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* $\Omega$: set of all possible “elementary” or “finest grain” outcomes of the random experiment.
  - *Set of events* $\mathcal{F}$: set of (all?) subsets of $\Omega$ — an event $A \subseteq \Omega$ 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 $(\Omega, \mathcal{F}, P)$. 

Axioms of Probability

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

• 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 $\Omega$ is said to be discrete if it is countable

• Examples:
  ◦ 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: $\Omega = \{H, TH, TTH, TTTH, \ldots\}$

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

• 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:

\[
P(\{\omega\}) \geq 0 \quad \text{for every } \omega \in \Omega \\
\sum_{\omega \in \Omega} P(\{\omega\}) = 1
\]

• The probability of any other event \( A \) is simply

\[
P(A) = \sum_{\omega \in A} P(\{\omega\})
\]

• Example: For the die rolling experiment, assign

\[
P(\{i\}) = \frac{1}{6} \quad \text{for } i = 1, 2, \ldots, 6
\]

The probability of the event “the outcome is even,” \( A = \{2, 4, 6\} \), is

\[
P(A) = P(\{2\}) + P(\{4\}) + P(\{6\}) = \frac{3}{6} = \frac{1}{2}
\]

### Continuous Probability Spaces

• A continuous sample space \( \Omega \) has an uncountable number of elements

• Examples:
  - Random number between 0 and 1: \( \Omega = (0, 1] \)
  - Point in the unit disk: \( \Omega = \{(x, y) : x^2 + y^2 \leq 1\} \)
  - Arrival times of \( n \) packets: \( \Omega = (0, \infty)^n \)

• For continuous \( \Omega \), 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 $\Omega$: we cannot take the set of events $\mathcal{F}$ as the power set of $\Omega$. (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 $\Omega$. 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* ($\sigma$-algebra, $\sigma$-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 $\sigma$-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 $P((a, b)) = b - a$ to all intervals.

In EE 278 we do not deal with sigma fields or the Borel field beyond (kind of) knowing what they are.
Useful Probability Laws

- **Union of Events Bound:**
  \[ P \left( \bigcup_{i=1}^{n} A_i \right) \leq \sum_{i=1}^{n} P(A_i) \]

- **Law of Total Probability:** Let \( A_1, A_2, A_3, \ldots \) be events that partition \( \Omega \), i.e., disjoint \((A_i \cap A_j = \emptyset \text{ for } i \neq j)\) and \( \bigcup_i A_i = \Omega \). Then for any event \( B \)
  \[ P(B) = \sum_i P(A_i \cap B) \]
  The Law of Total Probability is very useful for finding probabilities of sets

Conditional Probability

- Let \( B \) be an event such that \( P(B) \neq 0 \). The conditional probability of event \( A \) given \( B \) is defined to be
  \[ P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A, B)}{P(B)} \]

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

- **Chain rule:** \( P(A, B) = P(A)P(B \mid A) = P(B)P(A \mid B) \) (this can be generalized to \( n \) events)

- The probability of event \( A \) given \( B \), a nonzero probability event — the \textit{a posteriori} probability of \( A \) — is related to the unconditional probability of \( A \) — the \textit{a priori} probability — by
  \[ P(A \mid B) = \frac{P(B \mid A)}{P(B)} P(A) \]
  This follows directly from the definition of conditional probability
Bayes Rule

- Let $A_1, A_2, \ldots, A_n$ be nonzero probability events that partition $\Omega$, and let $B$ be a nonzero probability event.
- We know $P(A_i)$ and $P(B \mid A_i)$, $i = 1, 2, \ldots, n$, and want to find the a posteriori probabilities $P(A_j \mid B)$, $j = 1, 2, \ldots, n$.
- We know that
  \[
P(A_j \mid B) = \frac{P(B \mid A_j)}{P(B)} P(A_j)
  \]
- By the law of total probability
  \[
P(B) = \sum_{i=1}^{n} P(A_i, B) = \sum_{i=1}^{n} P(A_i)P(B \mid A_i)
  \]
- Substituting, we obtain Bayes rule
  \[
P(A_j \mid B) = \frac{P(B \mid A_j)P(A_j)}{\sum_{i=1}^{n} P(A_i)P(B \mid A_i)} P(A_i), \quad j = 1, 2, \ldots, n
  \]
- Bayes rule also applies to a (countably) infinite number of events.

