EE 278B in EE Curriculum

- EE 278B is one of three core ISL graduate courses:
  - EE 261 and EE 263 deal with deterministic (linear) systems
  - EE 278B deals with statistical systems
- EE 278B is also a prerequisite to courses in signal processing, image and video processing, communications, stochastic control, and machine learning. It also provides a good background for other areas (e.g., noise in devices, circuits, biological systems, . . . )
Statistical Signal Processing

- Focus is on extracting *information (signals)* from *noisy observations*
- Applications are all around us—cell phones, digital cameras, base stations, digital TV, DVD, . . .
- Generic signal processing problem:
  \[ X(t) \xrightarrow{\text{noisy channel}} Y(t) \xrightarrow{\text{signal processor}} \hat{X}(t) \]
  - Signal: (coded) digital data, audio, image, video, geophysical, medical, sensor . . .
  - Channel: twisted pair, optical, wireless, satellite, computer memory, electronic circuit, layers of earth, biological, . . .
  - Channel is modeled as a *statistical system*—linear vs. nonlinear, time invariant vs. time varying
  - Noise (physically generated or due to interference) and often signals are modeled as *random processes*, i.e., sequences of random variables indexed by time

- Signal processor: attempts to recover the signal from observation via
  - *estimation*: find an estimate that is close to the signal \( X(t) \), for example, one that minimizes the *mean square error* (MSE)
  - *detection*: decide which signal out of a finite number of possible signals (e.g., 0 and 1) was sent—goal: minimize the *probability of error*

- Statistical signal processing deals with both *modeling* of signals and channels and *design* of "optimal" signal processing algorithms
- Example: many real channels (twisted pair, wireless model, . . .) are modeled as linear time-invariant (LTI) system with additive noise

\[ Y(t) = h(t) \ast X(t) + Z(t) \]
• Very simple case: \( h(t) = \delta(t) \), \( X(t) \) binary waveform

\[ X(t) = \begin{cases} 1 & \text{for} \ t = 2k, \\ -1 & \text{for} \ t = 2k + 1 \end{cases} \]

\[ Y(t) = \begin{cases} \text{random noise} & \text{for} \ t = 2k, \\ \text{signal} & \text{for} \ t = 2k + 1 \end{cases} \]

\[ \hat{X}(t) = \begin{cases} 1 & \text{for} \ t = 2k, \\ -1 & \text{for} \ t = 2k + 1 \end{cases} \]

---

**Course Goal**

• To provide the statistical signal processing background:
  ○ provide relevant random vectors and processes background
  ○ introduce statistical models for noise and signals
  ○ introduce detection and estimation

• Courses that require EE278B as prerequisite include:
  ○ communications courses: EE276, EE279, EE376A,B and EE379A,B
  ○ signal and image processing and estimation courses: EE378, EE355, EE359, EE363, EE368, EE372

• Prerequisites:
  EE178/278A or equivalent,
  linear systems and transforms, e.g., EE102A,B

• EE 278B may not provide sufficient background for research in communication or signal processing; more mathematical courses, e.g., Stats 217, 218, 310A,B, C may be needed
Course Topics

- **Probability and random variables:** Lectures Notes 1 and 2; lectures 1–4. Axioms, basic laws, conditional probability, Bayes rule, and independence. Random variables; cumulative distribution function, probability mass function, probability density function, joint, marginal and conditional distributions, functions of random variables. Applications: Generation of random variables, scalar detection. Expectation; mean, variance, covariance and correlation. Inequalities; Markov and Chebyshev. Scalar MSE estimation; linear estimation and orthogonality principle.


- **Convergence and limit theorems:** Lecture notes 5; lecture 10.


Lecture Notes

- Help to organize and reduce note taking in lectures
- You will need to take some notes, e.g., clarifications, missing steps in derivations, solutions to additional examples
- Slide title indicates a topic that often continues over several consecutive slides
- Lecture notes + your notes + review sessions should be sufficient. (You may want to refer to textbooks for more explanations or different approaches)
- These lecture notes are always evolving. Please give me feedback on what can be improved
The following books should be on reserve at the Engineering Library:

- Leon-Garcia, *Probability and Random Processes for Electrical Engineers*


In addition, the following resources are available online:


- Gray and Davisson, *An Introduction to Statistical Signal Processing*, also available through Prof. Robert Gray’s webpage and EE 278 webpage.
Lecture Notes 1  
Probability and Random Variables

- Probability Spaces
- Conditional Probability and Independence
- Random Variables
- Functions of a Random Variable
- Generation of a Random Variable
- Jointly Distributed Random Variables
- Scalar detection

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:
  - \textit{Sample space} $\Omega$: set of all possible “elementary” or “finest grain” outcomes of the random experiment
  - \textit{Set of events} $\mathcal{F}$: set of (all?) subsets of $\Omega$ — an event $A \subseteq \Omega$ occurs if the outcome $\omega \in A$
  - \textit{Probability measure} $P$: function over $\mathcal{F}$ that assigns probabilities to events according to the axioms of probability (see below)
- Formally, a \textit{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:

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

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

• The probability of any other event $A$ is simply

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

• Example: For the die rolling experiment, assign

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

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

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

• For $\Omega = \mathbb{R} = (-\infty, \infty)$ (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 $\mathbb{R}$ 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 *a posteriori* probability of \( A \) — is related to the unconditional probability of \( A \) — the *a priori* probability — by
  \[ P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \]
  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
  \[ P(A_j \mid B) = \frac{P(B \mid A_j) P(A_j)}{P(B)} \]
  - 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
  \[ P(A_j \mid B) = \frac{P(B \mid A_j) P(A_j)}{\sum_{i=1}^{n} P(A_i) P(B \mid A_i)} P(A_j), \ j = 1, 2, \ldots, n \]
- 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., $X : \Omega \to \mathbb{R}$

\[ \Omega \]

\[ \omega \]

\[ X(\omega) \]

- 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 \( F_X(a) \geq F_X(b) \)

  \[ F_X(x) \]

  \[ 1 \]

  \[ x \]

- 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 **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 **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:

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

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

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

- **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

$$p_X(k) \to \frac{\lambda^k}{k!} e^{-\lambda} \quad \text{as } n \to \infty$$
A random variable is said to be *continuous* if its cdf is a continuous function

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

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

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

**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:
  
  ◦ **Uniform**: $X \sim U[a, b]$ where $a < b$ has pdf
    \[
    f_X(x) = \begin{cases} 
      \frac{1}{b-a} & \text{if } a \leq x \leq b \\
      0 & \text{otherwise}
    \end{cases}
    \]
  
  ◦ **Exponential**: $X \sim \text{Exp}(\lambda)$ where $\lambda > 0$ has pdf
    \[
    f_X(x) = \begin{cases} 
      \lambda e^{-\lambda x} & \text{if } x \geq 0 \\
      0 & \text{otherwise}
    \end{cases}
    \]
  
  ◦ **Laplace**: $X \sim \text{Laplace}(\lambda)$ where $\lambda > 0$ has pdf
    \[
    f_X(x) = \frac{1}{2} e^{-\lambda |x|}
    \]
  
  ◦ **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
  $$F_Y(-a) = F_X(-1) = \int_{-\infty}^{-1} \frac{1}{2} e^x \, dx = \frac{1}{2} e^{-1}$$
- $y = +1$: Here
  $$F_Y(+1) = F_X(+1) = \int_{+1}^{\infty} \frac{1}{2} e^x \, dx = \frac{1}{2} e^{-1}$$
- $y = +a$: Here
  $$F_Y(+a) = F_X(0) = \frac{1}{2}$$
- $y > +a$: Here
  $$F_Y(y) = 1$$

Diagram:

- $x$-axis from $-a$ to $+a$.
- $y$-axis from $-1$ to $+1$.
- Points at $(-1, -a)$, $(0, 0)$, $(+1, +a)$.
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

\[
x = F(y)
\]

\[
y = F^{-1}(x)
\]

• 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

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

\[
x = F(y)
\]

\[
y
\]

\[
1-p
\]

\[
0
\]

\[
1
\]

• 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 \( \mathbb{R}^2 \)

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

- \( \lim_{x,y \to \infty} F_{X,Y}(x, y) = 1 \)

• \( X \) 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:
  
  $$p_{X,Y}(x, y) = \Pr\{X = x, Y = y\}, \quad x \in \mathcal{X}, \ y \in \mathcal{Y}$$

  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

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

\[
= \lim_{\Delta x \to 0} \frac{P\{Y \leq y, x < X \leq x + \Delta x\}}{P\{x < X \leq x + \Delta x\}}
\]

\[
= \lim_{\Delta x \to 0} \int_{-\infty}^{y} \frac{f_{X,Y}(x,u) \, du \, \Delta x}{f_X(x) \, \Delta x}
\]

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

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

• Thus

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

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

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

• 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., \( F_{Y\mid \Theta}(y\mid \theta) \) is continuous for all \( \theta \). We define \( f_{Y\mid \Theta}(y\mid \theta) \) in the usual way.

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

\[
p_{\Theta\mid Y}(\theta \mid y) = \lim_{\Delta y \to 0} \frac{\frac{\Pr\{\Theta = \theta, y < Y \leq y + \Delta y\}}{\Pr\{y < Y \leq y + \Delta y\}}}{\frac{f_Y(y)\Delta y}{f_Y(y)} = \frac{f_{Y\mid \Theta}(y\mid \theta)p_{\Theta}(\theta)}{f_Y(y)}}
\]

This leads to the Bayes rule:

\[
p_{\Theta\mid Y}(\theta \mid y) = \frac{f_{Y\mid \Theta}(y\mid \theta)}{\sum_{\theta'} p_{\Theta}(\theta') f_{Y\mid \Theta}(y\mid \theta')} p_{\Theta}(\theta)
\]
Example: Additive Gaussian Noise Channel
Consider the following communication channel:

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

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

\[ \Theta \in \{ \theta_0, \theta_1 \} \quad \text{noisy channel} \quad Y \quad \rightarrow \quad \hat{\Theta}(Y) \in \{ \theta_0, \theta_1 \} \]

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:

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

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

- 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

\[
\hat{\Theta}(y) = \begin{cases} \theta_0 & \text{if } f_{Y|\Theta}(y|\theta_0) > f_{Y|\Theta}(y|\theta_1) \\ \theta_1 & \text{otherwise} \end{cases}
\]
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
Now to find the 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 signal-to-noise ratio (SNR)
Expectation

- Let $X \in \mathcal{X}$ be a discrete r.v. with pmf $p_X(x)$ and let $g(x)$ be a function of $x$. The expectation (or expected value or mean) of $g(X)$ can be defined as

$$E(g(X)) = \sum_{x \in \mathcal{X}} g(x)p_X(x)$$

- For a continuous r.v. $X \sim f_X(x)$, the expected value of $g(X)$ can be defined as

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f_X(x) \, dx$$

- Expectation is linear, i.e., for any constant $a$

$$E[ag_1(X) + g_2(X)] = a E(g_1(X)) + E(g_2(X))$$

In particular, $E(a) = a$
• Remark: We know that a r.v. is completely specified by its cdf (pdf, pmf), so why do we need expectation?
  ○ Expectation provides a summary or an estimate of the r.v. — a single number — instead of specifying the entire distribution
  ○ It is far easier to estimate the expectation of a r.v. from data than to estimate its distribution
  ○ Expectation can be used to bound or estimate probabilities of interesting events (as we shall see)

Mean and Variance

• The first moment (or mean) of $X \sim f_X(x)$ is
  \[ E(X) = \int_{-\infty}^{\infty} x f_X(x) \, dx \]

• The second moment (or mean squared or average power) of $X$ is
  \[ E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x) \, dx \]

• The variance of $X$ is
  \[ \text{Var}(X) = E[(X - E(X))^2] = E(X^2) - (E(X))^2 \]
  Hence $E(X^2) \geq (E(X))^2$

• The standard deviation of $X$ is defined as $\sigma_X = \sqrt{\text{Var}(X)}$, i.e., $\text{Var}(X) = \sigma_X^2$

• In general, the $k$th moment (\(k\) a positive integer) is
  \[ E(X^k) = \int_{-\infty}^{\infty} x^k f_X(x) \, dx \]
### Mean and Variance for Famous RVs:

<table>
<thead>
<tr>
<th>Random Variable</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bern($p$)</td>
<td>$p$</td>
<td>$p(1 - p)$</td>
</tr>
<tr>
<td>Geom($p$)</td>
<td>$\frac{1}{p}$</td>
<td>$\frac{1 - p}{p^2}$</td>
</tr>
<tr>
<td>Binom($n, p$)</td>
<td>$np$</td>
<td>$np(1 - p)$</td>
</tr>
<tr>
<td>Poisson($\lambda$)</td>
<td>$\lambda$</td>
<td>$\lambda$</td>
</tr>
<tr>
<td>$U[a, b]$</td>
<td>$\frac{a + b}{2}$</td>
<td>$\frac{(b - a)^2}{12}$</td>
</tr>
<tr>
<td>Exp($\lambda$)</td>
<td>$\frac{1}{\lambda}$</td>
<td>$\frac{1}{\lambda^2}$</td>
</tr>
<tr>
<td>Laplace($\lambda$)</td>
<td>$0$</td>
<td>$\frac{2}{\lambda^2}$</td>
</tr>
<tr>
<td>$N(\mu, \sigma^2)$</td>
<td>$\mu$</td>
<td>$\sigma^2$</td>
</tr>
</tbody>
</table>

### Expectation Can Be Infinite or May Not Exist

- **Expectation can be infinite.** For example,

  $$f_X(x) = \begin{cases} 
  \frac{1}{x^2} & 1 \leq x < \infty \\
  0 & \text{otherwise}
  \end{cases} \Rightarrow E(X) = \int_1^\infty \frac{x}{x^2} \, dx = \infty$$

- **Expectation may not exist.** To find conditions for expectation to exist, consider

  $$E(X) = \int_{-\infty}^{\infty} x f_X(x) \, dx = -\int_{-\infty}^{0} |x| f_X(x) \, dx + \int_{0}^{\infty} |x| f_X(x) \, dx,$$

  so either $\int_{-\infty}^{0} |x| f_X(x) \, dx$ or $\int_{0}^{\infty} |x| f_X(x) \, dx$ must be finite.

- **Example:** The *standard Cauchy* r.v. has the pdf

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

  Since both $\int_{-\infty}^{0} |x| f(x) \, dx$ and $\int_{0}^{\infty} |x| f(x) \, dx$ are infinite, its mean does not exist! (The second moment of the Cauchy is $E(X^2) = \infty$, so it exists)
Bounding Probability Using Expectation

- In many cases we do not know the distribution of a r.v. $X$ but want to find the probability of an event such as $\{X > a\}$ or $\{|X - E(X)| > a\}$

- The Markov and Chebyshev inequalities give upper bounds on the probabilities of such events in terms of the mean and variance of the random variable

- Example: Let $X \geq 0$ represent the age of a person in the Bay Area. If we know that $E(X) = 35$ years, what fraction of the population is $\geq 70$ years old?

  Clearly we cannot answer this question knowing only the mean, but we can say that $P\{X \geq 70\} \leq 0.5$, since otherwise the mean would be larger than 35

- This is an application of the Markov inequality

Markov Inequality

- For any r.v. $X \geq 0$ with finite mean $E(X)$ and any $a > 1$,

  $$P\{X \geq a E(X)\} \leq \frac{1}{a}$$

  **Proof:** Define the *indicator function* of the set $A = \{x \geq a E(X)\}$:

  $$I_A(x) = \begin{cases} 1 & x \geq a E(X) \\ 0 & \text{otherwise} \end{cases}$$

  Clearly, $I_A \leq \frac{X}{a E(X)}$

  Since $E(I_A) = P(A) = P\{X \geq a E(X)\}$, taking the expectations of both sides we obtain the Markov Inequality
Chebyshev Inequality

- Let $X$ be a device parameter in an integrated circuit (IC) with known mean and variance. The IC is out-of-spec if $X$ is more than, say, $3\sigma_X$ away from its mean. We wish to find the fraction of out-of-spec ICs, namely, $P\{|X - E(X)| \geq 3\sigma_X\}$.

The Chebyshev inequality gives us an upper bound on this fraction in terms the mean and variance of $X$.

- Let $X$ be a r.v. with known $E(X)$ and $\text{Var}(X) = \sigma_X^2$. The Chebyshev inequality states that for every $a > 1$,

$$P\{|X - E(X)| \geq a\sigma_X\} \leq \frac{1}{a^2}$$

Proof: We use the Markov inequality. Define the r.v. $Y = (X - E(X))^2 \geq 0$.

Since $E(Y) = \sigma_X^2$, the Markov inequality gives

$$P\{Y \geq a^2\sigma_X^2\} \leq \frac{1}{a^2}$$

But $\{|X - E(X)| \geq a\sigma_X\}$ occurs iff $\{Y \geq a^2\sigma_X^2\}$. Thus

$$P\{|X - E(X)| \geq a\sigma_X\} \leq \frac{1}{a^2}$$

Expectation Involving Two RVs

- Let $(X, Y) \sim f_{X,Y}(x, y)$ and let $g(x, y)$ be a function of $x$ and $y$. The expectation of $g(X, Y)$ is given by

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

The function $g(X, Y)$ may be $X, Y, X^2, X + Y$, etc.

- The correlation of $X$ and $Y$ is defined as $E(XY)$

$X$ and $Y$ are said to be orthogonal if $E(XY) = 0$.

- The covariance of $X$ and $Y$ is defined as

$$\text{Cov}(X, Y) = E(XY) - E(X) E(Y)$$

$X$ and $Y$ are said to be uncorrelated if $\text{Cov}(X, Y) = 0$.

- Note that $\text{Cov}(X, X) = \text{Var}(X)$.

- If $X$ and $Y$ are independent then they are uncorrelated.

- $X$ and $Y$ uncorrelated does not necessarily imply that they are independent.
• The correlation coefficient of $X$ and $Y$ is defined as
\[ \rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \]

Fact: $|\rho_{X,Y}| \leq 1$ with equality iff $(X - \text{E}(X))$ is a linear function of $(Y - \text{E}(Y))$

The correlation coefficient is a measure of how closely $(X - \text{E}(x))$ can be approximated by a linear function of $(Y - \text{E}(Y))$ (more on this soon)

---

Conditional Expectation

• Let $(X, Y) \sim f_{X,Y}(x, y)$. If $f_Y(y) \neq 0$, the conditional pdf of $X$ given $Y = y$ is given by
\[ f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \]

• We know that $f_{X|Y}(x|y)$ is a pdf for $X$ (function of $y$), so we can define the expectation of any function $g(X, Y)$ w.r.t. $f_{X|Y}(x|y)$ as
\[ \text{E}(g(X, Y) \mid Y = y) = \int_{-\infty}^{\infty} g(x, y) f_{X|Y}(x|y) \, dx \]

• Example: If $g(X, Y) = X$, then the conditional expectation of $X$ given $Y = y$ is
\[ \text{E}(X \mid Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) \, dx \]

• Example: If $g(X, Y) = XY$, then $\text{E}(XY \mid Y = y) = y \text{E}(X \mid Y = y)$

• We define the conditional expectation of $g(X, Y)$ given $Y$ as the random variable $\text{E}(g(X, Y) \mid Y)$, which is a function of the random variable $Y$
• In particular, $E(X \mid Y)$ is the conditional expectation of $X$ given $Y$, a r.v. that is a function of $Y$.

• *Iterated expectation*: In general we can find $E(g(X, Y))$ as

\[
E(g(X, Y)) = E_Y \left[ E_X(g(X, Y) \mid Y) \right],
\]

where $E_X$ means expectation w.r.t. $f_{X \mid Y}(x \mid y)$ and $E_Y$ means expectation w.r.t. $f_Y(y)$.

• Example: *Coin with random bias*. A coin with random bias $P$ such that $E(P) = 1/3$ is flipped $n$ times independently. Let $X$ be the number of heads. Find $E(X)$.

