdots & \ddots & \vdots \\
0 & \ldots & \lambda_n
\end{bmatrix},
\end{equation}

and as before $\lambda_i$ is the $i$th eigenvalue of $C_X$. Recall that $\Phi^T C_X \Phi = \Lambda$. Observe that $B(\Lambda^{-1/2} \Phi^T) = \Phi \Lambda^{1/2} \Lambda^{-1/2} \Phi^T = \Phi \Phi^T = \Phi \Phi^T = I$. Hence, $B^{-1} = \Lambda^{-1/2} \Phi^T$.

Now take $A = B^{-1}$. Let $Y = AX$, then $C_Y = AC_X A^T = \Lambda^{-1/2} \Phi^T C_X \Phi \Lambda^{-1/2} = \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = I$, which is the desired result.

Let us have better insight into this transformation. Note that

\begin{equation}
Y = AX = \Lambda^{-1/2} \Phi^T X = \begin{bmatrix}
\frac{1}{\sqrt{\lambda_1}} v_1^T X \\
\vdots \\
\frac{1}{\sqrt{\lambda_n}} v_n^T X
\end{bmatrix}.
\end{equation}

If $Y_i$ is the $i$th component of $Y$, then, $Y_i = \frac{1}{\sqrt{\lambda_i}} \sum_{j=1}^n (v_i)_j X_j$. Therefore, we can think of $A$ as a linear filter with $h_{ij} = \frac{1}{\sqrt{\lambda_i}} (v_i)_j$ as its impulse response, whereby $Y_i = \sum_{j=1}^n h_{ij} X_j$.

![Figure 7.](image)

Now recall the representation $X = \sum_{k=1}^n (v_k^T X) v_k$ and use the fact $Y_k = \frac{\sqrt{X}}{\sqrt{\lambda_k}}$ to obtain $X = \sum_{k=1}^n \sqrt{\lambda_k} Y_k v_k$. This is the \textit{reconstruction formula} from whitened data.

7.3. Simultaneous Diagonalization of Two Covariance Matrices. In addition to $C_X$, suppose that $C_Z$ is a second covariance matrix corresponding to a random vector $Z$. Our goal is to find a transformation $\Gamma$ that simultaneously diagonalizes $C_X$ and $C_Z$. 

To begin, we may want to try $A$ again. We know that $A$ whitens $X$. Put $T = AZ$ and note that $C_T = AC_ZA^T = \Lambda^{-1/2}\Phi^T C_Z\Phi \Lambda^{-1/2}$, which is not necessarily a diagonal matrix (show this by example). Note, however, that if $W$ is a matrix consisting of the eigenvectors of $C_T$, corresponding to the eigenvalues $\eta_1, \ldots, \eta_n$, then $W^T C_T W = M$, where $M$ is the diagonal matrix whose diagonal elements are $\eta_1, \ldots, \eta_n$.

Consider $\Gamma = W^TA$ as a candidate for the desired transformation. We need to check if this new transformation diagonalizes both $C_X$ and $C_Z$. To do so, consider $D = \Gamma X$ and note that $C_D = W^T A C_X A^T W = W^T I W = W^T W = I$ (because $W$ is normal matrix, i.e., $W^{-1} = W^T$).

Next, let $F = \Gamma Z$ and write $C_F = \Gamma C_Z \Gamma^T W = W^T A C_Z A^T W$. But, $A C_Z A^T = C_T$. Thus, $C_F = W^T G W = M$.

7.3.1. Summary. Let $C_X$ and $C_Z$ be given. Let $B$ be as before and let $W$ be the matrix whose $i$th column is the eigenvector of $B^{-1} C_Z B$ corresponding to the $i$th eigenvalue. Let $\Gamma = W^T B^{-1}$, then $Y = \Gamma X$ and $T = \Gamma Z$ have the following properties:

1. $C_Y = I$,
2. $C_T = M$

\[
\begin{bmatrix}
\eta_1 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & \eta_n
\end{bmatrix},
\]

(47)

where the $\eta_i$'s are the eigenvalues of $B^{-1} C_Z B$.

Example 20.

\[
C_X = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}; C_Z = \begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}.
\]

Find the transformation $\Gamma$.