Independence

- Two events are said to be statistically independent if
  \[
P(A, B) = P(A)P(B)
  \]
- When $P(B) \neq 0$, this is equivalent to
  \[
P(A \mid B) = P(A)
  \]
  In other words, knowing whether $B$ occurs does not change the probability of $A$.
- The events $A_1, A_2, \ldots, A_n$ are said to be independent if for every subset $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$ of the events,
  \[
P(A_{i_1}, A_{i_2}, \ldots, A_{i_k}) = \prod_{j=1}^{k} P(A_{i_j})
  \]
- Note: $P(A_1, A_2, \ldots, A_n) = \prod_{j=1}^{n} P(A_i)$ is not sufficient for independence.
Random Variables

- A random variable (r.v.) is a real-valued function $X(\omega)$ over a sample space $\Omega$, i.e., $X : \Omega \rightarrow \mathbb{R}$

\[ \Omega \]

\[ \omega \]

\[ X(\omega) \]

- Notations:
  - We use upper case letters for random variables: $X, Y, Z, \Phi, \Theta, \ldots$
  - We use lower case letters for values of random variables: $X = x$ means that random variable $X$ takes on the value $x$, i.e., $X(\omega) = x$ where $\omega$ is the outcome

Specifying a Random Variable

- Specifying a random variable means being able to determine the probability that $X \in A$ for any Borel set $A \subset \mathbb{R}$, in particular, for any interval $(a, b]$
- To do so, consider the inverse image of $A$ under $X$, i.e., $\{\omega : X(\omega) \in A\}$

\[ \text{inverse image of } A \text{ under } X(\omega), \text{i.e., } \{\omega : X(\omega) \in A\} \]

- Since $X \in A$ iff $\omega \in \{\omega : X(\omega) \in A\}$,
  \[ P(\{X \in A\}) = P(\{\omega : X(\omega) \in A\}) = P(\omega : X(\omega) \in A) \]
  Shorthand: $P(\{\text{set description}\}) = P(\text{set description})$
Cumulative Distribution Function (CDF)

- We need to be able to determine \( P\{X \in A\} \) for any Borel set \( A \subset \mathbb{R} \), i.e., any set generated by starting from intervals and taking countable unions, intersections, and complements.
- Hence, it suffices to specify \( P\{X \in (a, b]\} \) for all intervals. The probability of any other Borel set can be determined by the axioms of probability.
- Equivalently, it suffices to specify its cumulative distribution function (cdf):
  \[
  F_X(x) = P\{X \leq x\} = P\{X \in (-\infty, x]\}, \quad x \in \mathbb{R}.
  \]

- Properties of cdf:
  - \( F_X(x) \geq 0 \)
  - \( F_X(x) \) is monotonically nondecreasing, i.e., if \( a > b \) then \( F_X(a) \geq F_X(b) \)
  
  ![CDF Diagram](image)

- Limits: \( \lim_{x \to +\infty} F_X(x) = 1 \) and \( \lim_{x \to -\infty} F_X(x) = 0 \)
- \( F_X(x) \) is right continuous, i.e., \( F_X(a^+) = \lim_{x \to a^+} F_X(x) = F_X(a) \)
- \( P\{X = a\} = F_X(a) - F_X(a^-) \), where \( F_X(a^-) = \lim_{x \to a^-} F_X(x) \)
- For any Borel set \( A \), \( P\{X \in A\} \) can be determined from \( F_X(x) \)
- Notation: \( X \sim F_X(x) \) means that \( X \) has cdf \( F_X(x) \)
A random variable is said to be \textit{discrete} if $F_X(x)$ consists only of steps over a countable set $\mathcal{X}$.