---

**Conditional Variance**

• Let $X$ and $Y$ be two r.v.s. We define the *conditional variance* of $X$ given $Y = y$ to be the variance of $X$ using $f_{X \mid Y}(x \mid y)$, i.e.,

\[
\text{Var}(X \mid Y = y) = E \left[ (X - E(X \mid Y = y))^2 \right] \mid Y = y
\]

\[
= E(X^2 \mid Y = y) - [E(X \mid Y = y)]^2
\]

• The r.v. $\text{Var}(X \mid Y)$ is simply a function of $Y$ that takes on the values $\text{Var}(X \mid Y = y)$. Its expected value is

\[
E_Y \left[ \text{Var}(X \mid Y) \right] = E_Y \left[ E(X^2 \mid Y) - (E(X \mid Y))^2 \right] = E(X^2) - E \left[ (E(X \mid Y))^2 \right]
\]

• Since $E(X \mid Y)$ is a r.v., it has a variance

\[
\text{Var}(E(X \mid Y)) = E_Y \left[ (E(X \mid Y) - E[E(X \mid Y)])^2 \right] = E \left[ (E(X \mid Y))^2 \right] - (E(X))^2
\]

• *Law of Conditional Variances*: Adding the above expressions, we obtain

\[
\text{Var}(X) = E(\text{Var}(X \mid Y)) + \text{Var}(E(X \mid Y))
\]
Scalar MSE Estimation

- Consider the following signal processing problem:

\[ X \xrightarrow{f_X(x)} \text{Noisy Channel} \xrightarrow{f_{Y|X}(y|x)} Y \xrightarrow{\text{Estimator}} \hat{X}(Y) \]

- \( X \) is a signal with known statistics, i.e., known pdf \( f_X(x) \)
- The signal is transmitted (or stored) over a noisy channel with known statistics, i.e., conditional pdf \( f_{Y|X}(y|x) \)
- We observe the channel output \( Y \) and wish to find the estimate \( \hat{X}(Y) \) of \( X \) that minimizes the mean squared error
  \[
  \text{MSE} = E[(X - \hat{X}(Y))^2]
  \]
- The \( \hat{X} \) that achieves the minimum MSE is called the MMSE estimate of \( X \) (given \( Y \))

MMSE Estimate

- Theorem: The MMSE estimate of \( X \) given the observation \( Y \) and complete knowledge of the joint pdf \( f_{X,Y}(x, y) \) is

\[
\hat{X}(Y) = E(X | Y),
\]

and the MSE of \( \hat{X} \), i.e., the minimum MSE, is

\[
\text{MMSE} = E_Y(\text{Var}(X | Y)) = \text{Var}(X) - \text{Var}(E(X | Y))
\]

- Properties of the minimum MSE estimator:
  - Since \( E(\hat{X}) = E_Y[E(X | Y)] = E(X) \), the MMSE estimate is unbiased
  - If \( X \) and \( Y \) are independent, then the MMSE estimate is \( E(X) \)
  - The conditional expectation of the estimation error \( E[(X - \hat{X}) | Y = y] = 0 \) for every \( y \), i.e., the error is unbiased for every \( Y = y \)
The estimation error and the estimate are orthogonal
\[
E \left[ (X - \hat{X})\hat{X} \right] = E_Y \left[ E \left( (X - \hat{X})\hat{X} | Y \right) \right]
\]
\[
= E_Y \left[ \hat{X} E((X - \hat{X}) | Y) \right]
\]
\[
= E_Y \left[ \hat{X}(E(X | Y) - \hat{X}) \right]
\]
\[
= 0
\]
In fact, the estimation error is orthogonal to any function \( g(Y) \) of \( Y \)

MMSE estimate is linear: Let \( X = aU + V \) and \( \hat{U} \) and \( \hat{V} \) be the MMSE estimates of \( U \) and \( V \), respectively

Then, the MMSE estimate of \( X \) is
\[
\hat{X} = a\hat{U} + \hat{V}
\]

Proof of Theorem: We first show that \( \min_a E (X - a)^2 = \text{Var}(X) \) and that the minimum is achieved for \( a = E(X) \), i.e., in the absence of any observations, the mean of \( X \) is its MMSE estimate

To show this, consider
\[
E \left[ (X - a)^2 \right] = E \left[ (X - E(X) + E(X) - a)^2 \right]
\]
\[
= E \left[ (X - E(X))^2 \right] + (E(X) - a)^2 + 2E(X - E(X))(E(X) - a)
\]
\[
= E \left[ (X - E(X))^2 \right] + (E(X) - a)^2 \geq E \left[ (X - E(X))^2 \right]
\]
Equality holds if and only if \( a = E(X) \)

We use this result to show that \( E(X | Y) \) is the MMSE estimate of \( X \) given \( Y \)

First write
\[
E \left[ (X - \hat{X}(Y))^2 \right] = E_Y \left[ E_X((X - \hat{X}(Y))^2 | Y) \right]
\]

From the previous result we know that for every \( Y = y \) the minimum value for
\[
E_X \left[ (X - \hat{X}(y))^2 | Y = y \right]
\]

is obtained when \( \hat{X}(y) = E(X | Y = y) \)

Therefore the overall MSE is minimized for \( \hat{X}(Y) = E(X | Y) \)

In fact, \( E(X | Y) \) minimizes the MSE conditioned on every \( Y = y \) and not just its average over \( Y \)
To find the minimum MSE, consider
\[
E \left[ (X - E(X | Y))^2 \right] = E_Y \left( E_X \left[ (X - E(X | Y))^2 | Y \right] \right)
= E_Y \left( \text{Var}(X | Y) \right)
\]

- Finally, by the law of conditional variance,
\[
E(\text{Var}(X | Y)) = \text{Var}(X) - E(\text{Var}(X | Y)),
\]
i.e., the minimum MSE is the difference between the variance of the signal and the variance of the MMSE estimate.

The Additive Gaussian Noise Channel

- Consider a communication channel with input \( X \sim \mathcal{N}(\mu, P) \), noise \( Z \sim \mathcal{N}(0, N) \), and output \( Y = X + Z \). \( X \) and \( Z \) are independent.

Find the MMSE estimate of \( X \) given \( Y \) and its MSE, i.e., \( E(X | Y) \) and \( E(\text{Var}(X | Y)) \)

- To find \( f_{X|Y}(x|y) \) we use Bayes rule:
\[
f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)}{f_Y(y)} f_X(x)
\]

We know that \( X \sim \mathcal{N}(\mu, P) \), and since \( X \) and \( Z \) are independent and Gaussian, \( Y = X + Z \sim \mathcal{N}(\mu, P + N) \) (to be proved later).

To find \( f_{Y|X}(y|x) \), since \( Y \) is the sum of two independent r.v.s, we have
\[
f_{Y|X}(y|x) = f_{Z|X}(y-x|x) = f_Z(y-x) = \frac{1}{\sqrt{2\pi N}} e^{-\frac{(y-x)^2}{2N}}
\]

In other words, \( Y | \{X = x\} \sim \mathcal{N}(x, N) \)
Substituting in the Bayes rule formula, we finally obtain

\[ f_{X|Y}(x|y) = \frac{1}{\sqrt{2\pi \sigma_{X|Y}^2}} e^{-\frac{(x - (\frac{P}{P+N} y + \frac{N}{P+N} \mu))^2}{2\sigma_{X|Y}^2}} \]

that is,

\[ X | \{ Y = y \} \sim \mathcal{N}\left( \frac{P}{P+N} y + \frac{N}{P+N} \mu, \frac{P N}{P+N} \right) \]

Thus

\[ E(X | Y) = \frac{P}{P+N} Y + \frac{N}{P+N} \mu \]

\[ E(\text{Var}(X | Y)) = \frac{P N}{P+N} \]

---

Scalar Linear Estimation

- To find the MMSE estimate one needs to know the statistics of the signal and the channel — \( f_{X,Y}(x, y) \) — which is rarely the case in practice

- We typically have estimates only of the first and second moments of the signal and the observation, i.e., means, variances, and covariance of \( X \) and \( Y \)

- This is not, in general, sufficient information for computing the MMSE estimate, but as we shall see is enough to compute the MMSE linear (or affine) estimate of the signal \( X \) given the observation \( Y \), i.e., the estimate of the form

\[ \hat{X} = aY + b \]

that minimizes the mean squared error

\[ \text{MSE} = E \left[ (X - \hat{X})^2 \right] \]
The MMSE Linear Estimate

- Theorem: The MMSE linear estimate of $X$ given $Y$ is
  \[
  \hat{X} = \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} (Y - \text{E}(Y)) + \text{E}(X)
  \]
  \[
  = \rho_{X,Y} \sigma_X \left( \frac{Y - \text{E}(Y)}{\sigma_Y} \right) + \text{E}(X)
  \]
  and its MSE is given by
  \[
  \text{MSE} = \text{Var}(X) - \frac{\text{Cov}^2(X, Y)}{\text{Var}(Y)} = (1 - \rho_{X,Y}^2) \text{Var}(X)
  \]

- Properties of MMSE linear estimate:
  - $\text{E}(\hat{X}) = \text{E}(X)$, i.e., estimate is unbiased (also true for MMSE estimate)
  - If $\rho_{X,Y} = 0$, i.e., $X$ and $Y$ are uncorrelated, then $\hat{X} = \text{E}(X)$ — the observation $Y$ is ignored!
  - If $\rho_{X,Y} = \pm 1$, i.e., $(X - \text{E}(X))$ and $(Y - \text{E}(Y))$ are linearly dependent, then the MMSE linear estimate is perfect

- MMSE linear estimate is linear: Let $X = aU + V$ and $\hat{U}$ and $\hat{V}$ be the MMSE linear estimates of $U$ and $V$, respectively.
  Then, the MMSE linear estimate of $X$ is
  \[
  \hat{X} = a\hat{U} + \hat{V}
  \]

- Proof of the Theorem: From MMSE estimate derivation with no observation, we know that the MMSE estimate of $(X - aY)$ is its mean $\text{E}(X) - a \text{E}(Y)$; hence we can replace $b$ with $\text{E}(X) - a \text{E}(Y)$, which reduces the linear estimation problem to finding the coefficient $a$ that minimizes
  \[
  \text{E}[(X - \text{E}(X)) - a(Y - \text{E}(Y))]^2 = \text{E}[(X - \text{E}(X)) - (\hat{X} - \text{E}(X))]^2,
  \]
  i.e., the problem reduces to finding $(\hat{X} - \text{E}(X)) = a(Y - \text{E}(Y))$ that minimizes the MSE
  This problem can be solved using calculus. Instead we use a geometric argument that will help us solve more involved linear estimation problems
Geometric Formulation of Linear Estimation

• First we introduce some needed background

• A vector space \( \mathcal{V} \), e.g., Euclidean space, consists of a set of vectors that are closed under two operations:
  ○ vector addition: if \( v_1, v_2 \in \mathcal{V} \) then \( v_1 + v_2 \in \mathcal{V} \)
  ○ scalar multiplication: if \( a \in \mathbb{R} \) and \( v \in \mathcal{V} \), then \( av \in \mathcal{V} \)

• An inner product, e.g., dot product in Euclidean space, is a real-valued operation \( u \cdot v \) satisfying the three conditions:
  ○ commutativity: \( u \cdot v = v \cdot u \)
  ○ linearity: \( (au + v) \cdot w = a(u \cdot w) + v \cdot w \)
  ○ nonnegativity: \( u \cdot u \geq 0 \) and \( u \cdot u = 0 \) iff \( u = 0 \)

• The norm of \( u \) is defined as \( \|u\| = \sqrt{u \cdot u} \)

• \( u \) and \( v \) are orthogonal (written \( u \perp v \)) if \( u \cdot v = 0 \)

• A vector space with an inner product is called an inner product space
  Example: Euclidean space with dot product

Back to Linear Estimation

• View \((X - \mathbb{E}(X))\) and \((Y - \mathbb{E}(Y))\) as vectors in an inner product space \( \mathcal{V} \) that consists of all zero mean random variables defined over the same probability space, with
  ○ vector addition: \( V_1 + V_2 \in \mathcal{V} \)
    adding two zero mean r.v.s yields a zero mean r.v.
  ○ scalar multiplication: \( aV \in \mathcal{V} \)
    multiplying a zero mean r.v. by a constant yields a zero mean r.v.
  ○ inner product: \( \mathbb{E}(V_1 V_2) \)
    exercise: check that this is a legitimate inner product
  ○ norm of \( V \): \( \|V\| = \sqrt{\mathbb{E}(V^2)} = \sigma_V \)
So we have the following picture for the r.v.s $(X - E(X))$ and $(Y - E(Y))$:

\begin{align*}
(X - E(X)) \\
\theta \\
(Y - E(Y))
\end{align*}

inner product $\Leftrightarrow \text{Cov}(X, Y)$

norm of $(X - E(X))$ $\Leftrightarrow \sigma_X$

norm of $(Y - E(Y))$ $\Leftrightarrow \sigma_Y$

$\cos \theta$ $\Leftrightarrow \rho_{X,Y}$

Note that $(X - E(X))$ and $(Y - E(Y))$ can live in a vector space of very high dimension. We are interested only in the 2-dimensional subspace spanned by these two vectors.

Orthogonality Principle

- The linear estimation problem can now be recast as a geometry problem

\begin{align*}
(X - E(X)) \\
\text{error } X - \hat{X} \\
(\hat{X} - E(X)) \\
(Y - E(Y))
\end{align*}

Find a vector $(\hat{X} - E(X)) = a(Y - E(Y))$ that minimizes $\|X - \hat{X}\|$

- Clearly $(X - \hat{X}) \perp (Y - E(Y))$ minimizes $\|X - \hat{X}\|$, i.e.,

\[
E((X - \hat{X})(Y - E(Y))) = 0 \Rightarrow a = \frac{\text{Cov}(X, Y)}{\text{Var}(Y)}
\]

- This argument is called the orthogonality principle. Later we will see that it is key to deriving the MMSE linear estimate in more complex settings.
Linear vs. MMSE (Nonlinear) Estimate

- The linear estimate is not, in general, as good as the MMSE estimate
- Example: Let \( Y \sim U[-1, 1] \) and \( X = Y^2 \)
  The MMSE estimate of \( X \) given \( Y \) is \( Y^2 \) — perfect!
  To find the MMSE linear estimate we compute
  \[
  E(Y) = 0 \\
  E(X) = \int_{-1}^{1} \frac{1}{2} y^2 \, dy = \frac{1}{3} \\
  \text{Cov}(X, Y) = E(XY) - 0 = E(Y^3) = 0
  \]
  Thus the MMSE linear estimate \( \hat{X} = E(X) = 1/3 \), i.e., the observation \( Y \) is totally ignored, even though it completely determines \( X \)!
- There is a very important class of r.v.s for which the MMSE estimate is linear, the jointly Gaussian random variables

Jointly Gaussian Random Variables

- Two r.v.s are jointly Gaussian if their joint pdf is of the form
  \[
  f(x, y) = \frac{1}{2\pi \sigma_X \sigma_Y \sqrt{1 - \rho^2_{X,Y}}} e^{-\frac{1}{2(1-\rho^2_{X,Y})} \left( \frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\rho_{X,Y} \frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X \sigma_Y} \right)}
  \]
- The pdf is a function only of \( \mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \) and \( \rho_{X,Y} \)
- Note: In Lecture Notes 3 we will define this in a more general way
- Example: For the additive Gaussian noise channel, where \( X \sim N(\mu, P) \) and \( Z \sim N(0, N) \) are independent and \( Y = X + Z \), show that (a) \( X \) and \( Z \) are jointly Gaussian, and (b) \( X \) and \( Y \) are jointly Gaussian
  Solution: (a) It is easy to show that if two Gaussian r.v.s are independent, their joint pdf has the above form with \( \rho_{X,Z} = 0 \). (b) Now consider
  \[
  f(x, y) = f_X(x) f_Y|X(y|x) \\
  = f_X(x) f_Z|X(y-x|x) = f_X(x) f_Z(y-x)
  \]
  Now we can write \( f(x, y) \) in the form of a jointly Gaussian pdf
• If $X$ and $Y$ are jointly Gaussian, contours of equal joint pdf are ellipses defined by the quadratic equation

$$\frac{(x - \mu_X)^2}{\sigma_X^2} + \frac{(y - \mu_Y)^2}{\sigma_Y^2} - 2\rho_{X,Y}\frac{(x - \mu_X)(y - \mu_Y)}{\sigma_X\sigma_Y} = c \geq 0$$

• Examples: In the following examples we plot contours of equal joint pdf $f(x, y)$ for zero mean jointly Gaussian r.v.s for different values of $\sigma_X$, $\sigma_Y$, and $\rho_{X,Y}$

The orientation of the major axis of the ellipse is $\theta = \frac{1}{2} \arctan \left( \frac{2\rho_{X,Y}\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2} \right)$
\[ \sigma_X = 1, \sigma_Y = 3, \rho_{X,Y} = 0.4 : \theta = 81.65^\circ \]

\[ \sigma_X = 1, \sigma_Y = 3, \rho_{X,Y} = 0.7 : \theta = 76.15^\circ \]

\[ \sigma_X = 1, \sigma_Y = 3, \rho_{X,Y} = 0.99 : \theta = 71.7^\circ \]

\[ \sigma_X = 1, \sigma_Y = 3, \rho_{X,Y} = -0.4 : \theta = -81.65^\circ \]
Properties of Jointly Gaussian Random Variables

- If \( X \) and \( Y \) are jointly Gaussian, they are individually Gaussian, i.e., the marginals of \( f_{X,Y}(x,y) \) are Gaussian, i.e.,
  \[ X \sim \mathcal{N}(\mu_X, \sigma_X^2), \quad Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2) \]

- The converse is not necessarily true, i.e., Gaussian marginals do not necessarily mean that the r.v.s are jointly Gaussian
  
  **Example:** Let \( X_1 \sim \mathcal{N}(0,1) \) and
  \[ X_2 = \begin{cases} +1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases} \]
  be independent r.v.s, and let \( X_3 = X_1X_2 \)
  - Clearly, \( X_3 \sim \mathcal{N}(0,1) \)
  - However, \( X_1, X_3 \) do not have a joint pdf. Using delta functions, “\( f_{X_1,X_3}(x_1,x_3) \)” has the form shown in the following figure

- If \( X \) and \( Y \) are jointly Gaussian, the conditional pdf is Gaussian:
  \[ X \mid \{Y = y\} \sim \mathcal{N}\left(\rho_{X,Y} \sigma_X \frac{(y - \mu_Y)}{\sigma_Y} + \mu_X, \ (1 - \rho_{X,Y}^2)\sigma_X^2 \right), \]
  which shows that the MMSE estimate is linear

- If \( X \) and \( Y \) are jointly Gaussian and uncorrelated, i.e., \( \rho_{X,Y} = 0 \), then they are also independent
Specifying a Random Vector

Let $X_1, X_2, \ldots, X_n$ be random variables defined on the same probability space. We define a random vector (RV) as

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

$X$ is completely specified by its joint cdf for $x = (x_1, x_2, \ldots, x_n)$:

$$F_X(x) = P\{X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n\}, \quad x \in \mathbb{R}^n$$

If $X$ is continuous, i.e., $F_X(x)$ is a continuous function of $x$, then $X$ can be specified by its joint pdf:

$$f_X(x) = f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n), \quad x \in \mathbb{R}^n$$

If $X$ is discrete then it can be specified by its joint pmf:

$$p_X(x) = p_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n), \quad x \in \mathcal{X}^n$$
• A marginal cdf (pdf, pmf) is the joint cdf (pdf, pmf) for a subset of \( \{X_1, \ldots, X_n\} \); e.g., for 
\[
\mathbf{X} = \begin{bmatrix}
X_1 \\
X_2 \\
X_3
\end{bmatrix}
\]
the marginals are
\[
f_{X_1}(x_1), \ f_{X_2}(x_2), \ f_{X_3}(x_3) \\
f_{X_1,X_2}(x_1,x_2), \ f_{X_1,X_3}(x_1,x_3), \ f_{X_2,X_3}(x_2,x_3)
\]
• The marginals can be obtained from the joint in the usual way. For the previous example,
\[
F_{X_1}(x_1) = \lim_{x_2,x_3 \to \infty} F_{\mathbf{X}}(x_1,x_2,x_3)
\]
\[
f_{X_1,X_2}(x_1,x_2) = \int_{-\infty}^{\infty} f_{X_1,X_2,X_3}(x_1,x_2,x_3) \, dx_3
\]

• Conditional cdf (pdf, pmf) can also be defined in the usual way. E.g., the conditional pdf of \( X_{k+1}^n = (X_{k+1}, \ldots, X_n) \) given \( X^k = (X_1, \ldots, X_k) \) is
\[
f_{X_{k+1}^n | X^k}(x_{k+1}^n | x^k) = \frac{f_{\mathbf{X}}(x_1, x_2, \ldots, x_n)}{f_{X^k}(x_1, x_2, \ldots, x_k)} = \frac{f_{\mathbf{X}}(x)}{f_{X^k}(x^k)}
\]
• **Chain Rule:** We can write
\[
f_{\mathbf{X}}(x) = f_{X_1}(x_1) f_{X_2 | X_1}(x_2 | x_1) f_{X_3 | X_1,X_2}(x_3 | x_1, x_2) \cdots f_{X_n | X_{n-1}}(x_n | x_{n-1})
\]
Proof: By induction. The chain rule holds for \( n = 2 \) by definition of conditional pdf. Now suppose it is true for \( n - 1 \). Then
\[
f_{\mathbf{X}}(x) = f_{X_{n-1}}(x_{n-1}) f_{X_n | X_{n-1}}(x_n | x_{n-1})
\]
\[
= f_{X_1}(x_1) f_{X_2 | X_1}(x_2 | x_1) \cdots f_{X_{n-1} | X_{n-2}}(x_{n-1} | x_{n-2}) f_{X_n | X_{n-1}}(x_n | x_{n-1}),
\]
which completes the proof.
Independence and Conditional Independence

- Independence is defined in the usual way; e.g., \( X_1, X_2, \ldots, X_n \) are independent if
  \[
  f_X(x) = \prod_{i=1}^{n} f_{X_i}(x_i) \quad \text{for all } (x_1, \ldots, x_n)
  \]

- Important special case, i.i.d. r.v.s: \( X_1, X_2, \ldots, X_n \) are said to be independent, identically distributed (i.i.d.) if they are independent and have the same marginals.
  