We first we calculate $\Phi$. Set $|C_X - \lambda I| = 0$. We find that $\lambda_1 = 1$ and $\lambda_2 = 3$. Next, we find $v_1$ and $v_2$ using $C_X v_1 = \lambda_1 v_1$ and $C_X v_2 = \lambda_2 v_2$. Normalize these so
that $\|v_1\| = \|v_2\| = 1$ and obtain

$$v_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix}; v_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{49}$$

We now form

$$B^{-1} = \Lambda^{-1/2} \Phi^T = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix}. \tag{50}$$

Next,

$$B^{-1}C_Z (B^{-1})^T = \frac{1}{2} \begin{bmatrix} 8 & 0 \\ 0 & 4/3 \end{bmatrix}, \tag{51}$$

and its eigenvalues are $\eta_1 = 4$ and $\eta_2 = \frac{2}{3}$, and the corresponding eigenvectors are $[1, 0]$ and $[0, 1]$. Hence, $W = I$ and $\Gamma = W^T B^{-1} = B^{-1}$. 
8. Continuous-time Wiener Filtering

Suppose that we have the observation model

\[ X(t) = S(t) + N(t), \]

where \( S(t) \) and \( X(t) \) are wide-sense stationary random processes. We desire to find a physically realizable (causal) filter \( f(t) \) with impulse response \( h(t) \) to minimize the mean-square error (MSE)

\[ E[\epsilon^2] = E[(S(t + \alpha) - \hat{S}(t))^2], \]

where \( \hat{S}(t) = h(t) \ast X(t) \) is the filter output.

If \( \alpha = 0 \), this problem is called the Wiener filtering problem; if \( \alpha > 0 \), it is called the Wiener prediction problem, and if \( \alpha < 0 \), it is called the Wiener smoothing problem. Let’s characterize the solution to this problem.

We begin by writing

\[
E[\epsilon^2] = E[|S(t + \alpha) - \int_0^\infty h(\tau)X(t - \tau)d\tau|^2] \\
= E[S^2(t + \tau)] - 2 \int_0^\infty h(\tau)E[S(t + \alpha)X(t - \tau)]d\tau \\
+ \int_0^\infty \int_0^\infty h(\tau)h(\sigma)E[X(t - \tau)X(t - \sigma)]d\tau d\sigma \\
= R_{SS}(0) - 2 \int_0^\infty h(\tau)R_{XS}(\tau + \alpha)d\tau + \int_0^\infty \int_0^\infty h(\tau)h(\sigma)R_{XX}(\sigma - \tau)d\tau d\sigma.
\]

Suppose that there exists an optimum filter denoted by \( h_0(\tau) \), and consider a variation from that optimal filter defined as \( h_\nu(\tau) = h_0(\tau) + rg(\tau) \), where \( g \) is some arbitrary function and \( r \geq 0 \). The condition \( \frac{\partial}{\partial r} E[\epsilon^2(0)] = 0 \) is a necessary condition for the optimality (but not sufficient), where \( \epsilon^2(r) \) is the error associated with the
filter \( h_r \). Thus, we can pursue this differentiation approach and write

\[
E[\epsilon^2(r)] = R_{SS}(0) - 2 \int_0^\infty [h_0(\tau) + rg(\tau)] R_{XS}(\tau + \alpha) d\tau \\
+ \int_0^\infty \int_0^\infty [h_0(\tau) + rg(\tau)] [h_0(\sigma) + rg(\sigma)] R_{XX}(\sigma - \tau) d\tau d\sigma \\
= R_{SS}(0) - 2 \int_0^\infty h_0(\tau) R_{XS}(\tau + \alpha) d\tau - 2 \int_0^\infty r g(\tau) R_{XS}(\tau + \alpha) d\tau \\
+ \int_0^\infty \int_0^\infty [rg(\tau) h_0(\sigma) + rg(\sigma) h_0(\tau)] R_{XX}(\sigma - \tau) d\tau d\sigma \\
+ \int_0^\infty \int_0^\infty [h_0(\tau) h_0(\sigma)] R_{XX}(\sigma - \tau) d\tau d\sigma + \int_0^\infty \int_0^\infty [r^2 g(\tau) g(\sigma)] R_{XX}(\sigma - \tau) d\tau d\sigma.
\]

Note that \( R_{SS}(0) \) does not depend on \( r \). Now take partial derivative with respect to \( r \) and obtain

\[
\frac{\partial}{\partial r} E[\epsilon^2(r)] = -2 \int_0^\infty R_{XS}(\tau + \alpha) g(\tau) d\tau \\
+ \int_0^\infty \int_0^\infty [g(\tau) h_0(\sigma) + g(\sigma) h_0(\tau)] R_{XX}(\sigma - \tau) d\tau d\sigma \\
+ \int_0^\infty \int_0^\infty g(\tau) g(\sigma) R_{XX}(\sigma - \tau) d\tau d\sigma.
\]