Hence, a discrete random variable can be completely specified by the \textit{probability mass function} (pmf)

$$p_X(x) = P\{X = x\} \text{ for every } x \in \mathcal{X}$$

Clearly $p_X(x) \geq 0$ and $\sum_{x \in \mathcal{X}} p_X(x) = 1$

Notation: We use $X \sim p_X(x)$ or simply $X \sim p(x)$ to mean that the discrete random variable $X$ has pmf $p_X(x)$ or $p(x)$.

Famous discrete random variables:

- \textit{Bernoulli}: $X \sim \text{Bern}(p)$ for $0 \leq p \leq 1$ has the pmf
  $$p_X(1) = p \quad \text{and} \quad p_X(0) = 1 - p$$

- \textit{Geometric}: $X \sim \text{Geom}(p)$ for $0 \leq p \leq 1$ has the pmf
  $$p_X(k) = p(1 - p)^{k-1}, \quad k = 1, 2, 3, \ldots$$

- \textit{Binomial}: $X \sim \text{Binom}(n, p)$ for integer $n > 0$ and $0 \leq p \leq 1$ has the pmf
  $$p_X(k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, 2, \ldots$$

- \textit{Poisson}: $X \sim \text{Poisson}(\lambda)$ for $\lambda > 0$ has the pmf
  $$p_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots$$

Remark: Poisson is the limit of Binomial for $np = \lambda$ as $n \to \infty$, i.e., for every $k = 0, 1, 2, \ldots$, the Binom$(n, \lambda/n)$ pmf

$$p_X(k) \to \frac{\lambda^k}{k!} e^{-\lambda} \text{ as } n \to \infty$$
Probability Density Function (PDF)

- A random variable is said to be continuous if its cdf is a continuous function

![image]

- If $F_X(x)$ is continuous and differentiable (except possibly over a countable set), then $X$ can be completely specified by a probability density function (pdf) $f_X(x)$ such that

$$F_X(x) = \int_{-\infty}^{x} f_X(u) \, du$$

- If $F_X(x)$ is differentiable everywhere, then (by definition of derivative)

$$f_X(x) = \frac{dF_X(x)}{dx} = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{P\{x < X \leq x + \Delta x\}}{\Delta x}$$

**Properties of pdf:**

- $f_X(x) \geq 0$
- $\int_{-\infty}^{\infty} f_X(x) \, dx = 1$
- For any event (Borel set) $A \subset \mathbb{R}$,

$$P\{X \in A\} = \int_{x \in A} f_X(x) \, dx$$

In particular,

$$P\{x_1 < X \leq x_2\} = \int_{x_1}^{x_2} f_X(x) \, dx$$

- Important note: $f_X(x)$ should not be interpreted as the probability that $X = x$. In fact, $f_X(x)$ is not a probability measure since it can be $> 1$

- Notation: $X \sim f_X(x)$ means that $X$ has pdf $f_X(x)$
• Famous continuous random variables:
  o \textit{Uniform}: \( X \sim U[a, b] \) where \( a < b \) has pdf
    \[
    f_X(x) = \begin{cases} 
    \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 
    0 & \text{otherwise} 
    \end{cases}
    \]
  o \textit{Exponential}: \( X \sim \text{Exp}(\lambda) \) where \( \lambda > 0 \) has pdf
    \[
    f_X(x) = \begin{cases} 
    \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 
    0 & \text{otherwise} 
    \end{cases}
    \]
  o \textit{Laplace}: \( X \sim \text{Laplace}(\lambda) \) where \( \lambda > 0 \) has pdf
    \[
    f_X(x) = \frac{1}{2\lambda} e^{-\lambda|x|}
    \]
  o \textit{Gaussian}: \( X \sim \mathcal{N}(\mu, \sigma^2) \) with parameters \( \mu \) (the mean) and \( \sigma^2 \) (the variance, \( \sigma \) is the standard deviation) has pdf
    \[
    f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
    \]

