Lecture Notes 1
Review of Basic Probability Theory

- Probability Space and Axioms
- Basic Laws
- Conditional Probability and Bayes Rule
- Independence

Probability Theory

- Probability theory provides the mathematical rules for assigning probabilities to outcomes of random experiments, e.g., coin flips, packet arrivals, noise voltage
- Basic elements of probability theory:
  - *Sample space* Ω: set of all possible "elementary" or "finest grain" outcomes of the random experiment
  - *Set of events* $\mathcal{F}$: set of (all?) subsets of Ω — an event $A \subset \Omega$ occurs if the outcome $\omega \in A$
  - *Probability measure* P: function over $\mathcal{F}$ that assigns probabilities to events according to the axioms of probability (see below)
- Formally, a *probability space* is the triple $(\Omega, \mathcal{F}, \mathrm{P})$

Axioms of Probability

- A probability measure P satisfies the following axioms:
  1. $\mathrm{P}(A) \geq 0$ for every event A in $\mathcal{F}$
  2. $\mathrm{P}(\Omega) = 1$
  3. If $A_1, A_2, \ldots$ are *disjoint events* — i.e., $A_i \cap A_j = \emptyset$, for all $i \neq j$ — then

$$\mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(A_i)$$

- Notes:
  - P is a *measure* in the same sense as *mass*, *length*, *area*, and *volume* — all satisfy axioms 1 and 3
  - Unlike these other measures, P is bounded by 1 (axiom 2)
  - This analogy provides some intuition but is not sufficient to fully understand probability theory — other aspects such as conditioning and independence are unique to probability

Discrete Probability Spaces

- A sample space Ω is said to be *discrete* if it is countable
- Examples:
  - Flipping a coin: $\Omega = \{H, T\}$
  - Rolling a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - Flipping a coin n times: $\Omega = \{H, T\}^n$, sequences of heads/tails of length n
  - Flipping a coin until the first heads occurs: $\Omega = \{H, TH, TTH, TTTH, \ldots\}$
  - Number of packets arriving at a node in a communication network in time interval $(0, T\,]$: $\Omega = \{0, 1, 2, 3, \ldots\}$

  The first three examples have *finite* Ω, whereas the last two have *countably infinite* Ω. Both types are considered *discrete*

- For discrete sample spaces, the set of events $\mathcal{F}$ can be taken to be the set of all subsets of Ω, sometimes called the *power set* of Ω

- Example: For the coin flipping experiment,

$$\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \Omega\}$$

- $\mathcal{F}$ does not have to be the entire power set (more on this later)
- The probability measure P can be defined by assigning probabilities to individual outcomes — single outcome events $\{\omega\}$ — so that:

$$\mathrm{P}(\{\omega\}) \geq 0 \text{ for every } \omega \in \Omega$$

$$\sum_{\omega \in \Omega} \mathrm{P}(\{\omega\}) = 1$$

- The probability of any other event A is simply

$$\mathrm{P}(A) = \sum_{\omega \in A} \mathrm{P}(\{\omega\})$$

- Examples:
  - For the coin flipping experiment, for some p with $0 \leq p \leq 1$ assign

    $$\mathrm{P}(\{H\}) = p \quad \text{and} \quad \mathrm{P}(\{T\}) = 1 - p$$

    Note: p is called the *bias*. A coin is *fair* if $p = \frac{1}{2}$
  - For the die rolling experiment, assign

    $$\mathrm{P}(\{i\}) = \tfrac{1}{6} \quad \text{for } i = 1, 2, \ldots, 6$$

    The probability of the event "the outcome is even," $A = \{2, 4, 6\}$, is

    $$\mathrm{P}(A) = \mathrm{P}(\{2\}) + \mathrm{P}(\{4\}) + \mathrm{P}(\{6\}) = \tfrac{3}{6} = \tfrac{1}{2}$$
  - For the number of packets arriving in $(0, T\,]$, for parameter $\lambda > 0$ assign

    $$\mathrm{P}(\{k\}) = \frac{(\lambda T)^k}{k!} e^{-\lambda T} \quad \text{for } k = 0, 1, 2, \ldots$$

    This is the *Poisson* probability distribution with parameter λ, which is the average number of packets per unit time

Continuous Probability Spaces

- A *continuous* sample space Ω has an uncountable number of elements
- Examples:
  - Random number between 0 and 1: $\Omega = (0, 1\,]$
  - Packet arrival time: $\Omega = (0, \infty)$
  - Point in the unit disk: $\Omega = \{(x, y) \colon x^2 + y^2 \leq 1\}$
  - Arrival times of n packets: $\Omega = (0, \infty)^n$
- For continuous Ω, we cannot in general define the probability measure P by first assigning probabilities to outcomes
- To see why, consider assigning a uniform probability measure over $(0, 1\,]$
  - In this case the probability of each single outcome event is zero
  - How do we find the probability of an event such as $A = [0.25, 0.75]$?