  Example: if we flip a coin \( n \) times independently, we generate i.i.d. Bern(p) r.v.s \( X_1, X_2, \ldots, X_n \)

- R.v.s \( X_1 \) and \( X_3 \) are said to be conditionally independent given \( X_2 \) if
  \[
  f_{X_1,X_3|X_2}(x_1, x_3|x_2) = f_{X_1|X_2}(x_1|x_2)f_{X_3|X_2}(x_3|x_2) \quad \text{for all } (x_1, x_2, x_3)
  \]

- Conditional independence neither implies nor is implied by independence; \( X_1 \) and \( X_3 \) independent given \( X_2 \) does not mean that \( X_1 \) and \( X_3 \) are independent (or vice versa)

- Example: **Coin with random bias.** Given a coin with random bias \( P \sim f_P(p) \), flip it \( n \) times independently to generate the r.v.s \( X_1, X_2, \ldots, X_n \), where \( X_i = 1 \) if \( i \)-th flip is heads, 0 otherwise.
  
  - \( X_1, X_2, \ldots, X_n \) are not independent
  
  - However, \( X_1, X_2, \ldots, X_n \) are conditionally independent given \( P \); in fact, they are i.i.d. Bern(p) for every \( P = p \)

- Example: **Additive noise channel.** Consider an additive noise channel with signal \( X \), noise \( Z \), and observation \( Y = X + Z \), where \( X \) and \( Z \) are independent.
  
  - Although \( X \) and \( Z \) are independent, they are not in general conditionally independent given \( Y \)
The mean of the random vector \( \mathbf{X} \) is defined as
\[
E(\mathbf{X}) = [E(X_1) \ E(X_2) \ \cdots \ E(X_n)]^T
\]

Denote the covariance between \( X_i \) and \( X_j \), \( \text{Cov}(X_i, X_j) \), by \( \sigma_{ij} \) (so the variance of \( X_i \) is denoted by \( \sigma_{ii} \), \( \text{Var}(X_i) \), or \( \sigma_{X_i}^2 \)).

The covariance matrix of \( \mathbf{X} \) is defined as
\[
\Sigma_X = \begin{bmatrix}
\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\
\sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn}
\end{bmatrix}
\]

For \( n = 2 \), we can use the definition of correlation coefficient to obtain
\[
\Sigma_X = \begin{bmatrix}
\sigma_{11} & \sigma_{12} \\
\sigma_{21} & \sigma_{22}
\end{bmatrix} = \begin{bmatrix}
\sigma_{X_1}^2 & \rho_{X_1,X_2}\sigma_{X_1}\sigma_{X_2} \\
\rho_{X_1,X_2}\sigma_{X_1}\sigma_{X_2} & \sigma_{X_2}^2
\end{bmatrix}
\]

Properties of Covariance Matrix \( \Sigma_X \)

\( \Sigma_X \) is real and symmetric (since \( \sigma_{ij} = \sigma_{ji} \)).

\( \Sigma_X \) is positive semidefinite, i.e., the quadratic form
\[
a^T \Sigma_X a \geq 0 \quad \text{for every real vector } a
\]

Equivalently, all the eigenvalues of \( \Sigma_X \) are nonnegative, and also all leading principal minors are nonnegative.

To show that \( \Sigma_X \) is positive semidefinite we write
\[
\Sigma_X = E \left[ (\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))^T \right],
\]

i.e., as the expectation of an outer product. Thus
\[
a^T \Sigma_X a = a^T E \left[ (\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))^T \right] a
= E \left[ a^T (\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))^T a \right]
= E \left[ (a^T(\mathbf{X} - E(\mathbf{X})))^2 \right] \geq 0
\]
Which of the Following Can Be a Covariance Matrix?

1. \[
\begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{bmatrix}
\]

2. \[
\begin{bmatrix}
1 & 2 & 1 \\
2 & 1 & 1 \\
1 & 1 & 1
\end{bmatrix}
\]

3. \[
\begin{bmatrix}
1 & 0 & 1 \\
0 & 1 & 3
\end{bmatrix}
\]

4. \[
\begin{bmatrix}
-1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1
\end{bmatrix}
\]

5. \[
\begin{bmatrix}
1 & 1 & 1 \\
1 & 2 & 1 \\
1 & 1 & 3
\end{bmatrix}
\]

6. \[
\begin{bmatrix}
1 & 2 & 3 \\
2 & 4 & 6 \\
3 & 6 & 9
\end{bmatrix}
\]

Coloring and Whitening

- **Square root of covariance matrix:** Let \( \Sigma \) be a covariance matrix. Then there exists an \( n \times n \) matrix \( \Sigma^{1/2} \) such that \( \Sigma = \Sigma^{1/2}(\Sigma^{1/2})^T \). The matrix \( \Sigma^{1/2} \) is called the square root of \( \Sigma \).

- **Coloring:** Let \( \mathbf{X} \) be white RV, i.e., has zero mean and \( \Sigma_X = I \). Assume without loss of generality that \( a = 1 \).

  Let \( \Sigma \) be a covariance matrix, then the RV \( \mathbf{Y} = \Sigma^{1/2}\mathbf{X} \) has covariance matrix \( \Sigma \) (why?)

  Hence we can generate a RV with any prescribed covariance from a white RV.

- **Whitening:** Given a zero mean RV \( \mathbf{Y} \) with nonsingular covariance matrix \( \Sigma \), then the RV \( \mathbf{X} = \Sigma^{-1/2}\mathbf{Y} \) is white

  Hence, we can generate a white RV from any RV with nonsingular covariance matrix.

- Coloring and whitening have applications in simulations, detection, and estimation.
Finding Square Root of $\Sigma$

- For convenience, we assume throughout that $\Sigma$ is nonsingular.
- Since $\Sigma$ is symmetric, it has $n$ real eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and $n$ corresponding orthogonal eigenvectors $u_1, u_2, \ldots, u_n$.

  Further, since $\Sigma$ is positive definite, the eigenvalues are all positive.

- Thus, we have

$$\Sigma u_i = \lambda_i u_i, \quad \lambda_i > 0, \ i = 1, 2, \ldots, n$$

$$u_i^T u_j = 0 \quad \text{for every} \ i \neq j$$

Without loss of generality assume that the $u_i$ vectors are unit vectors.

- The first set of equations can be rewritten in the matrix form

$$\Sigma U = U \Lambda,$$

where

$$U = [u_1 \ u_2 \ \ldots \ u_n]$$

and $\Lambda$ is a diagonal matrix with diagonal elements $\lambda_i$.

- Note that $U$ is a unitary matrix ($U^T U = U U^T = I$), hence

$$\Sigma = U \Lambda U^T$$

and the square root of $\Sigma$ is

$$\Sigma^{1/2} = U \Lambda^{1/2},$$

where $\Lambda^{1/2}$ is a diagonal matrix with diagonal elements $\lambda_i^{1/2}$.

- The inverse of the square root is straightforward to find as

$$\Sigma^{-1/2} = \Lambda^{-1/2} U^T.$$

- Example: Let

$$\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$$

To find the eigenvalues of $\Sigma$, we find the roots of the polynomial equation

$$\det(\Sigma - \lambda I) = \lambda^2 - 5\lambda + 5 = 0,$$

which gives $\lambda_1 = 3.62$, $\lambda_2 = 1.38$.

To find the eigenvectors, consider

$$\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{12} \end{bmatrix} = 3.62 \begin{bmatrix} u_{11} \\ u_{12} \end{bmatrix},$$
and \( u_{11}^2 + u_{12}^2 = 1 \), which yields

\[
\mathbf{u}_1 = \begin{bmatrix} 0.53 \\ 0.85 \end{bmatrix}
\]

Similarly, we can find the second eigenvector

\[
\mathbf{u}_2 = \begin{bmatrix} -0.85 \\ 0.53 \end{bmatrix}
\]

Hence,

\[
\Sigma^{1/2} = \begin{bmatrix} 0.53 & -0.85 \\ 0.85 & 0.53 \end{bmatrix} \frac{\sqrt{3.62}}{\sqrt{1.38}} = \begin{bmatrix} 1 & -1 \\ 1.62 & 0.62 \end{bmatrix}
\]

The inverse of the square root is

\[
\Sigma^{-1/2} = \begin{bmatrix} 1/\sqrt{3.62} & 0 \\ 0 & 1/\sqrt{1.38} \end{bmatrix} \begin{bmatrix} 0.53 & 0.85 \\ -0.85 & 0.53 \end{bmatrix} = \begin{bmatrix} 0.28 & 0.45 \\ -0.72 & 0.45 \end{bmatrix}
\]

- **Geometric interpretation:** To generate a RV \( Y \) with covariance matrix \( \Sigma \) from white RV \( X \), we use the transformation \( Y = \mathbf{U} \Lambda^{1/2} X \).

  Equivalently, we first *scale* each component of \( X \) to obtain the RV \( Z = \Lambda^{1/2} X \).

  We then *rotate* \( Z \) using \( \mathbf{U} \) to obtain \( Y = \mathbf{U}Z \).

\[ \text{EE 278B: Random Vectors} \]

### Cholesky Decomposition

- **\( \Sigma \) has many square roots:**

  If \( \Sigma^{1/2} \) is a square root, then for any unitary matrix \( V \), \( \Sigma^{1/2}V \) is also a square root since \( \Sigma^{1/2}V V^T (\Sigma^{1/2})^T = \Sigma \)

- **The Cholesky decomposition is an efficient algorithm for computing lower triangle square root that can be used to perform coloring causally (sequentially):**

- **For \( n = 3 \), we want to find a lower triangle matrix (square root) \( A \) such that**

\[
\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{11} & a_{22} & a_{32} \\ 0 & 0 & a_{33} \end{bmatrix}
\]

The elements of \( A \) are computed in a *raster scan* manner:

\[
\begin{align*}
a_{11} & : \sigma_{11} = a_{11}^2 \Rightarrow a_{11} = \sqrt{\sigma_{11}} \\
a_{21} & : \sigma_{21} = a_{21}a_{11} \Rightarrow a_{21} = \sigma_{21}/a_{11} \\
a_{22} & : \sigma_{22} = a_{21}^2 + a_{22}^2 \Rightarrow a_{22} = \sqrt{\sigma_{22} - a_{21}^2} \\
a_{31} & : \sigma_{31} = a_{11}a_{31} \Rightarrow a_{31} = \sigma_{31}/a_{11}
\end{align*}
\]

\[ \text{EE 278B: Random Vectors} \]
• The inverse of a lower triangle square root is also lower triangular

• Coloring and whitening summary:
  ○ Coloring:
    \[
    \begin{array}{ccc}
    X & \xrightarrow{\Sigma^{1/2}} & Y \\
    \Sigma_X = I & & \Sigma_Y = \Sigma \\
    \end{array}
    \]

  ○ Whitening:
    \[
    \begin{array}{ccc}
    Y & \xrightarrow{\Sigma^{-1/2}} & X \\
    \Sigma_Y = \Sigma & & \Sigma_X = I \\
    \end{array}
    \]

  ○ Lower triangle square root and its inverse can be efficiently computed using Cholesky decomposition

Gaussian Random Vectors

• A random vector \( \mathbf{X} = (X_1, \ldots, X_n) \) is a Gaussian random vector (GRV) (or \(X_1, X_2, \ldots, X_n\) are jointly Gaussian r.v.s) if the joint pdf is of the form
  \[
  f_X(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} e^{-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)},
  \]
  where \( \mu \) is the mean and \( \Sigma \) is the covariance matrix of \( \mathbf{X} \), and \( |\Sigma| > 0 \), i.e., \( \Sigma \) is positive definite

• Verify that this joint pdf is the same as the case \( n = 2 \) from Lecture Notes 2

• Notation: \( \mathbf{X} \sim \mathcal{N}(\mu, \Sigma) \) denotes a GRV with given mean and covariance matrix

• Since \( \Sigma \) is positive definite, \( \Sigma^{-1} \) is positive definite. Thus if \( x - \mu \neq 0 \),
  \[
  (x - \mu)^T \Sigma^{-1} (x - \mu) > 0,
  \]
  which means that the contours of equal pdf are ellipsoids

• The GRV \( \mathbf{X} \sim \mathcal{N}(0, aI) \), where \( I \) is the identity matrix and \( a > 0 \), is called white; its contours of equal joint pdf are spheres centered at the origin
Properties of GRVs

• **Property 1**: For a GRV, uncorrelation implies independence
  This can be verified by substituting \( \sigma_{ij} = 0 \) for all \( i \neq j \) in the joint pdf.
  Then \( \Sigma \) becomes diagonal and so does \( \Sigma^{-1} \), and the joint pdf reduces to the product of the marginals \( X_i \sim \mathcal{N}(\mu_i, \sigma_{ii}) \).
  For the white GRV \( X \sim \mathcal{N}(0, aI) \), the r.v.s are i.i.d. \( \mathcal{N}(0, a) \).

• **Property 2**: Linear transformation of a GRV yields a GRV, i.e., given any \( m \times n \) matrix \( A \), where \( m \leq n \) and \( A \) has full rank \( m \), then
  \[ Y = AX \sim \mathcal{N}(A\mu, A\Sigma A^T) \]

• Example: Let
  \[ X \sim \mathcal{N} \left( 0, \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} \right) \]
  Find the joint pdf of
  \[ Y = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} X \]

Solution: From Property 2, we conclude that
  \[ Y \sim \mathcal{N} \left( 0, \begin{bmatrix} 7 & 3 \\ 3 & 2 \end{bmatrix} \right) = \mathcal{N} \left( 0, \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} \right) \]

Before we prove Property 2, let us show that
  \[ E(Y) = A\mu \quad \text{and} \quad \Sigma_Y = A\Sigma A^T \]
  These results follow from linearity of expectation. First, expectation:
  \[ E(Y) = E(AX) = A E(X) = A\mu \]
  Next consider the covariance matrix:
  \[ \Sigma_Y = E \left[ (Y - E(Y))(Y - E(Y))^T \right] \]
  \[ = E \left[ (AX - A\mu)(AX - A\mu)^T \right] \]
  \[ = A E \left[ (X - \mu)(X - \mu)^T \right] A^T = A\Sigma A^T \]

Of course this is not sufficient to show that \( Y \) is a GRV—we must also show that the joint pdf has the right form.
We do so using the *characteristic function* for a random vector.
• **Definition:** If \( X \sim f_X(x) \), the characteristic function of \( X \) is
\[
\Phi_X(\omega) = E\left(e^{i\omega^T x}\right),
\]
where \( \omega \) is an \( n \)-dimensional real valued vector and \( i = \sqrt{-1} \)

Thus
\[
\Phi_X(\omega) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(x)e^{i\omega^T x} dx
\]

This is the inverse of the multi-dimensional Fourier transform of \( f_X(x) \), which implies that there is a one-to-one correspondence between \( \Phi_X(\omega) \) and \( f_X(x) \).

The joint pdf can be found by taking the Fourier transform of \( \Phi_X(\omega) \), i.e.,
\[
f_X(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{(2\pi)^n} \Phi_X(\omega) e^{-i\omega^T x} d\omega
\]

• **Example:** The characteristic function for \( X \sim N(\mu, \sigma^2) \) is
\[
\Phi_X(\omega) = e^{-\frac{1}{2} \omega^T \sigma^2} + i\mu \omega,
\]
and for a GRV \( X \sim N(\mu, \Sigma) \),
\[
\Phi_X(\omega) = e^{-\frac{1}{2} \omega^T \Sigma \omega + i\omega^T \mu}
\]

• Now let’s go back to proving Property 2

Since \( A \) is an \( m \times n \) matrix, \( Y = AX \) and \( \omega \) are \( m \)-dimensional. Therefore the characteristic function of \( Y \) is
\[
\Phi_Y(\omega) = E\left(e^{i\omega^T Y}\right)
= E\left(e^{i\omega^T AX}\right)
= \Phi_X(A^T \omega)
= e^{-\frac{1}{2}(A^T \omega)^T \Sigma (A^T \omega) + i\omega^T A\mu}
\]

Thus \( Y = AX \sim N(A\mu, A\Sigma A^T) \)

• An equivalent definition of GRV: \( X \) is a GRV iff for any real vector \( a \neq 0 \), the r.v. \( Y = a^T X \) is Gaussian (see HW for proof)

• Whitening transforms a GRV to a white GRV; conversely, coloring transforms a white GRV to a GRV with prescribed covariance matrix
• **Property 3:** Marginals of a GRV are Gaussian, i.e., if \( X \) is GRV then for any subset \( \{i_1, i_2, \ldots, i_k\} \subset \{1, 2, \ldots, n\} \) of indexes, the RV

\[
Y = \begin{bmatrix}
X_{i_1} \\
X_{i_2} \\
\vdots \\
X_{i_k}
\end{bmatrix}
\]

is a GRV

• To show this we use Property 2. For example, let \( n = 3 \) and \( Y = \begin{bmatrix} X_1 \\ X_3 \end{bmatrix} \)

We can express \( Y \) as a linear transformation of \( X \):

\[
Y = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_3 \end{bmatrix}
\]

Therefore

\[
Y \sim \mathcal{N} \left( \begin{bmatrix} \mu_1 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{13} \\ \sigma_{31} & \sigma_{33} \end{bmatrix} \right)
\]

• As we have seen in Lecture Notes 2, the converse of Property 3 does not hold in general, i.e., Gaussian marginals do not necessarily mean that the r.v.s are jointly Gaussian


---

• **Property 4:** Conditionals of a GRV are Gaussian, more specifically, if

\[
X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right),
\]

where \( X_1 \) is a \( k \)-dim RV and \( X_2 \) is an \( n - k \)-dim RV, then

\[
X_2 \mid \{X_1 = x\} \sim \mathcal{N} \left( \Sigma_{21} \Sigma_{11}^{-1} (x - \mu_1) + \mu_2, \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \right)
\]

Compare this to the case of \( n = 2 \) and \( k = 1 \):

\[
X_2 \mid \{X_1 = x\} \sim \mathcal{N} \left( \frac{\sigma_{21}}{\sigma_{11}} (x - \mu_1) + \mu_2, \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}} \right)
\]

• Example:

\[
\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 1 \\ 2 & 5 & 2 \\ 1 & 2 & 9 \end{bmatrix} \right)
\]
From Property 4, it follows that

\[
E(X_2 \mid X_1 = x) = \begin{bmatrix} 2 \\ 1 \end{bmatrix} (x - 1) + \begin{bmatrix} 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 2x \\ x + 1 \end{bmatrix}
\]

\[
\Sigma_{X_2 \mid X_1 = x} = \begin{bmatrix} 5 & 2 \\ 2 & 9 \end{bmatrix} - \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \end{bmatrix}
\]

\[
= \begin{bmatrix} 1 & 0 \\ 0 & 8 \end{bmatrix}
\]

• The proof of Property 4 follows from properties 1 and 2 and the orthogonality principle (HW exercise)
Vector Detection

- Let the signal $\Theta = \theta_0$ with probability $p_0$ and $\Theta = \theta_1$ with probability $p_1 = 1 - p_0$
- We observe the RV $Y$, where $Y | \{\Theta = \theta_0\} \sim f_{Y|\Theta}(y|\theta_0)$ and $Y | \{\Theta = \theta_1\} \sim f_{Y|\Theta}(y|\theta_1)$
- We wish to find the estimate $\hat{\Theta}(Y)$ that minimizes the probability of detection error $P\{\hat{\Theta} \neq \Theta\}$
- The optimal estimate is obtained using the MAP decoder
  $$\hat{\Theta}(y) = \begin{cases} \theta_0 & \text{if } \frac{p_0 f_{Y|\Theta}(y|\theta_0)}{p_0 f_{Y|\Theta}(y|\theta_1)} > 1 \\ \theta_1 & \text{otherwise} \end{cases}$$
- When $p_0 = p_1 = 1/2$, the MAP decoder reduces to the ML decoder
  $$\hat{\Theta}(y) = \begin{cases} \theta_0 & \text{if } \frac{f_{Y|\Theta}(y|\theta_0)}{f_{Y|\Theta}(y|\theta_1)} > 1 \\ \theta_1 & \text{otherwise} \end{cases}$$
Consider a complete binary reconstruction tree of finite depth $k$.