The cdf of the standard normal random variable \( \mathcal{N}(0, 1) \) is
\[
\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du
\]

Define the function \( Q(x) = 1 - \Phi(x) = P\{X > x\} \)

\( \mathcal{N}(0, 1) \)

The \( Q(\cdot) \) function is used to compute \( P\{X > a\} \) for any Gaussian r.v. \( X \): Given \( Y \sim \mathcal{N}(\mu, \sigma^2) \), we represent it using the standard \( X \sim \mathcal{N}(0, 1) \) as
\[
Y = \sigma X + \mu
\]
Then
\[
P\{Y > y\} = P\left\{ X > \frac{y-\mu}{\sigma}\right\} = Q\left(\frac{y-\mu}{\sigma}\right)
\]

○ The complementary error function is \( \text{erfc}(x) = 2Q(\sqrt{2}x) \)
Functions of a Random Variable

- Suppose we are given a r.v. $X$ with known cdf $F_X(x)$ and a function $y = g(x)$. What is the cdf of the random variable $Y = g(X)$?

- We use

$$F_Y(y) = P\{Y \leq y\} = P\{x : g(x) \leq y\}$$

- Example: Square law detector. Let $X \sim F_X(x)$ and $Y = X^2$. We wish to find $F_Y(y)$

If $y < 0$, then clearly $F_Y(y) = 0$. Consider $y \geq 0$,

$$F_Y(y) = P \{-\sqrt{y} < X \leq \sqrt{y}\} = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

If $X$ is continuous with density $f_X(x)$, then

$$f_Y(y) = \frac{1}{2\sqrt{y}}\left(f_X(\sqrt{y}) + f_X(-\sqrt{y})\right)$$
• Remark: In general, let $X \sim f_X(x)$ and $Y = g(X)$ be differentiable. Then

$$f_Y(y) = \sum_{i=1}^{k} \frac{f_X(x_i)}{|g'(x_i)|},$$

where $x_1, x_2, \ldots$ are the solutions of the equation $y = g(x)$ and $g'(x_i)$ is the derivative of $g$ evaluated at $x_i$.

• Example: Limiter. Let $X \sim \text{Laplace}(1)$, i.e., $f_X(x) = (1/2)e^{-|x|}$, and let $Y$ be defined by the function of $X$ shown in the figure. Find the cdf of $Y$.

To find the cdf of $Y$, we consider the following cases:

- $y < -a$: Here clearly $F_Y(y) = 0$.
- $y = -a$: Here

$$F_Y(-a) = F_X(-1) = \int_{-\infty}^{-1} \frac{1}{2}e^x \, dx = \frac{1}{2}e^{-1}.$$
- $-a < y < a$: Here

$$F_Y(y) = P\{Y \leq y\}$$

$$= P\{aX \leq y\}$$

$$= P\left\{ X \leq \frac{y}{a} \right\} = F_X\left( \frac{y}{a} \right)$$

$$= \frac{1}{2}e^{-1} + \int_{-1}^{y/a} \frac{1}{2}e^{-|x|} \, dx$$

- $y \geq a$: Here $F_Y(y) = 1$

Combining the results, the following is a sketch of the cdf of $Y$

![Sketch of the cdf of Y](image)

**Generation of Random Variables**

- Generating a r.v. with a prescribed distribution is often needed for performing simulations involving random phenomena, e.g., noise or random arrivals.

- First let $X \sim F(x)$ where the cdf $F(x)$ is continuous and strictly increasing. Define $Y = F(X)$, a real-valued random variable that is a function of $X$.

What is the cdf of $Y$?

Clearly, $F_Y(y) = 0$ for $y < 0$, and $F_Y(y) = 1$ for $y > 1$

For $0 \leq y \leq 1$, note that by assumption $F$ has an inverse $F^{-1}$, so

$$F_Y(y) = P\{Y \leq y\} = P\{F(X) \leq y\} = P\{X \leq F^{-1}(y)\} = F(F^{-1}(y)) = y$$

Thus $Y \sim U[0,1]$, i.e., $Y$ is a uniformly distributed random variable.