- Another difference for continuous Ω: we cannot take the set of events $\mathcal{F}$ as the power set of Ω. (To learn why you need to study measure theory, which is beyond the scope of this course)
- The set of events $\mathcal{F}$ cannot be an arbitrary collection of subsets of Ω. It must make sense, e.g., if A is an event, then its complement A^c must also be an event, the union of two events must be an event, and so on
- Formally, $\mathcal{F}$ must be a *sigma algebra* (σ*-algebra*, σ*-field*), which satisfies the following axioms:
  1. $\emptyset \in \mathcal{F}$
  2. If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$
  3. If $A_1, A_2, \ldots \in \mathcal{F}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$
- Of course, the power set is a sigma algebra. But we can define smaller σ-algebras. For example, for rolling a die, we could define the set of events as

$$\mathcal{F} = \{\emptyset, \text{odd}, \text{even}, \Omega\}$$

- For $\Omega = R = (-\infty, \infty)$ (or $(0, \infty)$, $(0, 1)$, etc.) $\mathcal{F}$ is typically defined as the family of sets obtained by starting from the intervals and taking countable unions, intersections, and complements
- The resulting $\mathcal{F}$ is called the *Borel field*
- Note: Amazingly there are subsets in R that cannot be generated in this way! (Not ones that you are likely to encounter in your life as an engineer or even as a mathematician)
- To define a probability measure over a Borel field, we first assign probabilities to the intervals in a consistent way, i.e., in a way that satisfies the axioms of probability

  For example to define uniform probability measure over $(0, 1)$, we first assign $\mathrm{P}((a, b)) = b - a$ to all intervals
- In EE 278 we do not deal with sigma fields or the Borel field beyond (kind of) knowing what they are

Basic Probability Properties

- $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$
- If $A \subset B$, then $\mathrm{P}(A) \leq \mathrm{P}(B)$
- $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$
- $\mathrm{P}(A \cup B) \leq \mathrm{P}(A) + \mathrm{P}(B)$
- More generally, the *Union of Events or Bonferroni Bound*:

$$\mathrm{P}\Big(\bigcup_{i=1}^{n} A_i\Big) \;\leq\; \sum_{i=1}^{n} \mathrm{P}(A_i)$$

- *Law of Total Probability*: Let $A_1, A_2, A_3, \ldots$ be events that partition Ω, i.e., disjoint ($A_i \cap A_j = \emptyset$ for $i \neq j$) and $\bigcup_i A_i = \Omega$. Then for any event B

$$\mathrm{P}(B) \;=\; \textstyle\sum_i \mathrm{P}(A_i \cap B)$$

  The Law of Total Probability is very useful for finding probabilities of sets

Conditional Probability

- Let B be an event such that $\mathrm{P}(B) \neq 0$. The *conditional probability* of event A given B is defined to be

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(A, B)}{\mathrm{P}(B)}$$

- The function $\mathrm{P}(\cdot \mid B)$ is a probability measure over $\mathcal{F}$, i.e., it satisfies the axioms of probability

- Chain rule: $\mathrm{P}(A, B) = \mathrm{P}(A)\mathrm{P}(B \mid A) = \mathrm{P}(B)\mathrm{P}(A \mid B)$ (this can be generalized to n events.

- The probability of event A given B, a nonzero probability event — the *a posteriori* probability of A — is related to the unconditional probability of A — the *a priori* probability — by

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(B \mid A)}{\mathrm{P}(B)} \mathrm{P}(A)$$

  This follows directly from the definition of conditional probablility

Bayes Rule

- Let $A_1, A_2, \ldots, A_n$ be nonzero probability events that partition Ω, and let B be a nonzero probability event

- We know $\mathrm{P}(A_i)$ and $\mathrm{P}(B \mid A_i)$, $i = 1, 2, \ldots, n$, and want to find the a posteriori probabilities $\mathrm{P}(A_j \mid B)$, $j = 1, 2, \ldots, n$

- We know that

$$\mathrm{P}(A_j \mid B) = \frac{\mathrm{P}(B \mid A_j)}{\mathrm{P}(B)} \mathrm{P}(A_j)$$

- By the law of total probability

$$\mathrm{P}(B) = \sum_{i=1}^{n} \mathrm{P}(A_i, B) = \sum_{i=1}^{n} \mathrm{P}(A_i)\mathrm{P}(B \mid A_i)$$

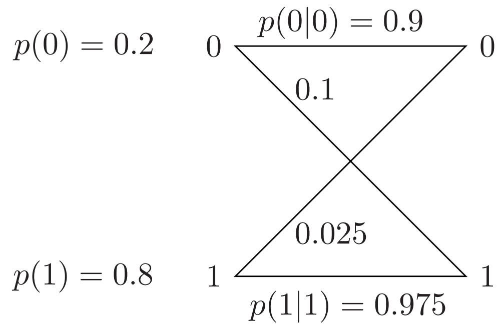
- Substituting, we obtain *Bayes rule*

$$\mathrm{P}(A_j \mid B) = \frac{\mathrm{P}(B \mid A_j)}{\sum_{i=1}^{n} \mathrm{P}(A_i)\mathrm{P}(B \mid A_i)} \mathrm{P}(A_j), \; j = 1, 2, \ldots, n$$