- The root node is assigned a r.v. $X_1 \sim \text{Bern}(1/2)$ (the signal).
- Denote the two children of each non-leaf node $i$ as $l_i$ and $r_i$ (e.g., for $i = 1$, $l_1 = 2$ and $r_1 = 3$).
- A random variable is assigned to each non-root node as follows:
  \[ X_{l_i} = X_i \oplus Z_{l_i}, \]
  \[ X_{r_i} = X_i \oplus Z_{r_i}, \]
  where $Z_1, Z_2, \ldots$ are i.i.d. Bern($\epsilon$), $\epsilon \leq 1/2$, r.v.s independent of $X_1$.

That is, the r.v. assigned to a node is the output of a binary symmetric channel (BSC) whose input is the r.v. of its parent.

- Denote the set of leaf r.v.s that are descendants of node $i$ as $X_i$ (e.g., for $i = 1$, $X_1 = (X_4, X_5, X_6, X_7)$, and for $i = 4$, $X_4 = X_4$ in figure).
- We observe the leaf node r.v.s $X_1$ and wish to find the estimate $\hat{X}_1(X_1)$ that minimizes the probability of error $P_e = P\{\hat{X}_1 \neq X_1\}$.
- This problem is a simple example of the reconstruction on the tree problem, which arises in computational evolutionary biology (phylogenetic reconstruction), statistical physics, and theoretical computer science. A question of interest in these fields is under what condition on the channel noise can $X_1$ be reconstructed with $P_e < 1/2$ as the tree depth $k \to \infty$.

The reconstruction problem itself is an example of graphical models in which random variables dependencies are specified by a graph (STAT 375, CS 228).
• Since $X_1 \sim \text{Bern}(1/2)$, the optimal estimate is obtained using the ML decoder

$$
\hat{X}_1(X_1) = \begin{cases} 
0 & \text{if } \frac{p_{X_1|X_1}(x_1|0)}{p_{X_1|X_1}(x_1|1)} > 1 \\
1 & \text{otherwise}
\end{cases}
$$

• Because of the special structure of the observation r.v.s, the optimal estimate can be computed using a fast iterative message passing algorithm

• Define

$$
L_{i,0} = p_{X_i|X_i}(x_i|0) \\
L_{i,1} = p_{X_i|X_i}(x_i|1)
$$

• Then the ML estimate can be written as

$$
\hat{X}_1(X_1) = \begin{cases} 
0 & \text{if } L_{1,0} > L_{1,1} \\
1 & \text{otherwise}
\end{cases}
$$

• We now show that $L_{1,0}, L_{1,1}$ can be computed (in order of the number of nodes in the tree) by iteratively computing the intermediate likelihoods $L_{i,0}, L_{i,1}$ beginning with the leaf nodes for which $L_{i,0} = 1 - x_i$ $L_{i,1} = x_i$

• By the law of total probability, for a non-leaf node $i$, we can write

$$
L_{i,0} = p_{X_i|X_i}(x_i|0) \\
= p_{X_{i_1}|X_{i_1}}(x_{i_1}|0) p_{X_{i_2}|X_{i_2}}(x_{i_2}|0) p_{X_{i_3}|X_{i_3}}(x_{i_3}|0) \\
+ p_{X_{i_1}|X_{i_1}}(x_{i_1}|0) p_{X_{i_2}|X_{i_2}}(x_{i_2}|0) p_{X_{i_3}|X_{i_3}}(x_{i_3}|0) \\
+ p_{X_{i_1}|X_{i_1}}(x_{i_1}|0) p_{X_{i_2}|X_{i_2}}(x_{i_2}|1) p_{X_{i_3}|X_{i_3}}(x_{i_3}|0) \\
+ p_{X_{i_1}|X_{i_1}}(x_{i_1}|0) p_{X_{i_2}|X_{i_2}}(x_{i_2}|1) p_{X_{i_3}|X_{i_3}}(x_{i_3}|0) \\
+ p_{X_{i_1}|X_{i_1}}(x_{i_1}|1) p_{X_{i_2}|X_{i_2}}(x_{i_2}|0) p_{X_{i_3}|X_{i_3}}(x_{i_3}|0) \\
+ p_{X_{i_1}|X_{i_1}}(x_{i_1}|1) p_{X_{i_2}|X_{i_2}}(x_{i_2}|0) p_{X_{i_3}|X_{i_3}}(x_{i_3}|1) \text{ conditional independence}
$$

$$
= \hat{\epsilon}^2 \cdot p_{X_{i_1}|X_{i_1}}(x_{i_1}|0) p_{X_{i_2}|X_{i_2}}(x_{i_2}|0) p_{X_{i_3}|X_{i_3}}(x_{i_3}|0, 0, 0) \\
+ \hat{\epsilon} \cdot p_{X_{i_1}|X_{i_1}}(x_{i_1}|0) p_{X_{i_2}|X_{i_2}}(x_{i_2}|0) p_{X_{i_3}|X_{i_3}}(x_{i_3}|0, 0, 1) \\
+ \hat{\epsilon} \cdot p_{X_{i_1}|X_{i_1}}(x_{i_1}|0) p_{X_{i_2}|X_{i_2}}(x_{i_2}|0) p_{X_{i_3}|X_{i_3}}(x_{i_3}|1, 0, 0) \\
+ \hat{\epsilon} \cdot p_{X_{i_1}|X_{i_1}}(x_{i_1}|0) p_{X_{i_2}|X_{i_2}}(x_{i_2}|0) p_{X_{i_3}|X_{i_3}}(x_{i_3}|1, 0, 1) \\
+ \hat{\epsilon} \cdot p_{X_{i_1}|X_{i_1}}(x_{i_1}|0) p_{X_{i_2}|X_{i_2}}(x_{i_2}|0) p_{X_{i_3}|X_{i_3}}(x_{i_3}|1, 0, 1) \\
+ \hat{\epsilon} \cdot p_{X_{i_1}|X_{i_1}}(x_{i_1}|0) p_{X_{i_2}|X_{i_2}}(x_{i_2}|0) p_{X_{i_3}|X_{i_3}}(x_{i_3}|1, 0, 1)
$$

$L_{i,1}$ can be expressed similarly
• Now, since \( X_i = (X_{li}, X_{ri}) \), by conditional independence,
\[
p_{X_i|X_i, X_{li}, X_{ri}}(x_i, x_{li}, x_{ri}) = p_{X_{li}|X_i, X_{ri}}(x_{li}|x_i, x_{ri}) p_{X_{ri}|X_i, X_{li}}(x_{ri}|x_i, x_{li})
\]

• Hence we obtain the following iteratively equations
\[
L_{i,0} = (\bar{\epsilon}L_{li,0} + \epsilon L_{li,1})(\bar{\epsilon}L_{ri,0} + \epsilon L_{ri,1}),
\]
\[
L_{i,1} = (\epsilon L_{li,0} + \bar{\epsilon}L_{li,1})(\epsilon L_{ri,0} + \bar{\epsilon}L_{ri,1}),
\]
where, at the leaf nodes
\[
L_{i,0} = p_{X_i|X_i}(x_i|0) = 1 - x_i
\]
\[
L_{i,1} = p_{X_i|X_i}(x_i|1) = x_i
\]

• Hence to compute \( L_{1,0} \) and \( L_{1,1} \), we start with the likelihoods at each leaf node, then compute the likelihoods for the nodes at level \( k - 1 \), and so on until we arrive at node 1

Detection for Vector Additive Gaussian Noise Channel

• Consider the vector additive Gaussian noise (AGN) channel
\[
Y = \Theta + Z,
\]
where the signal \( \Theta = \theta_0 \), an \( n \)-dimensional real vector, with probability \( 1/2 \) and \( \Theta = \theta_1 \) with probability \( 1/2 \), and the noise \( Z \sim \mathcal{N}(0, \Sigma_Z) \) are independent

• We observe \( y \) and wish to find the estimate \( \hat{\Theta}(Y) \) that minimizes the probability of decoding error \( P\{\hat{\Theta} \neq \Theta\} \)

• First assume that \( \Sigma_Z = NI \), i.e., additive white Gaussian noise channel

• The optimal decoding rule is the ML decoder. Define the log likelihood ratio
\[
\Lambda(y) = \ln \frac{f(y|\theta_0)}{f(y|\theta_1)}
\]

Then, the ML decoder is
\[
\hat{\Theta}(y) = \begin{cases} 
\theta_0 & \text{if } \Lambda(y) > 0 \\
\theta_1 & \text{otherwise}
\end{cases}
\]
Now,
\[ \Lambda(y) = \frac{1}{2N} [(y - \theta_1)^T(y - \theta_1) - (y - \theta_0)^T(y - \theta_0)] \]

- Hence, the ML decoder reduces to the minimum distance decoder
  \[ \hat{\Theta}(y) = \begin{cases} 
  \theta_0 & \text{if } ||y - \theta_0|| < ||y - \theta_1|| \\
  \theta_1 & \text{otherwise}
  \end{cases} \]

- We can simplify this further to
  \[ \hat{\Theta}(y) = \begin{cases} 
  \theta_0 & \text{if } y^T(\theta_1 - \theta_0) < \frac{1}{2}(\theta_1^T\theta_1 - \theta_0^T\theta_0) \\
  \theta_1 & \text{otherwise}
  \end{cases} \]

Hence, the decision depends only on the value of a scalar r.v. 
\( W = Y^T(\theta_1 - \theta_0) \). Such r.v. is referred to as a sufficient statistic for the optimal decoder. Further,
\[ W|\{\Theta = \theta_0\} \sim \mathcal{N}(\theta_0^T(\theta_1 - \theta_0), N(\theta_1 - \theta_0)^T(\theta_1 - \theta_0)) \]
\[ W|\{\Theta = \theta_1\} \sim \mathcal{N}(\theta_1^T(\theta_1 - \theta_0), N(\theta_1 - \theta_0)^T(\theta_1 - \theta_0)) \]

- Assuming that the signals have the same power, i.e., \( \theta_0^T\theta_0 = \theta_1^T\theta_1 = P \), the optimal decoding rule reduces to the matched filter decoder (receiver)
  \[ \hat{\Theta}(y) = \begin{cases} 
  \theta_0 & \text{if } y^T(\theta_1 - \theta_0) < 0 \\
  \theta_1 & \text{otherwise}
  \end{cases} \]
  that is,
  \[ \hat{\Theta}(y) = \begin{cases} 
  \theta_0 & \text{if } w < 0 \\
  \theta_1 & \text{if } w \geq 0
  \end{cases} \]

This is the same as the optimal rule for the scalar case discussed in Lecture notes 1! The minimum probability of error is
\[ P_e = Q \left( \frac{P - \theta_0^T\theta_1}{\sqrt{2N(P - \theta_0^T\theta_1)}} \right) \]
\[ = Q \left( \frac{P - \theta_0^T\theta_1}{2N} \right) \]
This is minimized by using antipodal signals $\theta_0 = -\theta_1$, which yields

$$P_e = Q\left(\sqrt{\frac{P}{N}}\right)$$

Exactly the same as scalar antipodal signals

- Now suppose that the noise is not white, i.e., $\Sigma_Z \neq NI$. Then the ML decoder reduces to

$$\hat{\Theta}(y) = \begin{cases} 
\theta_0 & \text{if } (y - \theta_0)^T \Sigma_Z^{-1} (y - \theta_0) < (y - \theta_1)^T \Sigma_Z^{-1} (y - \theta_1) \\
\theta_1 & \text{otherwise}
\end{cases}$$

Now, let $y' = \Sigma_Z^{-1/2} y$ and $\theta'_i = \Sigma_Z^{-1/2} \theta_i$ for $i = 0, 1$, then the rule becomes the same as that for the white noise case

$$\hat{\Theta}(y) = \begin{cases} 
\theta_0 & \text{if } ||y' - \theta'_0|| < ||y' - \theta'_1|| \\
\theta_1 & \text{otherwise}
\end{cases}$$

and can be simplified to the scalar case as before

- Thus, the optimal decoder is to first multiply $Y$ by $\Sigma_Z^{-1/2}$ to obtain $Y'$ and then to apply the optimal rule for the white noise case with the transformed signals $\theta'_i = \Sigma_Z^{-1/2} \theta_i$, $i = 0, 1$

**Vector Linear Estimation**

- Let $X \sim f_X(x)$ be a r.v. representing the signal and let $Y$ be an $n$-dimensional RV representing the observations

- The minimum MSE estimate of $X$ given $Y$ is the conditional expectation $E(X \mid Y)$. This is often not practical to compute either because the conditional pdf of $X$ given $Y$ is not known or because of high computational cost

- The MMSE linear (or affine) estimate is easier to find since it depends only on the means, variances, and covariances of the r.v.s involved

- To find the MMSE linear estimate, first assume that $E(X) = 0$ and $E(Y) = 0$. The problem reduces to finding a real $n$-vector $h$ such that

$$\hat{X} = h^T Y = \sum_{i=1}^n h_i Y_i$$

minimizes the MSE $E[(X - \hat{X})^2]$
MMSE Linear Estimate via Orthogonality Principle

• To find $\hat{X}$ we use the orthogonality principle: we view the r.v.s $X, Y_1, Y_2, \ldots, Y_n$ as vectors in the inner product space consisting of all zero mean r.v.s defined over the underlying probability space.

• The linear estimation problem reduces to a geometry problem: find the vector $\hat{X}$ that is closest to $X$ (in norm of error $X - \hat{X}$).

\[ X \text{ signal} \]
\[ \text{error vector } X - \hat{X} \]
\[ \hat{X} \]
\[ \text{subspace spanned by } Y_1, Y_2, \ldots, Y_n \]

• To minimize $\text{MSE} = \|X - \hat{X}\|^2$, we choose $\hat{X}$ so that the error vector $X - \hat{X}$ is orthogonal to the subspace spanned by the observations $Y_1, Y_2, \ldots, Y_n$, i.e.,

\[
E[(X - \hat{X})Y_i] = 0, \quad i = 1, 2, \ldots, n,
\]

hence

\[
E(Y_iX) = E(Y_i\hat{X}) = \sum_{j=1}^{n} h_j E(Y_jY_i), \quad i = 1, 2, \ldots, n
\]

• Define the cross covariance of $Y$ and $X$ as the $n$-vector

\[
\Sigma_{YX} = E[(Y - E(Y))(X - E(X))] = \begin{bmatrix}
\sigma_{Y_1X} \\
\sigma_{Y_2X} \\
\vdots \\
\sigma_{Y_nX}
\end{bmatrix}
\]

For $n = 1$ this is simply the covariance.

• The above equations can be written in vector form as $\Sigma_Y h = \Sigma_{YX}$.

• If $\Sigma_Y$ is nonsingular, we can solve the equations to obtain $h = \Sigma_Y^{-1}\Sigma_{YX}$. 

EE 278B: Random Vectors
• Thus, if $\Sigma_Y$ is nonsingular then the best linear MSE estimate is:

$$\hat{X} = h^T Y = \Sigma_{YX}^T \Sigma_Y^{-1} Y$$

• Compare this to the scalar case, where $\hat{X} = \frac{\text{Cov}(X,Y)}{\sigma_Y^2} Y$

• Now to find the minimum MSE, consider

$$\text{MSE} = E[(X - \hat{X})^2]$$

$$= E[(X - \hat{X})X] - E[(X - \hat{X})\hat{X}]$$

$$= E[(X - \hat{X})X]$$, since by orthogonality $(X - \hat{X}) \perp \hat{X}$

$$= E(X^2) - E(\hat{X}X)$$

$$= \text{Var}(X) - E(\Sigma_{YX}^T \Sigma_Y^{-1} Y X) = \text{Var}(X) - \Sigma_{YX}^T \Sigma_Y^{-1} \Sigma_Y X$$

• Compare this to the scalar case, where minimum MSE is $\text{Var}(X) - \frac{\text{Cov}(X,Y)^2}{\sigma_Y^2}$

• If $X$ or $Y$ have nonzero mean, the MMSE affine estimate $\hat{X} = h_0 + h^T Y$ is determined by first finding the MMSE linear estimate of $X - E(X)$ given $Y - E(Y)$ (minimum MSE for $\hat{X}'$ and $\hat{X}$ are the same), which is

$$\hat{X}' = \Sigma_{YX}^T \Sigma_Y^{-1} (Y - E(Y))$$, and then setting $\hat{X} = \hat{X}' + E(X)$ (since $E(\hat{X}) = E(X)$ is necessary)

Example

• Let $X$ be the r.v. representing a signal with mean $\mu$ and variance $P$. The observations are $Y_i = X + Z_i$, for $i = 1, 2, \ldots, n$, where the $Z_i$ are zero mean uncorrelated noise with variance $N$, and $X$ and $Z_i$ are also uncorrelated

Find the MMSE linear estimate of $X$ given $Y$ and its MSE

• For $n = 1$, we already know that $\hat{X}_1 = \frac{P}{P + N} Y_1 + \frac{N}{P + N} \mu$

• To find the MMSE linear estimate for general $n$, first let $X' = X - \mu$ and $Y'_i = Y_i - \mu$. Thus $X'$ and $Y'$ are zero mean

• The MMSE linear estimate of $X'$ given $Y'$ is given by $\hat{X}'_n = h^T Y'$, where

$$\Sigma_Y h = \Sigma_{YX}$$, thus

$$\begin{bmatrix} P + N & P & \cdots & P \\ P & P + N & \cdots & P \\ \vdots & \vdots & \ddots & \vdots \\ P & P & \cdots & P + N \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix} = \begin{bmatrix} P \\ P \\ \vdots \\ P \end{bmatrix}$$
• By symmetry, \( h_1 = h_2 = \cdots = h_n = \frac{P}{nP + N} \). Thus

\[
\hat{X}'_n = \frac{P}{nP + N} \sum_{i=1}^{n} Y'_i
\]

Therefore

\[
\hat{X}_n = \frac{P}{nP + N} \left( \sum_{i=1}^{n} (Y_i - \mu) \right) + \mu = \frac{P}{nP + N} \left( \sum_{i=1}^{n} Y_i \right) + \frac{N}{nP + N} \mu
\]

• The mean square error of the estimate:

\[
\text{MSE}_n = P - \mathbf{E}(\hat{X}'X') = \frac{PN}{nP + N}
\]

Thus as \( n \to \infty \), \( \text{MSE}_n \to 0 \), i.e., the linear estimate becomes perfect (even though we don’t know the complete statistics of \( X \) and \( Y \)).