- Note: $F(x)$ does not need to be invertible. If $F(x) = a$ is constant over some interval, then the probability that $X$ lies in this interval is zero. Without loss of generality, we can take $F^{-1}(a)$ to be the leftmost point of the interval.

- Conclusion: We can generate a $U[0,1]$ r.v. from any continuous r.v.
Now, let’s consider the opposite scenario where we are given $X \sim U[0, 1]$ (a random number generator) and wish to generate a random variable $Y$ with prescribed cdf $F(y)$, e.g., Gaussian or exponential.

If $F$ is continuous and strictly increasing, set $Y = F^{-1}(X)$. To show $Y \sim F(y)$,

$$F_Y(y) = P\{Y \leq y\} = P\{F^{-1}(X) \leq y\} = P\{X \leq F(y)\} = F(y),$$

since $X \sim U[0, 1]$ and $0 \leq F(y) \leq 1$

**Example:** To generate $Y \sim \text{Exp}(\lambda)$, set

$$Y = -\frac{1}{\lambda} \ln(1 - X)$$

**Note:** $F$ does not need to be continuous for the above to work. For example, to generate $Y \sim \text{Bern}(p)$, we set

$$Y = \begin{cases} 
0 & X \leq 1 - p \\
1 & \text{otherwise}
\end{cases}$$

**Conclusion:** We can generate a r.v. with any desired distribution from a $U[0, 1]$ r.v.
• A pair of random variables defined over the same probability space are specified by their joint cdf

\[ F_{X,Y}(x, y) = P\{X \leq x, Y \leq y\}, \quad x, y \in \mathbb{R} \]

\( F_{X,Y}(x, y) \) is the probability of the shaded region of \( \mathbb{R}^2 \)

\[ (x, y) \]

\[ x \]

\[ y \]

• Properties of the cdf:
  
  o \( F_{X,Y}(x, y) \geq 0 \)
  
  o If \( x_1 \leq x_2 \) and \( y_1 \leq y_2 \) then \( F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2) \)
  
  o \( \lim_{y \to -\infty} F_{X,Y}(x, y) = 0 \) and \( \lim_{x \to -\infty} F_{X,Y}(x, y) = 0 \)
  
  o \( \lim_{y \to \infty} F_{X,Y}(x, y) = F_X(x) \) and \( \lim_{x \to \infty} F_{X,Y}(x, y) = F_Y(y) \)

  \( F_X(x) \) and \( F_Y(y) \) are the marginal cdfs of \( X \) and \( Y \)
  
  o \( \lim_{x, y \to \infty} F_{X,Y}(x, y) = 1 \)

• \( X \) and \( Y \) are independent if for every \( x \) and \( y \)

\[ F_{X,Y}(x, y) = F_X(x)F_Y(y) \]
Joint, Marginal, and Conditional PMFs

- Let $X$ and $Y$ be discrete random variables on the same probability space.
- They are completely specified by their joint pmf:
  \[ p_{X,Y}(x, y) = P\{X = x, Y = y\}, \quad x \in \mathcal{X}, \ y \in \mathcal{Y} \]
  By axioms of probability, \[ \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y) = 1 \]
- To find $p_X(x)$, the marginal pmf of $X$, we use the law of total probability:
  \[ p_X(x) = \sum_{y \in \mathcal{Y}} p(x, y), \quad x \in \mathcal{X} \]
- The conditional pmf of $X$ given $Y = y$ is defined as:
  \[ p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}, \quad p_Y(y) \neq 0, \ x \in \mathcal{X} \]
- Chain rule: \[ p_{X,Y}(x, y) = p_X(x)p_{Y|X}(y|x) = p_Y(y)p_{X|Y}(x|y) \]