- Bayes rule also applies to a (countably) infinite number of events

Example: Binary Communication Channel

- Consider the following *probability transition diagram* for a noisy binary channel



  Given that 0 is received, find the probability that 0 was sent

- This is a random experiment with sample space

$$\Omega = \{(0,0),(0,1),(1,0),(1,1)\},$$

  where the first entry is the bit sent and the second is the bit received

- Define the two events

$$A = \{0 \text{ is sent}\} = \{(0,1),(0,0)\}$$
$$B = \{0 \text{ is received}\} = \{(0,0),(1,0)\}$$

- The probability measure on Ω is determined by $\mathrm{P}(A)$, $\mathrm{P}(B \mid A)$, and $\mathrm{P}(B^c|A^c)$, which are given on the probability transition diagram of the channel

- To find $\mathrm{P}(A \mid B)$, the *a posteriori* probability that 0 was sent, use Bayes rule:

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(B \mid A)}{\mathrm{P}(A)\mathrm{P}(B \mid A) + \mathrm{P}(A^c)\mathrm{P}(B \mid A^c)}\,\mathrm{P}(A)$$

  to obtain

$$\mathrm{P}(A \mid B) = \frac{0.9}{0.2 \cdot 0.9 + 0.8 \cdot 0.025} \cdot 0.2 = \frac{0.9}{0.2} \cdot 0.2 = 0.9$$

  Note that $\mathrm{P}(A \mid B) = 0.9$; the a posteriori probability of A, is much larger than $\mathrm{P}(A) = 0.2$, the a priori probability of A

Independence

- Two events are said to be *statistically independent* if
$$\mathrm{P}(A, B) = \mathrm{P}(A)\mathrm{P}(B)$$
- When $\mathrm{P}(B) \neq 0$, this is equivalent to
$$\mathrm{P}(A \mid B) = \mathrm{P}(A)$$
In other words, knowing whether B occurs does not change the probability of A
- Example: Assuming that the binary channel of the previous example is used to send two bits independently, what is the probability that both bits are in error?
  - Define the two events
$$E_1 = \{\text{First bit is in error}\}\,, \quad E_2 = \{\text{Second bit is in error}\}$$
  - Since the bits are sent independently, the probability that both are in error is
$$\mathrm{P}(E_1 \cap E_2) \;=\; \mathrm{P}(E_1, E_2) \;=\; \mathrm{P}(E_1)\mathrm{P}(E_2)$$

  - To find $\mathrm{P}(E_1)$, we express E_1 in terms of the events A and B as
$$E_1 = (A_1 \cap B_1^c) \cup (A_1^c \cap B_1)$$
  - Since E_1 has been expressed as the union of disjoint events,
$$\begin{aligned}\mathrm{P}(E_1) &= \mathrm{P}(A_1, B_1^c) + \mathrm{P}(A_1^c, B_1) \\ &= \mathrm{P}(A_1)\mathrm{P}(B_1^c \mid A_1) + \mathrm{P}(A_1^c)\mathrm{P}(B_1 \mid A_1^c) \\ &= 0.2 \cdot 0.1 \;+\; 0.8 \cdot 0.025 \;=\; 0.04\end{aligned}$$
  - The probability that the two bits are in error is
$$\mathrm{P}(E_1, E_2) \;=\; \mathrm{P}(E_1)\mathrm{P}(E_2) \;=\; (0.04)^2 \;=\; 1.6 \times 10^{-3}$$
- In general $A_1, A_2, \ldots, A_n$ are defined to be independent if for *every* subset $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$ of the events,
$$\mathrm{P}(A_{i_1}, A_{i_2}, \ldots, A_{i_k}) \;= \prod_{j=1}^{k} \mathrm{P}(A_{i_j})$$
- Note: $\mathrm{P}(A_1, A_2, \ldots, A_n) = \prod_{j=1}^{n} \mathrm{P}(A_i)$ is *not* sufficient for independence

Example: Roll two fair dice independently. Define the events

$$A = \{\text{First die is 1, 2, or 3}\}$$
$$B = \{\text{First die is 2, 3, or 6}\}$$
$$C = \{\text{Sum of outcomes is 9}\} \, = \, \{(3,6),(4,5),(5,4),(6,3)\}$$

Are A, B, and C independent?

Since the dice are fair and the experiments are done independently, the probability of any pair of outcomes is $\frac{1}{36}$. Therefore

$$\mathrm{P}(A) = \tfrac{1}{2}, \quad \mathrm{P}(B) = \tfrac{1}{2}, \quad \mathrm{P}(C) = \tfrac{1}{9}$$

Since $A \cap B \cap C = \{(3,6)\}$,

$$\mathrm{P}(A,B,C) = \tfrac{1}{36} = \mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C)$$

But A, B, and C are *not* independent because

$$\mathrm{P}(A,B) = \tfrac{2}{6} = \tfrac{1}{3} \neq \tfrac{1}{4} = \mathrm{P}(A)\mathrm{P}(B)$$