---

**Linear Innovation Sequence**

• Let \( X \) be the signal and \( Y \) be the observation vector (all zero mean)

• Suppose the \( Y_i \)'s are orthogonal, i.e., \( \mathbf{E}(Y_i Y_j) = 0 \) for all \( i \neq j \), and let \( \hat{X}(Y) \) be the best linear MSE estimate of \( X \) given \( Y \) and \( \hat{X}(Y_i) \) be the best linear MSE estimate of \( X \) given only \( Y_i \) for \( i = 1, \ldots, n \), then we can write

\[
\hat{X}(Y) = \sum_{i=1}^{n} \hat{X}(Y_i),
\]

\[
\text{MSE} = \text{Var}(X) - \sum_{i=1}^{n} \frac{\text{Cov}^2(X, Y_i)}{\text{Var}(Y_i)}
\]

Hence the computation of the best linear MSE estimate and its MSE are very simple

In fact, we can compute the estimates and the MSE *causally* (recursively)

\[
\hat{X}(Y^{i+1}) = \hat{X}(Y^i) + \hat{X}(Y_{i+1})
\]

\[
\text{MSE}_{i+1} = \text{MSE}_i - \frac{\text{Cov}^2(X, Y_{i+1})}{\text{Var}(Y_{i+1})}
\]
• This can be proved by direct evaluation of MMSE linear estimate or using orthogonality:

\[
\hat{X}(Y_1) = \sum_{i=1}^{n} \hat{X}(Z_i)
\]

\[
\text{MSE} = \text{Var}(X) - \sum_{i=1}^{n} \text{Cov}^2(X, Z_i)
\]

• Now suppose the \(Y_i\)s are not orthogonal. We can still express the estimate and its MSE as sums
  - We first whiten \(Y\) to obtain \(Z\). The best linear MSE estimate of \(X\) given \(Y\) is exactly the same as that given \(Z\) (why?)
  - The estimate and its MSE can then be computed as
    \[
    \hat{X}(Y) = \sum_{i=1}^{n} \hat{X}(Z_i)
    \]
    \[
    \text{MSE} = \text{Var}(X) - \sum_{i=1}^{n} \text{Cov}^2(X, Z_i)
    \]
  - We can compute an orthogonal observation sequence \(\tilde{Y}\) from \(Y\) causally:
    - Given \(Y^i\), we compute the error of the best linear MSE estimate of \(Y_{i+1}\),
      \[
      \tilde{Y}_{i+1}(Y^i) = Y_{i+1} - \hat{Y}_{i+1}(Y^i)
      \]
    - Clearly, \(\tilde{Y}_{i+1} \perp (\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_i)\), hence we can write
      \[
      \hat{Y}_{i+1}(Y^i) = \sum_{j=1}^{i} \hat{Y}_{i+1}(\tilde{Y}_j)
      \]
Interpretation: $\hat{Y}_{i+1}$ is the part of $Y_{i+1}$ predictable by $Y^i$, hence carries no useful new information for estimating $X$ beyond $Y^i$. 

$\tilde{Y}_{i+1}$ by comparison is the unpredictable part, hence carries new information.

As such, $\tilde{Y}$ is called the linear innovation sequence of $Y$.

Remark: If we normalize $\tilde{Y}$ (by dividing each $\tilde{Y}_i$ by its standard deviation), we obtain the same sequence as using the Cholesky decomposition in Lecture notes 3.

Example: Let the observation sequence be $Y_i = X + Z_i$ for $i = 1, 2, \ldots, n$, where $X, Z_1, \ldots, Z_n$ are zero mean, uncorrelated r.v.s with $E(X^2) = P$ and $E(Z_i^2) = N$ for $i = 1, 2, \ldots, n$. Find the linear innovation sequence of $Y$.

Using the innovation sequence, the MMSE linear estimate of $X$ given $\tilde{Y}^{i+1}$ and its MSE can be computed causally:

$$\hat{X}(\tilde{Y}^{i+1}) = \hat{X}(\tilde{Y}^i) + \hat{X}(\tilde{Y}_{i+1}),$$

$$\text{MSE}_{i+1} = \text{MSE}_i - \frac{\text{Cov}^2(X, \tilde{Y}_{i+1})}{\text{Var}(\tilde{Y}_{i+1})}$$

The innovation sequence will prove useful in deriving the Kalman filter.

**Kalman Filter**

- The Kalman filter is an efficient, recursive algorithm for computing the MMSE linear estimate and its MSE when the signal $X$ and observations $Y$ evolve according to a state-space model.

- Consider a linear dynamical system described by the state-space model:

$$X_{i+1} = A_i X_i + U_i, \quad i = 0, 1, \ldots, n$$

with noisy observations (output)

$$Y_i = X_i + V_i, \quad i = 0, 1, \ldots, n,$$

where $X_0, U_0, U_1, \ldots, U_n, V_0, V_1, \ldots, V_n$ are zero mean, uncorrelated RVs with $\Sigma_{X_0} = P_0, \Sigma_{U_i} = Q_i, \Sigma_{V_i} = N_i; A_i$ is a known sequence of matrices.
• This state space model is used in many applications:
  ○ Navigation, e.g., of a car:
    State: location, speed, heading, acceleration, tilt, steering wheel position of vehicle
    Observations: inertial (accelerometer, gyroscopes), electronic compass, GPS
  ○ Phase locked loop:
    State: phase and frequency offsets
    Observations: noisy observation of phase
  ○ Computer vision, e.g., face tracking:
    State: pose, motion, shape (size, articulation), appearance (light, color)
    Observations: video frame sequence

• The goal is to compute the MMSE linear estimate of the state from causal observations:
  ○ Prediction: Find the estimate \( \hat{X}_{i+1|i} \) of \( X_{i+1} \) from \( Y^i \) and its MSE \( \Sigma_{i+1|i} \)
  ○ Filtering: Find the estimate \( \hat{X}_{i|\hat{X}} \) of \( X_i \) from \( Y^i \) and its MSE \( \Sigma_{i|\hat{X}} \)

• The Kalman filter provides clever recursive equations for computing these estimates and their error covariance matrices
Scalar Kalman Filter

- Consider the scalar state space system:
  \[ X_{i+1} = a_i X_i + U_i, \quad i = 0, 1, \ldots, n \]
  with noisy observations
  \[ Y_i = X_i + V_i, \quad i = 0, 1, \ldots, n, \]
  where \( X_0, U_0, U_1, \ldots, U_n, V_0, V_1, \ldots, V_n \) are zero mean, uncorrelated r.v.s with \( \text{Var}(X_0) = P_0, \text{Var}(U_i) = Q_i, \text{Var}(V_i) = N_i \), and \( a_i \) is a known sequence.

- Kalman filter (prediction):
  Initialization: \( \hat{X}_{0|-1} = 0, \sigma_{0|-1}^2 = P_0 \)
  Update equations: For \( i = 0, 1, 2, \ldots, n \), the estimate is
  \[ \hat{X}_{i+1|i} = a_i \hat{X}_{i|i-1} + k_i (Y_i - \hat{X}_{i|i-1}), \]
  where the filter gain is
  \[ k_i = \frac{a_i \sigma_{i|i-1}^2}{\sigma_{i|i-1}^2 + N_i} \]
  The MSE of \( \hat{X}_{i+1|i} \) is
  \[ \sigma_{i+1|i}^2 = a_i (a_i - k_i) \sigma_{i|i-1}^2 + Q_i \]
• Example: Let \( a_i = 1, Q_i = 0, N_i = N, \) and \( P_0 = P \) (so \( X_0 = X_1 = X_2 = \cdots = X \)), and \( Y_i = X + V_i \) (this is the same as the earlier estimation example)

Kalman filter:
Initialization: \( \hat{X}_{0|−1} = 0 \) and \( \sigma^2_{0|−1} = P \)

The update in each step is
\[
\hat{X}_{i+1|i} = (1 - k_i) \hat{X}_{i|i−1} + k_i Y_i
\]
with
\[
k_i = \frac{\sigma^2_{i|i−1}}{\sigma^2_{i|i−1} + N},
\]
and the MSE is
\[
\sigma^2_{i+1|i} = (1 - k_i)\sigma^2_{i|i−1}
\]

We can solve for \( \sigma^2_{i+1|i} \) explicitly
\[
\sigma^2_{i+1|i} = \left(1 - \frac{\sigma^2_{i|i−1}}{\sigma^2_{i|i−1} + N}\right)\sigma^2_{i|i−1} = \frac{N\sigma^2_{i|i−1}}{\sigma^2_{i|i−1} + N}
\]
\[
\frac{1}{\sigma^2_{i+1|i}} = \frac{1}{N} + \frac{1}{\sigma^2_{i|i−1}}
\]
\[
\sigma^2_{i+1|i} = \frac{1}{i/N + 1/P} = \frac{NP}{iP + N}
\]

The gain is
\[
k_i = \frac{P}{iP + N}
\]

The recursive estimate is
\[
\hat{X}_{i+1|i} = \frac{(i - 1)P + N}{iP + N} \hat{X}_{i|i−1} + \frac{P}{iP + N} Y_i
\]
We thus obtain the previous result in a recursive form
Example: Let $n = 200$, $P_0 = 1$, $N_i = 1$

For $i = 1$ to 100: $a_i = \alpha^2$, $Q_i = P_0(1 - \alpha^2)$ with $\alpha = 0.95$ (memory factor)

For $i = 100$ to 200: $a_i = 1$, $Q_i = 0$ (i.e., state remains constant)
Derivation of the Kalman Filter

- We use innovations. Let \( \tilde{Y}_i \) be the innovation r.v. for \( Y_i \), then we can write

\[
\begin{align*}
\hat{X}_{i+1|i} &= \hat{X}_{i+1|i-1} + k_i \tilde{Y}_i, \\
\sigma_{i+1|i} &= \sigma_{i+1|i-1} + k_i \text{Cov}(X_{i+1}, \tilde{Y}_i)
\end{align*}
\]

where \( \hat{X}_{i+1|i-1} \) and \( \sigma_{i+1|i-1} \) are the MMSE linear estimate of \( X \) given \( Y_{i-1} \) and its MSE, and

\[
k_i = \frac{\text{Cov}(X_{i+1}, \tilde{Y}_i)}{\text{Var}(\tilde{Y}_i)}
\]

- Now, since \( X_{i+1} = a_i X_i + U_i \), by linearity of MMSE linear estimate, we have

\[
\hat{X}_{i+1|i-1} = a_i \hat{X}_{i|i-1}
\]

and

\[
\sigma_{i+1|i-1}^2 = a_i^2 \sigma_{i|i-1}^2 + Q_i
\]

- Now, the innovation r.v. for \( Y_i \) is \( \tilde{Y}_i = Y_i - \hat{Y}_i(Y_{i-1}) \)

Since \( Y_i = X_i + V_i \) and \( V_i \) is uncorrelated with \( Y_j, j = 1, 2, \ldots, i-1 \),

\[
\hat{Y}_i(Y_{i-1}) = \hat{X}_{i|i-1}
\]

Hence,

\[
\tilde{Y}_i = Y_i - \hat{X}_{i|i-1}
\]

This yields

\[
\begin{align*}
\hat{X}_{i+1|i} &= a_i \hat{X}_{i|i-1} + k_i \tilde{Y}_i = a_i \hat{X}_{i|i-1} + k_i (Y_i - \hat{X}_{i|i-1}) \\
\sigma_{i+1|i}^2 &= \sigma_{i+1|i-1}^2 - k_i \text{Cov}(X_{i+1}, \tilde{Y}_i),
\end{align*}
\]

Now, consider

\[
k_i = \frac{\text{Cov}(X_{i+1}, \tilde{Y}_i)}{\text{Var}(\tilde{Y}_i)} = \frac{\text{Cov}(a_i X_i + U_i, X_i - \hat{X}_{i|i-1} + V_i)}{\text{Var}(X_i - \hat{X}_{i|i-1} + V_i)} = \frac{\text{Cov}(a_i X_i, X_i - \hat{X}_{i|i-1})}{\text{Var}(X_i - \hat{X}_{i|i-1} + V_i)}
\]
= \frac{a_i \text{Cov}(X_i, X_i - \hat{X}_{i|i-1})}{\text{Var}(X_i - \hat{X}_{i|i-1} + V_i)}
\begin{align*}
&= \frac{a_i \text{Cov}(X_i - \hat{X}_{i|i-1}, X_i - \hat{X}_{i|i-1})}{\text{Var}(X_i - \hat{X}_{i|i-1} + V_i)} \\
&= \frac{a_i \text{Var}(X_i - \hat{X}_{i|i-1})}{\text{Var}(X_i - \hat{X}_{i|i-1} + N_i)} \\
&= \frac{a_i \sigma_i^2}{\sigma_{i|i-1}^2 + N_i}
\end{align*}

The MSE is
\begin{align*}
\sigma_{i+1|i}^2 &= \sigma_{i+1|i-1}^2 - k_i \text{Cov}(a_i X_i + U_i, X_i - \hat{X}_{i|i-1} + V_i) \\
&= \sigma_{i+1|i-1}^2 - k_i a_i \sigma_{i|i-1}^2 \\
&= a_i (a_i - k_i) \sigma_{i|i-1}^2 + Q_i
\end{align*}

This completes the derivation of the scalar Kalman filter.

---

**Vector Kalman Filter**

- The above scalar Kalman filter can be extended to the vector state space model:
  - **Initialization:** \( \hat{X}_{0|-1} = 0, \Sigma_{0|-1} = P_0 \)
  - **Update equations:** For \( i = 0, 1, 2, \ldots, n \), the estimate is
    \( \hat{X}_{i+1|i} = A_i \hat{X}_{i|i-1} + K_i (Y_i - \hat{X}_{i|i-1}) \),
  - where the filter gain matrix
    \( K_i = A_i \Sigma_{i|i-1} (\Sigma_{i|i-1} + N_i)^{-1} \)
  - The covariance of the error is
    \( \Sigma_{i+1|i} = A_i \Sigma_{i|i-1} A_i^T - K_i \Sigma_{i|i-1} A_i^T + Q_i \)
  - **Remark:** If \( X_0, U_0, U_1, \ldots, U_n \) and \( V_0, V_1, \ldots, V_n \) are Gaussian (zero mean, uncorrelated), then the Kalman filter yields the best MSE estimate of \( X_i \), \( i = 0, \ldots, n \)
Filtering

• Now assume the goal is to compute the MMSE linear estimate of $X_i$ given $Y_i$, i.e., instead of predicting the next state, we are interested in estimating the current state

• We denote this estimate by $\hat{X}_{i|i}$ and its MSE by $\sigma^2_{i|i}$

• The Kalman filter can be adapted to this case as follows:
  Initialization:
  \[
  \hat{X}_{0|0} = \frac{P_0}{P_0 + N_0} Y_0 \\
  \sigma^2_{0|0} = \frac{P_0 N_0}{P_0 + N_0}
  \]
  Update equations: For $i = 1, 2, \ldots, n$, the estimate is
  \[
  \hat{X}_{i|i} = a_{i-1}(1 - k_i)\hat{X}_{i-1|i-1} + k_i Y_i
  \]
  with filter gain
  \[
  k_i = \frac{a_{i-1}^2 \sigma^2_{i-1|i-1} + Q_{i-1}}{a_{i-1}^2 \sigma^2_{i-1|i-1} + Q_{i-1} + N_i}
  \]
  and MSE recursion
  \[
  \sigma^2_{i|i} = (1 - k_i) \left( a_{i-1}^2 \sigma^2_{i-1|i-1} + Q_{i-1} \right)
  \]

• Vector case
  Initialization:
  \[
  \hat{X}_{0|0} = P_0(P_0 + N_0)^{-1} Y_0 \\
  \Sigma_{0|0} = P_0(I - (P_0 + N_0)^{-1}P_0)
  \]
  Update equations: For $i = 1, 2, \ldots, n$, the estimate is
  \[
  \hat{X}_{i|i} = (I - K_i)A_{i-1}\hat{X}_{i-1|i-1} + K_i Y_i
  \]
  with filter gain
  \[
  K_i = (A_{i-1} \Sigma_{i-1|i-1} A_{i-1}^T + Q_{i-1})^{-1} (A_{i-1} \Sigma_{i-1|i-1} A_{i-1}^T + Q_{i-1} + N_i)^{-1}
  \]
  and MSE recursion
  \[
  \Sigma_{i|i} = (A_{i-1} \Sigma_{i-1|i-1} A_{i-1}^T + Q_{i-1})(I - K_i^T)
  \]
Motivation

- One of the key questions in statistical signal processing is how to estimate the *statistics* of a r.v., e.g., its mean, variance, distribution, etc.

  To estimate such a statistic, we collect *samples* and use an *estimator* in the form of a *sample average*

  - How good is the *estimator*? Does it *converge* to the true statistic?
  - How many samples do we need to ensure with some *confidence* that we are within a certain range of the true value of the statistic?

- Another key question in statistical signal processing is how to estimate a signal from noisy observations, e.g., using MSE or linear MSE

  - Does the estimator converge to the true signal?
  - How many observations do we need to achieve a desired estimation accuracy?

- The subject of convergence and limit theorems for r.v.s addresses such questions
Example: Estimating the Mean of a R.V.

- Let $X$ be a r.v. with finite but unknown mean $E(X)$
- To estimate the mean we generate $X_1, X_2, \ldots, X_n$ i.i.d. samples drawn according to the same distribution as $X$ and compute the sample mean
  \[ S_n = \frac{1}{n} \sum_{i=1}^{n} X_i \]
- Does $S_n$ converge to $E(X)$ as we increase $n$? If so, how fast?
  But what does it mean to say that a r.v. sequence $S_n$ converges to $E(X)$?
- First we give an example: Let $X_1, X_2, \ldots, X_n, \ldots$ be i.i.d. $\mathcal{N}(0,1)$
  - We use MATLAB to generate 6 sets of outcomes of $X_1, \ldots, X_n, \ldots, X_{10000}$
  - We then plot $s_n$ for the 6 sets of outcomes as a function of $n$
  - Note that each $s_n$ sequence appears to be converging to 0, the mean of the r.v., as $n$ increases
Convergence With Probability 1

- Recall that a sequence of numbers \( x_1, x_2, \ldots, x_n, \ldots \) converges to \( x \) if for every \( \epsilon > 0 \), there exists an \( m(\epsilon) \) such that \( |x_n - x| < \epsilon \) for every \( n \geq m(\epsilon) \)

- Now consider a sequence of r.v.s \( X_1, X_2, \ldots, X_n, \ldots \) all defined on the same probability space \( \Omega \). For every \( \omega \in \Omega \) we obtain a sample sequence (sequence of numbers) \( X_1(\omega), X_2(\omega), \ldots, X_n(\omega), \ldots \)

- A sequence \( X_1, X_2, X_3, \ldots \) of r.v.s is said to converge to a random variable \( X \) with probability 1 (w.p.1, also called almost surely) if
  \[
P\{\omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)\} = 1
  \]

- This means that the set of sample paths that converge to \( X(\omega) \), in the sense of a sequence converging to a limit, has probability 1

- Equivalently, \( X_1, X_2, \ldots, X_n, \ldots \) converges w.p.1 if for every \( \epsilon > 0 \),
  \[
  \lim_{m \to \infty} P\{|X_n - X| < \epsilon \text{ for every } n \geq m\} = 1
  \]

Example 1: Let \( X_1, X_2, \ldots, X_n \) be i.i.d. Bern(1/2), and define \( Y_n = 2^n \prod_{i=1}^{n} X_i \). Show that the sequence \( Y_n \) converges to 0 w.p.1

Solution: To show this, let \( \epsilon > 0 \) (and \( \epsilon < 2^m \)), and consider
  \[
P\{|Y_n - 0| < \epsilon \text{ for all } n \geq m\} = P\{X_n = 0 \text{ for some } n \leq m\}
  = 1 - P\{X_n = 1 \text{ for all } n \leq m\}
  = 1 - \left(\frac{1}{2}\right)^m \to 1 \text{ as } m \to \infty
  \]

- An important example of convergence w.p.1: the Strong Law of Large Numbers (SLLN), which says that if \( X_1, X_2, \ldots, X_n, \ldots \) are i.i.d. with finite mean \( E(X) \), then the sequence of sample means \( S_n \to E(X) \) w.p.1
  - The previous MATLAB example is a good demonstration of the SLLN — each of the 6 sample paths appears to be converging to 0, which is \( E(X) \)
  - The proof of the SLLN and other convergence w.p.1 results are beyond the scope of this course. Take Stats 310 if you want to learn a lot more about this
Convergence in Mean Square

- A sequence of r.v.s $X_1, X_2, \ldots, X_n, \ldots$ converges to a random variable $X$ in mean square (m.s.) if
  \[
  \lim_{n \to \infty} E[(X_n - X)^2] = 0
  \]

- Example: Estimating the mean.
  Let $X_1, X_2, \ldots, X_n, \ldots$ be i.i.d. with finite mean $E(X)$ and variance $\text{Var}(X)$. Then $S_n \to E(X)$ in m.s.