- **Independence:** $X$ and $Y$ are said to be independent if for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$,
  \[ p_{X,Y}(x, y) = p_X(x)p_Y(y), \]
  which is equivalent to $p_{X|Y}(x|y) = p_X(x)$ for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ such that $p_Y(y) \neq 0$.
Joint, Marginal, and Conditional PDF

- $X$ and $Y$ are jointly continuous random variables if their joint cdf is continuous in both $x$ and $y$

In this case, we can define their joint pdf, provided that it exists, as the function $f_{X,Y}(x,y)$ such that

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v) \, du \, dv, \quad x, y \in \mathbb{R}$$

- If $F_{X,Y}(x,y)$ is differentiable in $x$ and $y$, then

$$f_{X,Y}(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y} = \lim_{\Delta x, \Delta y \to 0} \frac{P\{x < X \leq x + \Delta x, y < Y \leq y + \Delta y\}}{\Delta x \Delta y}$$

- Properties of $f_{X,Y}(x,y)$:
  - $f_{X,Y}(x,y) \geq 0$
  - $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) \, dx \, dy = 1$

- The marginal pdf of $X$ can be obtained from the joint pdf via the law of total probability:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) \, dy$$

- $X$ and $Y$ are independent iff $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for every $x, y$

- Conditional cdf and pdf: Let $X$ and $Y$ be continuous random variables with joint pdf $f_{X,Y}(x,y)$. We wish to define $F_{Y|X}(y \mid X = x) = P\{Y \leq y \mid X = x\}$

We cannot define the above conditional probability as

$$\frac{P\{Y \leq y, X = x\}}{P\{X = x\}}$$

because both numerator and denominator are equal to zero. Instead, we define conditional probability for continuous random variables as a limit

$$F_{Y|X}(y \mid x) = \lim_{\Delta x \to 0} \frac{P\{Y \leq y \mid x < X \leq x + \Delta x\}}{P\{x < X \leq x + \Delta x\}}$$

$$= \lim_{\Delta x \to 0} \frac{P\{y \leq Y \leq y \mid x < X \leq x + \Delta x\}}{P\{x < X \leq x + \Delta x\}}$$

$$= \lim_{\Delta x \to 0} \frac{\int_{-\infty}^{y} f_{X,Y}(x,u) \, du \, \Delta x}{f_X(x) \Delta x} = \int_{-\infty}^{y} \frac{f_{X,Y}(x,u)}{f_X(x)} \, du$$
• We then define the conditional pdf in the usual way as

\[ f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \quad \text{if } f_X(x) \neq 0 \]

• Thus

\[ F_{Y|X}(y|x) = \int_{-\infty}^{y} f_{Y|X}(u|x) \, du \]

which shows that \( f_{Y|X}(y|x) \) is a pdf for \( Y \) given \( X = x \), i.e.,

\[ Y | \{X = x\} \sim f_{Y|X}(y|x) \]

• Independence: \( X \) and \( Y \) are independent if \( f_{X,Y}(x,y) = f_X(x)f_Y(y) \) for every \((x,y)\)

---

One Discrete and One Continuous Random Variables

• Let \( \Theta \) be a discrete random variable with pmf \( p_\Theta(\theta) \)

• For each \( \Theta = \theta \) with \( p_\Theta(\theta) \neq 0 \), let \( Y \) be a continuous random variable, i.e., \( F_{Y|\Theta}(y|\theta) \) is continuous for all \( \theta \). We define \( f_{Y|\Theta}(y|\theta) \) in the usual way

• The conditional pmf of \( \Theta \) given \( y \) can be defined as a limit

\[
p_{\Theta|Y}(\theta \mid y) = \lim_{\Delta y \to 0} \frac{\text{P}\{\Theta = \theta, y < Y \leq y + \Delta y\}}{\text{P}\{y < Y \leq y + \Delta y\}}
\]

\[
= \lim_{\Delta y \to 0} \frac{p_\Theta(\theta)f_{Y|\Theta}(y|\theta)\Delta y}{f_Y(y)\Delta y} = \frac{f_{Y|\Theta}(y|\theta)p_\Theta(\theta)}{f_Y(y)}
\]