Set \( \frac{\partial}{\partial r} E[\epsilon^2(0)] = 0 \) and obtain

\[
-2 \int_0^\infty R_{XS}(\tau + \alpha) g(\tau) d\tau + \int_0^\infty \int_0^\infty [g(\tau) h_0(\sigma) + g(\sigma) h_0(\tau)] R_{XX}(\sigma - \tau) d\tau d\sigma = 0.
\]

Using the fact that \( R_{XX}(\tau) \) is symmetric, i.e.,

\[
\int_0^\infty \int_0^\infty g(\tau) h_0(\sigma) R_{XX}(\sigma - \tau) d\tau d\sigma = \int_0^\infty \int_0^\infty g(\sigma) h_0(\tau) R_{XX}(\sigma - \tau) d\tau d\sigma,
\]

we get

\[
\int_0^\infty g(\tau) [-R_{XS}(\tau + \alpha) + \int_0^\infty h_0(\sigma) R_{XX}(\sigma - \tau) d\sigma] d\tau = 0.
\]

Since this is true for all \( g \), the only way this can be true is that the multiplier of \( g \) inside the outer integral must be zero for almost all \( \tau \geq 0 \) (with respect to Lebesgue measure).

Thus, we get the so-called Wiener Integral Equation:

\[
R_{XS}(\tau + \alpha) = \int_0^\infty R_{XX}(\tau - \sigma) h_0(\sigma) d\sigma, \quad \tau \geq 0
\]
8.1. **Non-causal Wiener Filtering.** If we replace \( \int_0^\infty \) with \( \int_{-\infty}^\infty \) in Weiner’s Equation, i.e.,

\[
R_{XS}(\tau + \alpha) = \int_{-\infty}^{\infty} R_{XX}(\tau - \sigma) h_0(\sigma) d\sigma, \quad \tau \in \mathbb{R},
\]
we obtain the characterization of the optimal non-causal (non-realizable) filter. In the frequency domain, we have

\[
S_{XS}(\omega)e^{j\omega\alpha} = S_{XX}(\omega)H_0(\omega),
\]
and hence

\[
H_0(\omega) = \frac{S_{XS}(\omega)e^{j\omega\alpha}}{S_{XX}(\omega)},
\]
which is the optimum non-realizable filter.

Suppose the signal and noise are uncorrelated, then

\[
R_{XX}(\tau) = E[X(t)X(t+\tau)] = E[(S(t)+N(t))(S(t+\tau)+N(t+\tau))] = R_{SS}(\tau) + R_{NN}(\tau).
\]
Similarly, \( R_{XS}(\tau) = R_{SS}(\tau) \). Then,

\[
H_0(\omega) = \frac{S_{SS}(\omega)}{S_{SS}(\omega)+S_{NN}(\omega)}
\]

**Example:** Suppose that \( N(t) \equiv 0 \) and \( \alpha > 0 \) (ideal prediction problem), then \( H_0(\omega) = e^{j\omega\alpha}[1] \), or \( h_0(t) = \delta(t + \alpha) \), which is the trivial non-realizable predictor.

Before we attempt solving the Wiener integral equation for the realizable optimal filter, we will review germane aspects of spectral factorization.

8.2. **Review of Spectral Factorization.** Note that we can approximate any power spectrum as white noise passed through a linear time-invariant (LTI) filter (why?). Suppose the transfer function of the filter is

\[
H(\omega) = \frac{\omega^n + a_{n-1}\omega^{n-1} + \ldots + a_1\omega + a_0}{\omega^d + b_{d-1}\omega^{d-1} + \ldots + b_1\omega + b_0}
\]
If the input is white noise with power spectrum \( \frac{N_0}{2} \), the output power spectrum is \( \frac{N_0}{2}|H(\omega)|^2 \). Note that we can write

\[
H(\omega) = \frac{A(\omega^2) + j\omega B(\omega^2)}{C(\omega^2) + j\omega D(\omega^2)},
\]
where $A$, $B$, $C$ and $D$ are polynomials with real coefficients. Thus,
\[
|H(\omega)|^2 = \frac{A^2(\omega^2) + \omega^2 B^2(\omega^2)}{C^2(\omega^2) + \omega^2 D^2(\omega^2)}.
\]

Observe that the output power spectrum will have only real coefficients in its numerator and denominator polynomials.