- Proof: Here we need to show that
  \[
  \lim_{n \to \infty} E[(S_n - E(X))^2] = 0
  \]
  First note that
  \[
  E(S_n) = E \left[ \frac{1}{n} \sum_{i=1}^{n} X_i \right] = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{1}{n} \sum_{i=1}^{n} E(X) = E(X)
  \]
  So, $S_n$ is an unbiased estimate of $E(X)$

Now to prove convergence in m.s., consider
\[
E[(S_n - E(X))^2] = E[(S_n - E(S_n))^2]
\]
\[
= E \left[ \left( \frac{1}{n} \sum_{i=1}^{n} X_i - \frac{1}{n} \sum_{i=1}^{n} E(X) \right)^2 \right]
\]
\[
= \frac{1}{n^2} E \left[ \left( \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} E(X) \right)^2 \right]
\]
\[
= \frac{1}{n^2} \text{Var} \left( \sum_{i=1}^{n} X_i \right)
\]
\[
= \frac{1}{n^2} \left( \sum_{i=1}^{n} \text{Var}(X_i) \right) \quad \text{since } \{X_i\} \text{ are independent}
\]
\[
= \frac{1}{n^2} (n \text{Var}(X))
\]
\[
= \frac{1}{n} \text{Var}(X) \to 0 \text{ as } n \to \infty
\]
• Note that the proof works even if the r.v.s are only pairwise independent or even only uncorrelated

• Example: Consider the best linear MSE estimates found in the first estimation example of Lecture Notes 4 as a sequence of r.v.s \( \hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n, \ldots \), where \( \hat{X}_n \) is the best linear estimate of \( X \) given the first \( n \) observations. This sequence converges in m.s. to \( X \) since \( \text{MSE}_n \to 0 \)

• Convergence in m.s. does not necessarily imply convergence w.p.1

• Example 2: Let \( X_1, X_2, \ldots, X_n, \ldots \) be a sequence of independent r.v.s such that

\[
X_n = \begin{cases} 
0 & \text{with probability } 1 - \frac{1}{n} \\
1 & \text{with probability } \frac{1}{n} 
\end{cases}
\]

Clearly this sequence converges to 0 in m.s., but does it converge w.p.1?

It actually does not, since for \( 0 < \epsilon < 1 \) and any \( m \)

\[
P\{|X_n - 0| < \epsilon \text{ for all } n \geq m\} = \lim_{n \to \infty} \prod_{i=m}^{n} \left(1 - \frac{1}{i}\right)
\]

\[
= \lim_{n \to \infty} \prod_{i=m}^{n} \left(1 - \frac{1}{i}\right)
\]

\[
= \lim_{n \to \infty} \frac{(m-1)}{m} \frac{m}{(m+1)} \cdots \frac{(n-1)}{n}
\]

\[
= \lim_{n \to \infty} \frac{m-1}{n} \to 0 \neq 1
\]

• Also convergence w.p.1 does not imply convergence in m.s.

Consider the sequence in Example 1. Since

\[
\mathbb{E}[(Y_n - 0)^2] = \left(\frac{1}{2}\right)^n 2^{2n} = 2^n,
\]

the sequence does not converge in m.s. even though it converges w.p.1
• Example: *Convergence to a random variable*:

Flip a coin with random bias $P$ conditionally independently to obtain the sequence $X_1, X_2, \ldots, X_n, \ldots$, where as usual $X_i = 1$ if the $i$th coin flip is heads and $X_i = 0$ otherwise.

As we already know, the r.v.s $X_1, X_2, \ldots, X_n$ are not independent, but given $P = p$ they are i.i.d. Bern($p$).

It is easy to show using iterated expectation that $E(S_n) = E(X_1) = E(P)$.

In a homework exercise, you will show that $S_n \to P$ (not to $E(P)$) in m.s.

---

**Convergence in Probability**

• A sequence of r.v.s $X_1, X_2, \ldots, X_n, \ldots$ converges to a r.v. $X$ *in probability* if for any $\epsilon > 0$,

$$\lim_{n \to \infty} P\{|X_n - X| < \epsilon\} = 1$$

• Convergence w.p.1 implies convergence in probability. The converse is not necessarily true, so convergence w.p.1 is stronger than in probability.

• Example 3: Let $X_1, X_2, \ldots, X_n, \ldots$ be independent such that

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

Clearly, this sequence converges in probability to 0, since

$$P\{|X_n - 0| > \epsilon\} = P\{X_n > \epsilon\} = \frac{1}{n} \to 0 \text{ as } n \to \infty$$

But does it converge w.p.1? The answer is no (see Example 2)
• Convergence in m.s. implies convergence in probability. To show this we use the Markov inequality. For any $\epsilon > 0$,

$$P\{|X_n - X| > \epsilon\} = P\{(X_n - X)^2 > \epsilon^2\} \leq \frac{E[(X_n - X)^2]}{\epsilon^2}$$

If $X_n \rightarrow X$ in m.s., then

$$\lim_{n \rightarrow \infty} E[(X_n - X)^2] = 0 \Rightarrow \lim_{n \rightarrow \infty} P\{|X_n - X| > \epsilon\} = 0,$$

i.e., $X_n \rightarrow X$ in probability

• The converse is not necessarily true. In Example 3, $X_n$ converges in probability. Now consider

$$E[(X_n - 0)^2] = 0 \cdot \left(1 - \frac{1}{n}\right) + n^2 \cdot \frac{1}{n} = n \rightarrow \infty \text{ as } n \rightarrow \infty$$

Thus $X_n$ does not converge in m.s.

• So convergence in probability is weaker than both convergence w.p.1 and in m.s.

The Weak Law of Large Numbers

• The WLLN states that if $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of i.i.d. r.v.s with finite mean $E(X)$ and variance $\text{Var}(X)$, then

$$S_n = \frac{1}{n} \sum_{i=1}^{n} X_i \rightarrow E(X) \text{ in probability}$$

• We already proved that $S_n \rightarrow E(X)$ in m.s., and since convergence in m.s. implies convergence in probability, $S_n \rightarrow E(X)$ in probability

So, WLLN requires only uncorrelation of the r.v.s (SLLN requires independence)
Confidence Intervals

- Given $\epsilon, \delta > 0$, how large should $n$, the number of samples, be so that
  $$P\{|S_n - E(X)| \leq \epsilon\} \geq 1 - \delta,$$
  i.e., $S_n$ is within $\pm \epsilon$ of $E(X)$ with probability $\geq 1 - \delta$?

- Let's use the Chebyshev inequality:
  $$P\{|S_n - E(X)| \leq \epsilon\} = P\{|S_n - E(S_n)| \leq \epsilon\} \geq 1 - \frac{\text{Var}(S_n)}{\epsilon^2} = 1 - \frac{\text{Var}(X)}{n\epsilon^2}.$$

  So $n$ should be large enough that: $\text{Var}(X)/n\epsilon^2 \leq \delta \Rightarrow n \geq \frac{\text{Var}(X)}{\delta \epsilon^2}$

- Example: Let $\epsilon = 0.1\sigma_X$ and $\delta = 0.001$. The number of samples should satisfy
  $$n \geq \frac{\sigma_X^2}{0.001 \times 0.01\sigma_X^2} = 10^5,$$
  i.e., $10^5$ samples ensure that $S_n$ is within $\pm 0.1\sigma_X$ of $E(X)$ with probability $\geq 0.999$, independent of the distribution of $X$.

Convergence in Distribution

- A sequence of r.v.s $X_1, X_2, \ldots, X_n, \ldots$ converges in distribution to a r.v. $X$ if
  $$\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$$
  for every $x$ at which $F_X(x)$ is continuous.

- Convergence in probability implies convergence in distribution — so convergence in distribution is the weakest form of convergence we discuss.

- The most important example of convergence in distribution is the Central Limit Theorem (CLT). Let $X_1, X_2, \ldots, X_n, \ldots$ be i.i.d. r.v.s with finite mean $E(X)$ and variance $\sigma_X^2$. Consider the normalized sum
  $$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - E(X)}{\sigma_X}$$

  The sum is called normalized because $E(Z_n) = 0$ and $\text{Var}(Z_n) = 1$.

  The Central Limit Theorem states that $Z_n \to Z \sim \mathcal{N}(0,1)$ in distribution, i.e.,
  $$\lim_{n \to \infty} F_{Z_n}(z) = \Phi(z) = \begin{cases} 1 - Q(z) & z \geq 0 \\ Q(-z) & z < 0 \end{cases}$$
• Example: Let $X_1, X_2, \ldots$ be i.i.d. $U[-1, 1]$ r.v.s. The normalized sum is $Z_n = \sum_{i=1}^n X_i / \sqrt{n/3}$. The following plots show the pdf of $Z_n$ for $n = 1, 2, 4, 16$. Note how quickly the pdf of $Z_n$ approaches the Gaussian pdf.

![PDF plots](image1)

• Example: Let $X_1, X_2, \ldots$ be i.i.d. $\text{Bern}(1/2)$. The normalized sum is $Z_n = \sum_{i=1}^n (X_i - 0.5) / \sqrt{n/4}$. The following plots show the cdf of $Z_n$ for $n = 10, 20, 160$. $Z_n$ is discrete and thus has no pdf, but its cdf converges to the Gaussian cdf.

![CDF plots](image2)
Application: Confidence Intervals

- Let \(X_1, X_2, \ldots, X_n\) be i.i.d. with finite mean \(E(X)\) and variance \(\text{Var}(X)\) and let \(S_n\) be the sample mean

- Given \(\epsilon, \delta > 0\), how large should \(n\), the number of samples, be so that
\[
P\{|S_n - E(X)| \leq \epsilon\} \geq 1 - \delta?
\]

- We can use the CLT to find an estimate of \(n\) as follows:
\[
P\{|S_n - E(S_n)| \leq \epsilon\} = P\left\{\left|\frac{1}{n} \sum_{i=1}^{n} (X_i - E(X))\right| \leq \epsilon\right\}
= P\left\{\left|\frac{1}{\sigma_X \sqrt{n}} \sum_{i=1}^{n} (X_i - E(X))\right| \leq \frac{\epsilon \sqrt{n}}{\sigma_X}\right\}
\approx 1 - 2Q\left(\frac{\epsilon \sqrt{n}}{\sigma_X}\right)
\]

- Example: For \(\epsilon = 0.1\sigma_X\), \(\delta = 0.001\), set \(2Q(0.1\sqrt{n}) = 0.001\), so \(0.1\sqrt{n} = 3.3\) or \(n = 1089\) — much smaller than \(n \geq 10^5\) obtained by the Chebyshev inequality

CLT for Random Vectors

- The CLT applies to i.i.d. sequences of random vectors

- Let \(X_1, X_2, \ldots, X_n, \ldots\) be a sequence of i.i.d. \(k\)-dimensional random vectors with finite mean \(\mu\) and nonsingular covariance matrix \(\Sigma\). Define the sequence of random vectors \(Z_1, Z_2, \ldots, Z_n, \ldots\) by
\[
Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)
\]

- The Central Limit Theorem for random vectors states that as \(n \to \infty\)
\[
Z_n \to Z \sim N(0, \Sigma) \text{ in distribution}
\]

- Example: Let \(X_1, X_2, \ldots, X_n, \ldots\) be a sequence of i.i.d. 2-dimensional random vectors with
\[
f_{X_1}(x_{11}, x_{12}) = \begin{cases} x_{11} + x_{12} & 0 < x_{11} < 1, 0 < x_{12} < 1 \\ 0 & \text{otherwise} \end{cases}
\]
The following plots show the joint pdf of \(Y_n = \sum_{i=1}^{n} X_i\) for \(n = 1, 2, 3, 4\). Note how quickly it looks Gaussian.
Relationships Between Types of Convergence

- The following figure summarizes the relationships between the different types of convergence we discussed:

  
  
  with probability 1
  
  \[\iff\]
  
  in probability \[\implies\]
  
  in distribution
  
  in mean square
Random Process

- A random process (RP) (or stochastic process) is an infinite indexed collection of random variables \( \{ X(t) : t \in T \} \), defined over a common probability space.
- The index parameter \( t \) is typically time, but can also be a spatial dimension.
- Random processes are used to model random experiments that evolve in time:
  - Received sequence/waveform at the output of a communication channel
  - Packet arrival times at a node in a communication network
  - Thermal noise in a resistor
  - Scores of an NBA team in consecutive games
  - Daily price of a stock
  - Winnings or losses of a gambler
Questions Involving Random Processes

- Dependencies of the random variables of the process
  - How do future received values depend on past received values?
  - How do future prices of a stock depend on its past values?

- Long term averages
  - What is the proportion of time a queue is empty?
  - What is the average noise power at the output of a circuit?

- Extreme or boundary events
  - What is the probability that a link in a communication network is congested?
  - What is the probability that the maximum power in a power distribution line is exceeded?
  - What is the probability that a gambler will lose all his capital?

- Estimation/detection of a signal from a noisy waveform

Two Ways to View a Random Process

- A random process can be viewed as a function \( X(t, \omega) \) of two variables, time \( t \in T \) and the outcome of the underlying random experiment \( \omega \in \Omega \)
  - For fixed \( t \), \( X(t, \omega) \) is a random variable over \( \Omega \)
  - For fixed \( \omega \), \( X(t, \omega) \) is a deterministic function of \( t \), called a sample function
A random process is said to be *discrete time* if $T$ is a countably infinite set, e.g.,

- $\mathcal{N} = \{0, 1, 2, \ldots\}$
- $\mathbb{Z} = \{\ldots, -2, -1, 0, +1, +2, \ldots\}$

In this case the process is denoted by $X_n$, for $n \in \mathcal{N}$, a countably infinite set, and is simply an infinite sequence of random variables.

A sample function for a discrete time process is called a *sample sequence* or *sample path*.

A discrete-time process can comprise discrete, continuous, or mixed r.v.s.

**Example**

- Let $Z \sim U[0, 1]$, and define the discrete time process $X_n = Z^n$ for $n \geq 1$.

Sample paths:
First-order pdf of the process: For each \( n \), \( X_n = Z^n \) is a r.v.; the sequence of pdfs of \( X_n \) is called the first-order pdf of the process.

Since \( X_n \) is a differentiable function of the continuous r.v. \( Z \), we can find its pdf as

\[
f_{X_n}(x) = \frac{1}{nx^{(n-1)/n}} = \frac{1}{n} x^{\frac{1}{n}-1}, \quad 0 \leq x \leq 1
\]

Continuous Time Random Process

- A random process is continuous time if \( T \) is a continuous set.
- Example: Sinusoidal Signal with Random Phase

\[
X(t) = \alpha \cos(\omega t + \Theta), \quad t \geq 0
\]

where \( \Theta \sim U[0, 2\pi] \) and \( \alpha \) and \( \omega \) are constants.

- Sample functions:
• The first-order pdf of the process is the pdf of $X(t) = \alpha \cos(\omega t + \Theta)$. In an earlier homework exercise, we found it to be

$$f_{X(t)}(x) = \frac{1}{\alpha \pi \sqrt{1 - (x/\alpha)^2}}, \quad -\alpha < x < +\alpha$$

The graph of the pdf is shown below

Note that the pdf is independent of $t$. (The process is stationary)

**Specifying a Random Process**

• In the above examples we specified the random process by describing the set of sample functions (sequences, paths) and explicitly providing a probability measure over the set of events (subsets of sample functions)

• This way of specifying a random process has very limited applicability, and is suited only for very simple processes

• A random process is typically specified (directly or indirectly) by specifying all its $n$-th order cdfs (pdfs, pmfs), i.e., the joint cdf (pdf, pmf) of the samples

$$X(t_1), X(t_2), \ldots, X(t_n)$$

for every order $n$ and for every set of $n$ points $t_1, t_2, \ldots, t_n \in T$

• The following examples of important random processes will be specified (directly or indirectly) in this manner
**Important Classes of Random Processes**

- **IID process**: \( \{X_n : n \in \mathcal{N}\} \) is an IID process if the r.v.s \( X_n \) are i.i.d.
  
  Examples:
  
  - Bernoulli process: \( X_1, X_2, \ldots, X_n, \ldots \) i.i.d. \( \sim \) Bern(\( p \))
  
  - Discrete-time white Gaussian noise (WGN): \( X_1, \ldots, X_n, \ldots \) i.i.d. \( \sim \mathcal{N}(0, N) \)

- Here we specified the \( n \)-th order pmfs (pdfs) of the processes by specifying the first-order pmf (pdf) and stating that the r.v.s are independent

- It would be quite difficult to provide the specifications for an IID process by specifying the probability measure over the subsets of the sample space

---

**The Random Walk Process**

- Let \( Z_1, Z_2, \ldots, Z_n, \ldots \) be i.i.d., where
  
  \[
  Z_n = \begin{cases} 
  +1 \text{ with probability } \frac{1}{2} \\
  -1 \text{ with probability } \frac{1}{2} 
  \end{cases}
  \]

- The random walk process is defined by
  
  \[
  X_0 = 0
  
  X_n = \sum_{i=1}^{n} Z_i, \quad n \geq 1
  \]

- Again this process is specified by (indirectly) specifying all \( n \)-th order pmfs

- Sample path: The sample path for a random walk is a sequence of integers, e.g.,
  
  \[
  0, +1, 0, -1, -2, -3, -4, \ldots
  \]

  or
  
  \[
  0, +1, +2, +3, +4, +3, +4, +3, +4, \ldots
  \]
Example:

\begin{align*}
\text{z}_n & : \quad 1 \quad 1 \quad 1 \quad -1 \quad 1 \quad -1 \quad -1 \quad -1 \quad -1 \quad -1 \\
\text{• First-order pmf:} \quad \text{The first-order pmf is } & \quad P\{X_n = k\} \quad \text{as a function of } n. \quad \text{Note that} \\
& \quad k \in \{-n, -(n-2), \ldots, -2, 0, +2, \ldots, +(n-2), +n\} \quad \text{for } n \text{ even} \\
& \quad k \in \{-n, -(n-2), \ldots, -1, +1, +3, \ldots, +(n-2), +n\} \quad \text{for } n \text{ odd} \\
& \quad \text{Hence, } \quad P\{X_n = k\} = 0 \quad \text{if } n+k \text{ is odd, or if } k < -n \text{ or } k > n
\end{align*}

Now for \( n + k \) even, let \( a \) be the number of \(+1\)'s in \( n \) steps, then the number of \(-1\)'s is \( n - a \), and we find that

\[ k = a - (n - a) = 2a - n \quad \Rightarrow \quad a = \frac{n + k}{2} \]

Thus

\[ P\{X_n = k\} = P\{\frac{1}{2}(n+k) \text{ heads in } n \text{ independent coin tosses}\} \]

\[ = \left( \frac{n}{n+k} \right) \cdot 2^{-n} \quad \text{for } n + k \text{ even and } -n \leq k \leq n \]
**Markov Processes**

- A discrete-time random process $X_n$ is said to be a Markov process if the *future and past are conditionally independent given its present value*.