This leads to the Bayes rule:

\[
p_{\Theta|Y}(\theta \mid y) = \frac{f_{Y|\Theta}(y|\theta)}{\sum_{\theta'} p_\Theta(\theta')f_{Y|\Theta}(y|\theta')} p_\Theta(\theta)
\]
• Example: *Additive Gaussian Noise Channel*

Consider the following communication channel:

\[
Z \sim \mathcal{N}(0, N)
\]

\[
\Theta \rightarrow + \rightarrow Y
\]

The signal transmitted is a binary random variable \( \Theta \):

\[
\Theta = \begin{cases} 
+1 & \text{with probability } p \\
-1 & \text{with probability } 1 - p
\end{cases}
\]

The received signal, also called the *observation*, is \( Y = \Theta + Z \), where \( \Theta \) and \( Z \) are independent.

Given \( Y = y \) is received (observed), find \( p_{\Theta|Y}(\theta|y) \), the a posteriori pmf of \( \Theta \).

• In some cases we are given \( f_Y(y) \) and \( p_{\Theta|Y}(\theta|y) \) for every \( y \).

• We can find the a posteriori pdf of \( Y \) using the Bayes rule:

\[
f_{Y|\Theta}(y|\theta) = \frac{p_{\Theta|Y}(\theta|y)}{\int f_Y(y')p_{\Theta|Y}(\theta|y')dy'} f_Y(y)
\]

• Example: *Coin with random bias*

Consider a coin with random bias \( P \sim f_P(p) \). Flip the coin and let \( X = 1 \) if the outcome is heads and \( X = 0 \) if the outcome is tails.

Given that \( X = 1 \) (i.e., outcome is heads), find \( f_{P|X}(p|1) \), the a posteriori pdf of \( P \).
Scalar Detection

Consider the following general digital communication system

\[ \Theta \in \{ \theta_0, \theta_1 \} \]

\[ Y \] noisy channel

\[ \hat{\Theta}(Y) \in \{ \theta_0, \theta_1 \} \]

where the signal sent is

\[ \Theta = \begin{cases} 
\theta_0 & \text{with probability } p \\
\theta_1 & \text{with probability } 1 - p 
\end{cases} \]

and the observation (received signal) is

\[ Y | \{ \Theta = \theta \} \sim f_{Y|\Theta}(y | \theta), \quad \theta \in \{ \theta_0, \theta_1 \} \]

We wish to find the estimate \( \hat{\Theta}(Y) \) (i.e., design the decoder) that minimizes the probability of error:

\[ P_e = P \{ \hat{\Theta} \neq \Theta \} = P \{ \Theta = \theta_0, \hat{\Theta} = \theta_1 \} + P \{ \Theta = \theta_1, \hat{\Theta} = \theta_0 \} \]

\[ = P \{ \Theta = \theta_0 \} P \{ \hat{\Theta} = \theta_1 | \Theta = \theta_0 \} + P \{ \Theta = \theta_1 \} P \{ \hat{\Theta} = \theta_0 | \Theta = \theta_1 \} \]

We define the maximum a posteriori probability (MAP) decoder as

\[ \hat{\Theta}(y) = \begin{cases} 
\theta_0 & \text{if } p_{\Theta|Y}(\theta_0|y) > p_{\Theta|Y}(\theta_1|y) \\
\theta_1 & \text{otherwise} 
\end{cases} \]

The MAP decoding rule minimizes \( P_e \), since

\[ \min_{\hat{\Theta}} P_e = 1 - \max_{\Theta} P \{ \hat{\Theta}(Y) = \Theta \} \]

\[ = 1 - \max_{\Theta} \int_{-\infty}^{\infty} f_Y(y) P \{ \Theta = \hat{\Theta}(y) | Y = y \} dy \]

\[ = 1 - \int_{-\infty}^{\infty} f_Y(y) \max_{\hat{\Theta}(y)} P \{ \Theta = \hat{\Theta}(y) | Y = y \} dy \]

and the probability of error is minimized if we pick the largest \( p_{\Theta|Y}(\hat{\Theta}(y)|y) \) for every \( y \), which is precisely the MAP decoder