Now recall that the roots of polynomials with real coefficients are either real or complex-conjugate pairs. Consider the complex variable $s = \sigma + j\omega$. If $s_k = \sigma_k + j\omega_k$ is a root, then $s_k^* = \sigma_k - j\omega_k$ is also a root. Since the polynomials are even in $\omega$, we also know that if $s_k = \sigma_k + j\omega_k$ is a root, then $-s_k = -\sigma_k - j\omega_k$ is also a root.

Now replace everywhere $j\omega$ appears (in the expression for the output power spectrum) with $s$. Then write
\[
S(s) = S(\omega)|_{j\omega=s} = \frac{\prod_p(s - \lambda_p)}{\prod_k(s - \lambda_k)} \equiv S^+(s)S^-(s),
\]
where $S^+(\lambda)$ has all its zeroes and poles in the left-hand plane (LHP) and $S^-(s)$ has all its zeroes and poles in the right-hand plane (RHP). Furthermore, $|S^+(s)| = |S^-(s)^*|$ since poles and zeroes occur in complex-conjugate pairs. Finally, $S(s) = S^+(s)S^-(s) = |S^\pm(s)|^2$.

**Example:**
\[
S(\omega) = \frac{2\omega^2 + 3}{\omega^2 + 1} = \left[\frac{\sqrt{2}(j\omega + \sqrt{3}/2)}{j\omega + 1}\right]\left[\frac{\sqrt{2}(j\omega - \sqrt{3}/2)}{j\omega - 1}\right]
\]
That is,
\[
S^+(s) = \frac{\sqrt{2}(s + \sqrt{3}/2)}{s + 1} \quad S^-(s) = \frac{\sqrt{2}(s - \sqrt{3}/2)}{s - 1}
\]

We next investigate the connection between causality and the location of poles.

Suppose that $f(t)$ is a function that vanishes on the negative half time, i.e., $f(t) = 0$ for $t < 0$ and suppose that $\int_{-\infty}^{\infty} |f(t)| dt < \infty$. Then, the magnitude of the Fourier transform of this function is $|F(\omega)| = \int_{-\infty}^{\infty} e^{-j\omega t} f(t) dt \leq \int_{0}^{\infty} |e^{-j\omega t} f(t)| dt < \infty$, $\forall \omega$.

Let $s = \sigma + j\omega$ and observe that the magnitude of the Laplace transform has the property that $|F(s)| \leq \int_{0}^{\infty} |f(t)e^{-\sigma t} e^{-j\omega t}| dt < \infty$; thus, $F$ has no singularities in the RHP.
Similarly, if \( f(t) = 0 \) for \( t > 0 \), we can show that \( F(s) \) has no singularities in the LHP.

The converses of these statements are also true. We therefore conclude that the inverse Laplace transform of \( S^+(\lambda) \) is causal and that of \( S^-(-\lambda) \) is anticausal.

### 8.3. Back to Causal Wiener Filtering.

Recall the Wiener integral equation and define

\[
g(\tau) = R_{XS}(\tau + \alpha) - \int_0^\infty R_{XX}(\tau - \sigma) h_0(\sigma) d\sigma
\]

\[
= R_{XS}(\tau + \alpha) - \int_{-\infty}^\infty R_{XX}(\tau - \sigma) h_0(\sigma) d\sigma.
\]

Since \( g(\tau) \) is anticausal (by invoking the Wiener equation), \( G(s) \) has no singularities in the LHP. And,

\[
G(s) = S_{XS}(s)e^{s\alpha} - S_{XX}(s)h_0(s)
\]

\[
= S_{XS}(s)e^{s\alpha} - S_{XX}^+(s)S_{XX}^-(s)h_0(s).
\]

Now define \( a(\tau) = b(\tau) - c(\tau) \) where

\[
a(\tau) = \frac{1}{2\pi} \int_{-\infty}^\infty \frac{G(\omega)}{S_{XX}(\omega)} e^{j\omega \tau} d\omega,
\]

\[
b(\tau) = \frac{1}{2\pi} \int_{-\infty}^\infty \frac{S_{XS}(\omega)}{S_{XX}(\omega)} e^{j\omega \alpha} e^{j\omega \tau} d\omega,
\]

and

\[
c(\tau) = \frac{1}{2\pi} \int_{-\infty}^\infty S_{XX}^+(\omega) h_0(\omega) e^{j\omega \tau} d\omega.
\]