- Mathematically this can be rephrased in several ways. For example, if the r.v.s $\{X_n : n \geq 1\}$ are discrete, then the process is Markov iff
  
  $$
  p_{X_{n+1}|X^n}(x_{n+1}|x_n, x^{n-1}) = p_{X_{n+1}|X^n}(x_{n+1}|x_n)
  $$
  
  for every $n$.

- IID processes are Markov.

- The random walk process is Markov. To see this consider
  
  $$
  P\{X_{n+1} = x_{n+1} \mid X^n = x^n\} = P\{X_n + Z_{n+1} = x_{n+1} \mid X^n = x^n\}
  = P\{X_n + Z_{n+1} = x_{n+1} \mid X_n = x_n\}
  = P\{X_{n+1} = x_{n+1} \mid X_n = x_n\}
  $$

**Independent Increment Processes**

- A discrete-time random process $\{X_n : n \geq 0\}$ is said to be *independent increment* if the *increment* random variables
  
  $$
  X_{n_1}, X_{n_2} - X_{n_1}, \ldots, X_{n_k} - X_{n_{k-1}}
  $$
  
  are independent for all sequences of indices such that $n_1 < n_2 < \cdots < n_k$.

- Example: Random walk is an independent increment process because
  
  $$
  X_{n_1} = \sum_{i=1}^{n_1} Z_i, \quad X_{n_2} - X_{n_1} = \sum_{i=n_1+1}^{n_2} Z_i, \quad \ldots, \quad X_{n_k} - X_{n_{k-1}} = \sum_{i=n_{k-1}+1}^{n_k} Z_i
  $$
  
  are independent because they are functions of independent random vectors.

- The independent increment property makes it easy to find the $n$-th order pmfs of a random walk process from knowledge only of the first-order pmf.
- Example: Find $P\{X_5 = 3, X_{10} = 6, X_{20} = 10\}$ for random walk process $\{X_n\}$

Solution: We use the independent increment property as follows

$$P\{X_5 = 3, X_{10} = 6, X_{20} = 10\} = P\{X_5 = 3, X_{10} - X_5 = 3, X_{20} - X_{10} = 4\}$$

$$= P\{X_5 = 3\}P\{X_5 = 3\}P\{X_{10} = 4\}$$

$$= \left(\frac{5}{4}\right)^2 \left(\frac{5}{4}\right)^2 \left(\frac{10}{7}\right)^2 \frac{1}{2^{10}} = 3000 \cdot 2^{-20}$$

- In general if a process is independent increment, then it is also Markov. To see this let $X_n$ be an independent increment process and define

$$\Delta X^n = [X_1, X_2 - X_1, \ldots, X_n - X_{n-1}]^T$$

Then

$$p_{X_{n+1}|X^n}(x_{n+1} | x^n) = P\{X_{n+1} = x_{n+1} | X^n = x^n\}$$

$$= P\{X_{n+1} - X_n + X_n = x_{n+1} | \Delta X^n = \Delta x^n, X_n = x_n\}$$

$$= P\{X_{n+1} = x_{n+1} | X_n = x_n\}$$

- The converse is not necessarily true, e.g., IID processes are Markov but not independent increment

EE 278B: Random Processes 6 – 17

- The independent increment property can be extended to continuous-time processes:

A process $X(t), t \geq 0,$ is said to be independent increment if $X(t_1), X(t_2) - X(t_1), \ldots, X(t_k) - X(t_{k-1})$ are independent for every $0 \leq t_1 < t_2 < \ldots < t_k$ and every $k \geq 2$

- Markovity can also be extended to continuous-time processes:

A process $X(t)$ is said to be Markov if $X(t_{k+1})$ and $(X(t_1), \ldots, X(t_{k-1}))$ are conditionally independent given $X(t_k)$ for every $0 \leq t_1 < t_2 < \ldots < t_k < t_{k+1}$ and every $k \geq 3$
Counting Processes and Poisson Process

- A continuous-time random process $N(t), t \geq 0$, is said to be a counting process if $N(0) = 0$ and $N(t) = n, n \in \{0, 1, 2, \ldots\}$, is the number of events from 0 to $t$ (hence $N(t_2) \geq N(t_1)$ for every $t_2 > t_1 \geq 0$)

Sample path of a counting process:

$t_1, t_2, \ldots$ are the arrival times or the wait times of the events

$t_1, t_2 - t_1, \ldots$ are the interarrival times of the events

The events may be:
- Photon arrivals at an optical detector
- Packet arrivals at a router
- Student arrivals at a class

- The Poisson process is a counting process in which the events are “independent of each other”

- More precisely, $N(t)$ is a Poisson process with rate (intensity) $\lambda > 0$ if:
  - $N(0) = 0$
  - $N(t)$ is independent increment
  - $(N(t_2) - N(t_1)) \sim \text{Poisson}(\lambda(t_2 - t_1))$ for all $t_2 > t_1 \geq 0$

- To find the $k$th order pmf, we use the independent increment property

  \[
  P\{N(t_1) = n_1, N(t_2) = n_2, \ldots, N(t_k) = n_k\} = P\{N(t_1) = n_1, N(t_2) - N(t_1) = n_2 - n_1, \ldots, N(t_k) - N(t_{k-1}) = n_k - n_{k-1}\}
  \]

  \[
  = p_{N(t_1)}(n_1)p_{N(t_2)-N(t_1)}(n_2-n_1) \cdots p_{N(t_k)-N(t_{k-1})}(n_k-n_{k-1})
  \]
• Example: Packets arrive at a router according to a Poisson process \( N(t) \) with rate \( \lambda \). Assume the service time for each packet \( T \sim \text{Exp}(\beta) \) is independent of \( N(t) \) and of each other. What is the probability that \( k \) packets arrive during a service time?

• **Merging**: The sum of independent Poisson process is Poisson. This is a consequence of the infinite divisibility of the Poisson r.v.

• **Branching**: Let \( N(t) \) be a Poisson process with rate \( \lambda \). We split \( N(t) \) into two counting subprocesses \( N_1(t) \) and \( N_2(t) \) such that \( N(t) = N_1(t) + N_2(t) \) as follows:

Each event is randomly and independently assigned to process \( N_1(t) \) with probability \( p \), otherwise it is assigned to \( N_2(t) \).

Then \( N_1(t) \) is a Poisson process with rate \( p\lambda \) and \( N_2(t) \) is a Poisson process with rate \( (1 - p)\lambda \).

This can be generalized to splitting a Poisson process into more than two processes.

---

### Related Processes

- **Arrival time process**: Let \( N(t) \) be Poisson with rate \( \lambda \). The arrival time process \( T_n, n \geq 0 \) is a discrete time process such that:
  - \( T_0 = 0 \)
  - \( T_n \) is the arrival time of the \( n \)th event of \( N(t) \)

- **Interarrival time process**: Let \( N(t) \) be a Poisson process with rate \( \lambda \). The interarrival time process is \( X_n = T_n - T_{n-1} \) for \( n \geq 1 \)

  - \( X_n \) is an IID process with \( X_n \sim \text{Exp}(\lambda) \)

  - \( T_n = \sum_{i=1}^{n} X_i \) is an independent increment process with \( T_{n_2} - T_{n_1} \sim \text{Gamma}(\lambda, n_2 - n_1) \) for \( n_2 > n_1 \geq 1 \), i.e.,

    \[
    f_{T_{n_2} - T_{n_1}}(t) = \frac{\lambda^{n_2-n_1} e^{-\lambda t}}{(n_2-n_1-1)!} t^{n_2-n_1-1}
    \]

- Example: Let \( N_1(t) \) and \( N_2(t) \) be two independent Poisson processes with rates \( \lambda_1 \) and \( \lambda_2 \), respectively. What is the probability that \( N_1(t) = 1 \) before \( N_2(t) = 1? \)
• **Random telegraph process**: A random telegraph process $Y(t), t \geq 0$, assumes values of $+1$ and $-1$ with $Y(0) = +1$ with probability $1/2$ and $-1$ with probability $1/2$, and $Y(t)$ changes polarities with each event of a Poisson process with rate $\lambda > 0$.

Sample path:

![Sample path](image)

---

**Mean and Autocorrelation Functions**

- For a random vector $X$ the first and second order moments are
  - mean $\mu = \mathbb{E}(X)$
  - correlation matrix $R_X = \mathbb{E}(XX^T)$

- For a random process $X(t)$ the first and second order moments are
  - *mean* function: $\mu_X(t) = \mathbb{E}(X(t))$ for $t \in \mathcal{T}$
  - *autocorrelation* function: $R_X(t_1, t_2) = \mathbb{E}(X(t_1)X(t_2))$ for $t_1, t_2 \in \mathcal{T}$

- The *autocovariance function* of a random process is defined as
  
  $C_X(t_1, t_2) = \mathbb{E} \left[ (X(t_1) - \mathbb{E}(X(t_1))) (X(t_2) - \mathbb{E}(X(t_2))) \right]$ 

  The autocovariance function can be expressed using the mean and autocorrelation functions as
  
  $C_X(t_1, t_2) = R_X(t_1, t_2) - \mu_X(t_1)\mu_X(t_2)$
Examples

- IID process:

\[
\mu_X(n) = E(X_1) \\
R_X(n_1, n_2) = E(X_{n_1}X_{n_2}) = \begin{cases} 
E(X_1^2) & n_1 = n_2 \\
(E(X_1))^2 & n_1 \neq n_2 
\end{cases}
\]

- Random phase signal process:

\[
\mu_X(t) = E(\alpha \cos(\omega t + \Theta)) = \int_0^{2\pi} \frac{\alpha}{2\pi} \cos(\omega t + \theta) d\theta = 0 \\
R_X(t_1, t_2) = E(X(t_1)X(t_2)) \\
= \int_0^{2\pi} \frac{\alpha^2}{2\pi} \cos(\omega t_1 + \theta) \cos(\omega t_2 + \theta) d\theta \\
= \int_0^{2\pi} \frac{\alpha^2}{4\pi} [\cos(\omega(t_1 + t_2) + 2\theta) + \cos(\omega(t_1 - t_2))] d\theta \\
= \frac{\alpha^2}{2} \cos(\omega(t_1 - t_2))
\]

- Random walk:

\[
\mu_X(n) = E\left(\sum_{i=1}^{n} Z_i\right) = \sum_{i=1}^{n} 0 = 0 \\
R_X(n_1, n_2) = E(X_{n_1}X_{n_2}) \\
= E[X_{n_1}(X_{n_2} - X_{n_1} + X_{n_1})] \\
= E(X_{n_1}^2) = n_1 \quad \text{assuming } n_2 \geq n_1 \\
= \min\{n_1, n_2\} \quad \text{in general}
\]

- Poisson process:

\[
\mu_N(t) = \lambda t \\
R_N(t_1, t_2) = E(N(t_1)N(t_2)) \\
= E[N(t_1)(N(t_2) - N(t_1) + N(t_1))] \\
= \lambda t_1 \times \lambda(t_2 - t_1) + \lambda t_1 + \lambda^2 t_1^2 = \lambda t_1 + \lambda^2 t_1 t_2 \quad \text{assuming } t_2 \geq t_1 \\
= \lambda \min\{t_1, t_2\} + \lambda^2 t_1 t_2
\]
Gaussian Random Processes

- A Gaussian random process (GRP) is a random process \( X(t) \) such that
  \[
  [X(t_1), X(t_2), \ldots, X(t_n)]^T
  \]
is a GRV for all \( t_1, t_2, \ldots, t_n \in T \)
- Since the joint pdf for a GRV is specified by its mean and covariance matrix, a GRP is specified by its mean \( \mu_X(t) \) and autocorrelation \( R_X(t_1, t_2) \) functions
- Example: The discrete time WGN process is a GRP

Gauss-Markov Process

- Let \( Z_n, n \geq 1, \) be a WGN process, i.e., an IID process with \( Z_1 \sim \mathcal{N}(0, N) \)
The Gauss-Markov process is a first-order autoregressive process defined by
  \[
  X_1 = Z_1 \\
  X_n = \alpha X_{n-1} + Z_n, \quad n > 1,
  \]
where \(|\alpha| < 1\)
- This process is a GRP, since \( X_1 = Z_1 \) and \( X_k = \alpha X_{k-1} + Z_k \) where \( Z_1, Z_2, \ldots \) are i.i.d. \( \mathcal{N}(0, N), \)
  \[
  \begin{bmatrix}
  X_1 \\
  X_2 \\
  X_3 \\
  \vdots \\
  X_n
  \end{bmatrix} =
  \begin{bmatrix}
  1 & 0 & \cdots & 0 & 0 \\
  \alpha & 1 & \cdots & 0 & 0 \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  \alpha^{n-2} & \alpha^{n-3} & \cdots & 1 & 0 \\
  \alpha^{n-1} & \alpha^{n-2} & \cdots & \alpha & 1
  \end{bmatrix}
  \begin{bmatrix}
  Z_1 \\
  Z_2 \\
  Z_3 \\
  \vdots \\
  Z_n
  \end{bmatrix}
  \]
is a linear transformation of a GRV and is therefore a GRV
- Clearly, the Gauss-Markov process is Markov. It is not, however, an independent increment process
• Mean and covariance functions:
\[ \mu_X(n) = E(X_n) = E(\alpha X_{n-1} + Z_n) \]
\[ = \alpha E(X_{n-1}) + E(Z_n) = \alpha E(X_{n-1}) = \alpha^{n-1} E(Z_1) = 0 \]

To find the autocorrelation function, for \( n_2 > n_1 \) we write
\[ X_{n_2} = \alpha^{n_2-n_1} X_{n_1} + \sum_{i=0}^{n_2-n_1-1} \alpha^i Z_{n_2-i} \]

Thus
\[ R_X(n_1, n_2) = E(X_{n_1}X_{n_2}) = \alpha^{n_2-n_1} E(X_{n_1}^2) + 0, \]

since \( X_{n_1} \) and \( Z_{n_2-i} \) are independent, zero mean for \( 0 \leq i \leq n_2 - n_1 - 1 \)

Next, to find \( E(X_{n_1}^2) \), consider
\[ E(X_{n_1}^2) = N \]
\[ E(X_{n_1}^2) = E[ (\alpha X_{n_1-1} + Z_{n_1})^2 ] = \alpha^2 E(X_{n_1-1}^2) + N \]

Thus
\[ E(X_{n_1}^2) = \frac{1 - \alpha^{2n_1}}{1 - \alpha^2} N \]

Finally the autocorrelation function is
\[ R_X(n_1, n_2) = \alpha^{n_2-n_1} \frac{1 - \alpha^{2 \min\{n_1,n_2\}}}{1 - \alpha^2} N \]

• Estimation of Gauss-Markov process: Suppose we observe a noisy version of the Gauss-Markov process,
\[ Y_n = X_n + W_n, \]

where \( W_n \) is a WGN process independent of \( Z_n \) with average power \( Q \)

We can use the Kalman filter from Lecture Notes 4 to estimate \( X_{i+1} \) from \( Y^i \) as follows:
Initialization:
\[ \hat{X}_{1|0} = 0 \]
\[ \sigma^2_{1|0} = N \]
Update: For $i = 2, 3, \ldots,$

$$\hat{X}_{i+1|i} = \alpha \hat{X}_{i|i-1} + k_i (Y_i - \hat{X}_{i|i-1}),$$

$$\sigma^2_{i+1|i} = \alpha (\alpha - k_i) \sigma^2_{i|i-1} + N,$$

where

$$k_i = \frac{\alpha \sigma^2_{i|i-1}}{\sigma^2_{i|i-1} + Q}$$

Substituting from the $k_i$ equation into the MSE update equation, we obtain

$$\sigma^2_{i+1|i} = \frac{\alpha^2 Q \sigma^2_{i|i-1}}{\sigma^2_{i|i-1} + Q} + N,$$

This is a Riccati recursion (a quadratic recursion in the MSE) and has a steady state solution:

$$\sigma^2 = \frac{\alpha^2 Q \sigma^2}{\sigma^2 + Q} + N$$

Solving this quadratic equation, we obtain

$$\sigma^2 = \frac{N - (1 - \alpha^2)Q + \sqrt{4NQ + (N - (1 - \alpha^2)Q)^2}}{2}$$

The Kalman gain $k_i$ converges to

$$k = \frac{-N - (1 - \alpha^2)Q + \sqrt{4NQ + (N - (1 - \alpha^2)Q)^2}}{2\alpha Q}$$

and the steady-state Kalman filter is

$$\hat{X}_{i+1|i} = \alpha \hat{X}_{i|i-1} + k (Y_i - \hat{X}_{i|i-1})$$
Stochastic Processes

- Stationarity refers to time invariance of some, or all, of the statistics of a random process, such as mean, autocorrelation, $n$-th-order distribution
- We define two types of stationarity: strict sense (SSS) and wide sense (WSS)
- A random process $X(t)$ (or $X_n$) is said to be SSS if all its finite order distributions are time invariant, i.e., the joint cdfs (pdfs, pmfs) of
  \[ X(t_1), X(t_2), \ldots, X(t_k) \quad \text{and} \quad X(t_1 + \tau), X(t_2 + \tau), \ldots, X(t_k + \tau) \]
  are the same for all $k$, all $t_1, t_2, \ldots, t_k$, and all time shifts $\tau$
- So for a SSS process, the first-order distribution is independent of $t$, and the second-order distribution — the distribution of any two samples $X(t_1)$ and $X(t_2)$ — depends only on $\tau = t_2 - t_1$
- To see this, note that from the definition of stationarity, for any $t$, the joint distribution of $X(t_1)$ and $X(t_2)$ is the same as the joint distribution of $X(t_1 + (t - t_1)) = X(t)$ and $X(t_2 + (t - t_1)) = X(t + (t_2 - t_1))$
• Example: The random phase signal \( X(t) = \alpha \cos(\omega t + \Theta) \) where \( \Theta \in U[0, 2\pi] \) is SSS
  ○ We already know that the first order pdf is
    \[
    f_{X(t)}(x) = \frac{1}{\pi \alpha \sqrt{1 - (x/\alpha)^2}}, \quad -\alpha < x < +\alpha
    \]
    which is independent of \( t \), and is therefore stationary
  ○ To find the second order pdf, note that if we are given the value of \( X(t) \) at one point, say \( t_1 \), there are (at most) two possible sample functions:

  ![Sample paths](image)

  The second order pdf can thus be written as
  \[
  f_{X(t_1), X(t_2)}(x_1, x_2) = f_{X(t_1)}(x_1) f_{X(t_2)}|_{X(t_1)}(x_2|x_1)
  = f_{X(t_1)}(x_1) \left( \frac{1}{2} \delta(x_2 - x_{21}) + \frac{1}{2} \delta(x_2 - x_{22}) \right),
  \]
  which depends only on \( t_2 - t_1 \), and thus the second order pdf is stationary
  ○ Now if we know that \( X(t_1) = x_1 \) and \( X(t_2) = x_2 \), the sample path is totally determined (except when \( x_1 = x_2 = 0 \), where two paths may be possible), and thus all \( n \)-th order pdfs are stationary
  
  • IID processes are SSS
  • Random walk and Poisson processes are not SSS
  • The Gauss-Markov process (as we defined it) is not SSS. However, if we set \( X_1 \) to the steady state distribution of \( X_n \), it becomes SSS (see homework exercise)
Wide-Sense Stationary Random Processes