If \( p = \frac{1}{2} \), i.e., equally likely signals, using Bayes rule, the MAP decoder reduces to the maximum likelihood (ML) decoder

\[ \hat{\Theta}(y) = \begin{cases} 
\theta_0 & \text{if } f_{Y|\Theta}(y|\theta_0) > f_{Y|\Theta}(y|\theta_1) \\
\theta_1 & \text{otherwise} 
\end{cases} \]
Additive Gaussian Noise Channel

- Consider the additive Gaussian noise channel with signal
  \[ \Theta = \begin{cases} 
  +\sqrt{P} & \text{with probability } \frac{1}{2} \\
  -\sqrt{P} & \text{with probability } \frac{1}{2} 
  \end{cases} \]
  noise \( Z \sim \mathcal{N}(0, N) \) (\( \Theta \) and \( Z \) are independent), and output \( Y = \Theta + Z \)

- The MAP decoder is
  \[ \hat{\Theta}(y) = \begin{cases} 
  +\sqrt{P} & \text{if } \frac{P\{\Theta = +\sqrt{P} | Y = y\}}{P\{\Theta = -\sqrt{P} | Y = y\}} > 1 \\
  -\sqrt{P} & \text{otherwise} 
  \end{cases} \]
  Since the two signals are equally likely, the MAP decoding rule reduces to the ML decoding rule
  \[ \hat{\Theta}(y) = \begin{cases} 
  +\sqrt{P} & \text{if } \frac{f_{Y|\Theta}(y | +\sqrt{P})}{f_{Y|\Theta}(y | -\sqrt{P})} > 1 \\
  -\sqrt{P} & \text{otherwise} 
  \end{cases} \]

- Using the Gaussian pdf, the ML decoder reduces to the minimum distance decoder
  \[ \hat{\Theta}(y) = \begin{cases} 
  +\sqrt{P} & (y - \sqrt{P})^2 < (y - (-\sqrt{P}))^2 \\
  -\sqrt{P} & \text{otherwise} 
  \end{cases} \]
  From the figure, this simplifies to
  \[ \hat{\Theta}(y) = \begin{cases} 
  +\sqrt{P} & y > 0 \\
  -\sqrt{P} & y < 0 
  \end{cases} \]
  Note: The decision when \( y = 0 \) is arbitrary

\[ f(y | -\sqrt{P}) \quad f(y | +\sqrt{P}) \]
\[ \begin{array}{c}
-\sqrt{P} \\
\sqrt{P} \\
\end{array} \]
• Now to find the \textit{minimum} probability of error, consider

\[ P_e = P\{\hat{\Theta}(Y) \neq \Theta\} \]

\[ = P\{\Theta = \sqrt{P}\}P\{\hat{\Theta}(Y) = -\sqrt{P} | \Theta = \sqrt{P}\} + \]

\[ P\{\Theta = -\sqrt{P}\}P\{\hat{\Theta}(Y) = \sqrt{P} | \Theta = -\sqrt{P}\} \]

\[ = \frac{1}{2}P\{Y \leq 0 | \Theta = \sqrt{P}\} + \frac{1}{2}P\{Y > 0 | \Theta = -\sqrt{P}\} \]

\[ = \frac{1}{2}P\{Z \leq -\sqrt{P}\} + \frac{1}{2}P\{Z > \sqrt{P}\} \]

\[ = Q\left(\sqrt{\frac{P}{N}}\right) = Q\left(\sqrt{\text{SNR}}\right) \]

The probability of error is a decreasing function of \( P/N \), the \textit{signal-to-noise ratio} (SNR)