Note that \( a(\tau) \) is anti-causal since \( \frac{G(s)}{S_{XX}(s)} \) has no singularities in the LHP. Thus, \( a(\tau) = 0, \forall \tau \geq 0 \), and therefore \( b(\tau) = c(\tau), \forall \tau \geq 0 \). Also note that \( S_{XX}^+(\omega) h_0(\omega) \) has no singularities in the RHP; thus, \( c(\tau) = 0, \forall \tau \leq 0 \). Hence,

\[
\int_{-\infty}^\infty c(\tau)e^{-j\omega \tau} d\tau = \int_0^\infty c(\tau)e^{-j\omega \tau} d\tau = S_{XX}^+(\omega)h_0(\omega) = \int_0^\infty b(\tau)e^{-j\omega \tau} d\tau,
\]

and we obtain

\[
H_0(\omega) = \frac{1}{2\pi} \frac{1}{S_{XX}(\omega)} \int_0^\infty e^{-j\omega \tau} \int_{-\infty}^\infty \frac{S_{XS}(\omega')}{S_{XX}(\omega')} e^{j\omega' \alpha} e^{j\omega' \tau} d\omega' d\tau.
\]

### 8.4. Bode-Shannon method.

A simple implementation of the above solution can be obtained as follows: Let \( H_1(\omega) = \frac{1}{S_{XX}(\omega)} \). Consider the unrealizable filter defined by \( H_3(\omega) = \frac{S_{XS}(\omega)}{S_{XX}(\omega)} e^{j\omega \alpha} \). Define \( h_2(t) = h_3(t)u(t) \). Then we will have

\[
H_0(\omega) = H_1(\omega)H_2(\omega)
\]

**Example:** \( X(t) = S(t) + N(t) \), \( N(t) \) is zero-mean white Gaussian noise that is independent of \( S(t) \).

\[
S_{SS}(\omega) = \frac{2k}{\omega^2 + k^2} = \frac{\sqrt{2k}}{j\omega + k} \frac{\sqrt{2k}}{-j\omega + k}
\]
\[
S_{XX}(\omega) = S_{SS}(\omega) + S_{NN}(\omega) = \frac{2k}{\omega^2 + k^2} + \frac{N_0}{2} = \frac{N_0 \omega^2 + k^2 (1 + \beta)}{\omega^2 + k^2}, \text{ where } \beta = \frac{4}{N_0 k}.
\]
Next,
\[
S^+_{XX}(\omega) = \sqrt{\frac{N_0}{2} j \omega + k \sqrt{1 + \beta}} + j \omega + k \sqrt{1 + \beta}, \quad S^-_{XX}(\omega) = \sqrt{\frac{N_0}{2} j \omega - k \sqrt{1 + \beta}} - j \omega + k \sqrt{1 + \beta}.
\]

We must now take the inverse Fourier transform to find \( h_3(t) \):
\[
\mathcal{F}^{-1}\left\{ \frac{S_{SS}(\omega)}{S_{XX}} e^{j \omega \alpha} \right\} = \mathcal{F}^{-1}\left\{ \frac{2k}{\sqrt{\frac{N_0}{2} (j \omega + k)(-j \omega + k \sqrt{1 + \beta})}} e^{j \omega \alpha} \right\}.
\]
Therefore,
\[
h_3(t) = \begin{cases} 
\frac{2}{\sqrt{N_0/2}} \frac{1}{1 + \sqrt{1 + \beta}} e^{-k(\tau + \alpha)}, & \tau > -\alpha \\
\frac{2}{\sqrt{N_0/2}} \frac{1}{1 + \sqrt{1 + \beta}} e^{k \sqrt{1 + \beta}(\tau + \alpha)}, & \tau \leq -\alpha
\end{cases}
\]

If \( \alpha = 0 \) (filtering problem), then \( H_2(\omega) = \int_0^\infty h_3(\tau) e^{-j \omega \tau} d\tau = \frac{1}{1 + \sqrt{1 + \beta}} \frac{2}{\sqrt{N_0/2} (j \omega + k)}, \)
and \( H_0(\omega) = H_1(\omega) H_2(\omega) = \frac{2}{\sqrt{\frac{N_0}{2} (1 + \sqrt{1 + \beta})} j \omega + k \sqrt{1 + \beta}} \). Finally, \( h_0(t) = \frac{2}{\sqrt{N_0/2} (1 + \sqrt{1 + \beta})} e^{-k \sqrt{1 + \beta} t} u(t). \)

Acknowledgement: I wish to thank Professor Jim Bucklew for developing most of the material of this section.