- A random process $X(t)$ is said to be *wide-sense stationary* (WSS) if its mean and autocorrelation functions are time invariant, i.e.,
  - $E(X(t)) = \mu$, independent of $t$
  - $R_X(t_1, t_2)$ is a function only of the time difference $t_2 - t_1$
  - $E[X(t)^2] < \infty$ (technical condition)

- Since $R_X(t_1, t_2) = R_X(t_2, t_1)$, for any wide sense stationary process $X(t)$, $R_X(t_1, t_2)$ is a function only of $|t_2 - t_1|$

- Clearly SSS $\Rightarrow$ WSS. The converse is not necessarily true

- Example: Let
  
  $$X(t) = \begin{cases} 
  + \sin t & \text{with probability } \frac{1}{4} \\
  - \sin t & \text{with probability } \frac{1}{4} \\
  + \cos t & \text{with probability } \frac{1}{4} \\
  - \cos t & \text{with probability } \frac{1}{4} 
  \end{cases}$$

  - $E(X(t)) = 0$ and $R_X(t_1, t_2) = \frac{1}{2} \cos(t_2 - t_1)$, thus $X(t)$ is WSS
  - But $X(0)$ and $X\left(\frac{\pi}{4}\right)$ do not have the same pmf (different ranges), so the first order pmf is not stationary, and the process is not SSS

- For Gaussian random processes, WSS $\Rightarrow$ SSS, since the process is completely specified by its mean and autocorrelation functions

- Random walk is not WSS, since $R_X(n_1, n_2) = \min\{n_1, n_2\}$ is not time invariant; similarly Poisson process is not WSS
Autocorrelation Function of WSS Processes

- Let $X(t)$ be a WSS process. Relabel $R_X(t_1, t_2)$ as $R_X(\tau)$ where $\tau = t_1 - t_2$

1. $R_X(\tau)$ is real and even, i.e., $R_X(\tau) = R_X(-\tau)$ for every $\tau$

2. $|R_X(\tau)| \leq R_X(0) = \mathbb{E}[X^2(t)]$, the “average power” of $X(t)$
   
   This can be shown as follows. For every $t$,
   $$
   (R_X(\tau))^2 = [\mathbb{E}(X(t)X(t + \tau))]^2
   \leq \mathbb{E}[X^2(t)] \mathbb{E}[X^2(t + \tau)] \quad \text{by Schwarz inequality}
   \leq (R_X(0))^2 \quad \text{by stationarity}
   $$

3. If $R_X(T) = R_X(0)$ for some $T \neq 0$, then $R_X(\tau)$ is periodic with period $T$ and so is $X(t)$ (with probability 1) !! That is,
   $$
   R_X(\tau) = R_X(\tau + T), \quad X(t) = X(t + T) \text{ w.p.1 for every } \tau
   $$

- Example: The autocorrelation function for the periodic signal with random phase $X(t) = \alpha \cos(\omega t + \Theta)$ is $R_X(\tau) = \frac{\alpha^2}{2} \cos \omega \tau$ (also periodic)

- To prove property 3, we again use the Schwarz inequality: For every $\tau$,
   $$
   [R_X(\tau) - R_X(\tau + T)]^2 = [\mathbb{E}(X(t)(X(t + \tau) - X(t + \tau + T)))]^2
   \leq \mathbb{E}[X^2(t)] \mathbb{E}[(X(t + \tau) - X(t + \tau + T))^2]
   = R_X(0)(2R_X(0) - 2R_X(T))
   = R_X(0)(2R_X(0) - 2R_X(0)) = 0
   $$
   Thus $R_X(\tau) = R_X(\tau + T)$ for all $\tau$, i.e., $R_X(\tau)$ is periodic with period $T$

- The above properties of $R_X(\tau)$ are necessary but not sufficient for a function to qualify as an autocorrelation function for a WSS process
The necessary and sufficient conditions for a function to be an autocorrelation function for a WSS process is that it be real, even, and nonnegative definite.

By nonnegative definite we mean that for any \( n \), any \( t_1, t_2, \ldots, t_n \) and any real vector \( a = (a_1, \ldots, a_n) \),

\[
\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j R(t_i - t_j) \geq 0
\]

To see why this is necessary, recall that the correlation matrix for a random vector must be nonnegative definite, so if we take a set of \( n \) samples from the WSS random process, their correlation matrix must be nonnegative definite.

The condition is sufficient since such an \( R(\tau) \) can specify a zero mean stationary Gaussian random process.

The nonnegative definite condition may be difficult to verify directly. It turns out, however, to be equivalent to the condition that the Fourier transform of \( R_X(\tau) \), which is called the power spectral density \( S_X(f) \), is nonnegative for all frequencies \( f \).

---

**Which Functions Can Be an** \( R_X(\tau) \)?

1. \( e^{-\alpha \tau} \)

2. \( e^{-\alpha |\tau|} \)

3. \( \text{sinc} \tau \)

4. \( \text{sinc} \tau \)
Which Functions can be an $R_X(\tau)$?

5. \[ \tau \]

6. \[ 2^{-|n|} \]

7. \[ \tau \]

8. \[ \tau \]

Interpretation of Autocorrelation Function

- Let $X(t)$ be WSS with zero mean. If $R_X(\tau)$ drops quickly with $\tau$, this means that samples become uncorrelated quickly as we increase $\tau$. Conversely, if $R_X(\tau)$ drops slowly with $\tau$, samples are highly correlated.

- So $R_X(\tau)$ is a measure of the rate of change of $X(t)$ with time $t$, i.e., the frequency response of $X(t)$.

- It turns out that this is not just an intuitive interpretation — the Fourier transform of $R_X(\tau)$ (the power spectral density) is in fact the average power density of $X(t)$ over frequency.
Power Spectral Density

- The power spectral density (psd) of a WSS random process $X(t)$ is the Fourier transform of $R_X(\tau)$:
  \[ S_X(f) = \mathcal{F}(R_X(\tau)) = \int_{-\infty}^{\infty} R_X(\tau) e^{-i2\pi \tau f} d\tau \]

- For a discrete time process $X_n$, the power spectral density is the discrete-time Fourier transform (DTFT) of the sequence $R_X(n)$:
  \[ S_X(f) = \sum_{n=-\infty}^{\infty} R_X(n) e^{-i2\pi nf}, \quad |f| < \frac{1}{2} \]

- $R_X(\tau)$ (or $R_X(n)$) can be recovered from $S_X(f)$ by taking the inverse Fourier transform or inverse DTFT:
  \[ R_X(\tau) = \int_{-\infty}^{\infty} S_X(f) e^{i2\pi \tau f} df \]
  \[ R_X(n) = \int_{-\frac{1}{2}}^{\frac{1}{2}} S_X(f) e^{i2\pi nf} df \]

Properties of the Power Spectral Density

1. $S_X(f)$ is real and even, since the Fourier transform of the real and even function $R_X(\tau)$ is real and even

2. $\int_{-\infty}^{\infty} S_X(f) df = R_X(0) = E(X^2(t))$, the average power of $X(t)$, i.e., the area under $S_X$ is the average power

3. $S_X(f)$ is the average power density, i.e., the average power of $X(t)$ in the frequency band $[f_1, f_2]$ is
  \[ \int_{-f_2}^{-f_1} S_X(f) df + \int_{f_1}^{f_2} S_X(f) df = 2 \int_{f_1}^{f_2} S_X(f) df \]
  (we will show this soon)

- From property 3, it follows that $S_X(f) \geq 0$. Why?

- In general, a function $S(f)$ is a psd if and only if it is real, even, nonnegative, and
  \[ \int_{-\infty}^{\infty} S(f) df < \infty \]
Examples

1. \( R_X(\tau) = e^{-\alpha|\tau|} \)
   \[ S_X(f) = \frac{2\alpha}{\alpha^2 + (2\pi f)^2} \]

2. \( R_X(\tau) = \frac{\alpha^2}{2} \cos \omega \tau \)
   \[ S_X(f) = \frac{\alpha^2}{4} \delta(2\pi f - \omega) + \frac{\alpha^2}{4} \delta(2\pi f + \omega) \]

3. \( R_X(n) = 2^{-|n|} \)
   \[ S_X(f) = \frac{3}{5 - 4 \cos 2\pi f} \]

4. *Discrete time white noise process*: \( X_1, X_2, \ldots, X_n, \ldots \) zero mean, uncorrelated, with average power \( N \)
   
   \[ R_X(n) = \begin{cases} 
   N & n = 0 \\
   0 & \text{otherwise} 
   \end{cases} \]
   \[ S_X(f) \quad \text{for } f = \frac{\pm \omega}{2\pi} \]

   If \( X_n \) is also a GRP, then we obtain a *discrete time WGN process*
5. **Bandlimited white noise process**: WSS zero mean process $X(t)$ with

\[
S_X(f) = \frac{N}{2} \quad \text{for all } f
\]

\[
R_X(\tau) = NB \text{sinc} 2B\tau
\]

For any $t$, the samples $X(t \pm \frac{n}{2B})$ for $n = 0, 1, 2, \ldots$ are uncorrelated.

6. **White noise process**: If we let $B \to \infty$ in the previous example, we obtain a white noise process, which has

\[
S_X(f) = \frac{N}{2} \quad \text{for all } f
\]

\[
R_X(\tau) = \frac{N}{2} \delta(\tau)
\]

If, in addition, $X(t)$ is a GRP, then we obtain the famous **white Gaussian noise (WGN) process**

- **Remarks on white noise**:
  - For a white noise process, all samples are uncorrelated
  - The process is not physically realizable, since it has infinite power
  - However, it plays a similar role in random processes to point mass in physics and delta function in linear systems
  - Thermal noise and shot noise are well modeled as white Gaussian noise, since they have very flat psd over very wide band (GHz)
Continuity and Integration of Random Processes

- We are all familiar with the definitions of continuity and integration for deterministic functions as limits.

- Using the notions of convergence discussed in Lecture Notes 5, we can define these notions for random processes. We focus only on m.s. convergence.

**Continuity**: A process \(X(t)\) is said to be *mean square continuous* if for every \(t\)

\[
\lim_{s \to t} E[(X(s) - X(t))^2] = 0
\]

- The continuity of \(X(t)\) depends only on its autocorrelation function \(R_X(t_1, t_2)\).

  In fact, the following statements are all equivalent:

1. \(R_X(t_1, t_2)\) is continuous at all points of the form \((t, t)\).
2. \(X(t)\) is m.s. continuous.
3. \(R_X(t_1, t_2)\) is continuous in \(t_1, t_2\).

Proof:

1. 1 implies 2: Since if \(R_X(t_1, t_2)\) is continuous at all points \((t, t)\),

\[
E[(X(t) - X(s))^2] = R_X(t, t) + R_X(s, s) - 2R_X(s, t) \to 0 \text{ as } s \to t
\]

2. 2 implies 3: Consider

\[
R_X(s_1, s_2) = E[X(s_1)X(s_2)]
= E[(X(t_1) + (X(s_1) - X(t_1)))(X(t_2) + (X(s_2) - X(t_2)))]
= R_X(t_1, t_2) + E[X(t_1)(X(s_2) - X(t_2))] + E[X(t_2)(X(s_1) - X(t_1))]
+ E[(X(s_1) - X(t_1))(X(s_2) - X(t_2))]
\leq R_X(t_1, t_2) + \sqrt{E[X^2(t_1)]} \sqrt{E[(X(s_2) - X(t_2))^2]}
+ \sqrt{E[X^2(t_2)]} \sqrt{E[(X(s_1) - X(t_1))^2]}
+ \sqrt{E[(X(s_1) - X(t_1))^2]} \sqrt{E[(X(s_2) - X(t_2))^2]}
\]

Schwartz inequality

\[
\to R_X(t_1, t_2) \text{ as } s_1 \to t_1 \text{ and } s_2 \to t_2
\]

since \(X(t)\) is m.s. continuous.

3. Since 3 implies 1, we are done.
Example: The Poisson process \( N(t) \) with rate \( \lambda > 0 \) is m.s. continuous, since its autocorrelation function,

\[
R_N(t_1, t_2) = \lambda \min\{t_1, t_2\} + \lambda^2 t_1 t_2
\]
is a continuous function.

Integration: Let \( X(t) \) be a RP and \( h(t) \) be a function. We can define the integral

\[
\int_a^b h(t)X(t)dt
\]
as the limit of a sum (as in Riemann integral of a deterministic function) in m.s.

Let \( \Delta > 0 \) such that \( b - a = n\Delta \) and

\[
a \leq \tau_1 \leq a + \Delta \leq \tau_2 \leq a + 2\Delta \leq \cdots \leq \tau_{n-1} \leq a + (n-1)\Delta \leq \tau_n \leq a + n\Delta = b,
\]

then the corresponding Riemann sum is

\[
\sum_{i=1}^{n-1} h(\tau_i)X(\tau_i)\Delta
\]
The above integral then exists if this sum has a limit in m.s. as \( \Delta \to 0 \)

Moreover, if the random integral exists for all \( a, b \), then we can define

\[
\int_{-\infty}^{\infty} h(t)X(t)dt = \lim_{a,b \to \infty} \int_a^b h(t)X(t)dt \quad \text{in m.s.}
\]

Fact: The existence of the m.s. integral depends only on \( R_X \) and \( h \)

More specifically, the above integral exists iff

\[
\int_a^b \int_a^b R_X(t_1, t_2)h(t_1)h(t_2)dt_1dt_2
\]
exists (in the normal sense)

Remark: We are skipping several mathematical details here. In what follows, we use the above fact to justify the existence of integrals involving random processes and in interchanging expectation and integration.
• Let $X(t)$ be SSS or only WSS

• Ergodicity of $X(t)$ means that certain time averages converge to their respective statistical averages

• Mean ergodic process: Let $X(t)$ be a WSS and m.s. continuous RP with mean $\mu_X$

  To estimate the mean of $X(t)$, we form the time average

  \[
  \bar{X}(t) = \frac{1}{t} \int_0^t X(\tau) d\tau
  \]

  The RP $X(t)$ is said to be mean ergodic if $\bar{X}(t) \to \mu_X$ as $t \to \infty$ in m.s.

  Similarly for a discrete RP, the time average (same as sample average) is

  \[
  \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_n
  \]

  and the RP is mean ergodic if $\bar{X}_n \to \mu_X$ in m.s.

• Example: Let $X_n$ be a WSS process with $C_X(n) = 0$ for $n \neq 0$, i.e., the $X_i$s are uncorrelated, then $X_n$ is mean ergodic

• The process does not need to have uncorrelated samples for it to be mean ergodic, however

• Whether a WSS process is mean ergodic again depends only on its autocorrelation function

  By definition, mean ergodicity means that

  \[
  \lim_{t \to \infty} E[(\bar{X}(t) - \mu_X)^2] \to 0
  \]

  Since $E(\bar{X}(t)) = \mu_X$, the condition for mean ergodicity is the same as

  \[
  \lim_{t \to \infty} \text{Var}(\bar{X}(t)) = 0
  \]
Now, consider

\[ E(\bar{X}^2(t)) = E\left( \left( \frac{1}{t} \int_0^t X(\tau) d\tau \right)^2 \right) \]

\[ = E\left( \frac{1}{t^2} \int_0^t \int_0^t X(\tau_1)X(\tau_2) d\tau_1 d\tau_2 \right) \]

\[ = \frac{1}{t^2} \int_0^t \int_0^t R_X(\tau_1, \tau_2) d\tau_1 d\tau_2 \]

\[ = \frac{1}{t^2} \int_0^t \int_0^t R_X(\tau_1 - \tau_2) d\tau_1 d\tau_2 \]

From figure below, this double integral reduces to the single integral

\[ E(\bar{X}^2(t)) = \frac{2}{t^2} \int_0^t (t - \tau) R_X(\tau) d\tau \]

- Hence, a WSS process \( X(t) \) is mean ergodic iff

\[
\lim_{t \to \infty} \frac{2}{t^2} \int_0^t (t - \tau) R_X(\tau) d\tau = \mu_X^2
\]
• Example: Let \( X(t) \) be a WSS with zero mean and \( R_X(\tau) = e^{-|\tau|} \). Evaluating the condition on mean ergodicity, we obtain

\[
\frac{2}{t^2} \int_0^t (t - \tau) R_X(\tau) d\tau = \frac{2}{t^2} (e^{-t} + t - 1),
\]

which \( \to 0 \) as \( t \to \infty \). Hence \( X(t) \) is mean ergodic.

• Example: Consider the coin with random bias \( P \) example in Lecture Notes 5. The random process \( X_1, X_2, \ldots \) is stationary. However, it is not mean ergodic, since \( \bar{X}_n \to P \) in m.s.

• Remarks:
  ○ The process in the above example can be viewed as a mixture of IID Bernoulli(\( p \)) processes, each of which is stationary ergodic (it turns out that every stationary process is a mixture of stationary ergodic processes).
  ○ Ergodicity can be defined for general (not necessarily stationary) processes (this is beyond the scope of this course, however).
Lecture Notes 8
Random Processes in Linear Systems

- Linear System with Random process Input
- LTI System with WSS Process Input
- Process Linear Estimation
  - Infinite smoothing filter
  - Spectral Factorization
  - Wiener Filter

Linear System with Random Process Input

- Consider a linear system with (time-varying) impulse response \( h(t, t - \tau) \) driven by a random process input \( X(t) \)

\[
\begin{align*}
X(t) & \quad \quad \quad h(t, t - \tau) \quad \quad \quad Y(t) \\
\end{align*}
\]

- The output of the system is

\[
Y(t) = \int_{-\infty}^{\infty} h(t, t - \tau) X(\tau) \, d\tau
\]

- We wish to specify the output random process \( Y(t) \)
- It is difficult to obtain a complete specification of the output process in general
- Important special case: If \( X(t) \) is a GRP, the output process \( Y(t) \) is also a GRP (since the integral above can be approximated by a sum and thus the output process is obtained via a linear transformation of \( X(t) \))
We focus on finding the mean and autocorrelation functions of \( Y(t) \) in terms of the mean and autocorrelation functions of the input process \( X(t) \) and the impulse response of the system \( h(t, t - \tau) \).

We are also interested in finding the crosscorrelation function between \( X(t) \) and \( Y(t) \) defined as

\[
R_{XY}(t_1, t_2) = \mathbb{E}(X(t_1)Y(t_2))
\]

Note that unlike \( R_X(t_1, t_2) \), \( R_{XY}(t_1, t_2) \) is not necessarily symmetric in \( t_1 \) and \( t_2 \). However, \( R_{XY}(t_1, t_2) = R_{YX}(t_2, t_1) \).

To find the mean of \( Y(t) \), consider

\[
\mathbb{E}(Y(t)) = \mathbb{E}\left(\int_{-\infty}^{\infty} h(t, t - \tau)X(\tau) \ d\tau\right) = \int_{-\infty}^{\infty} h(t, t - \tau) \mathbb{E}(X(\tau)) \ d\tau
\]

The crosscorrelation function between \( Y(t) \) and \( X(t) \) is

\[
R_{YX}(t_1, t_2) = \mathbb{E}(Y(t_1)X(t_2))
\]

\[
= \mathbb{E}\left(\int_{-\infty}^{\infty} h(t_1, t_1 - \tau)X(\tau)X(t_2) \ d\tau\right)
\]

\[
= \int_{-\infty}^{\infty} h(t_1, t_1 - \tau)R_X(\tau, t_2) \ d\tau
\]

The average power is

\[
\mathbb{E}(Y^2(t)) = R_Y(t, t)
\]

Example (Integrator): Let \( X(t) \) be a white noise process with autocorrelation function \( R_X(\tau) = (N/2)\delta(\tau) \) and let the linear system be an ideal integrator, i.e.,

\[
Y(t) = \int_{0}^{t} X(\tau) \ d\tau
\]

Find the mean and autocorrelation functions and the average power of the integrator output \( Y(t) \), for \( t > 0 \).
This example is motivated by several applications:

- Noise in an image sensor pixel: the white noise models the photodetector shot noise, which is integrated with the signal over a capacitor before sampling
- Noise in a voltage controlled oscillator (for phase locked loops)

- **Solution:** The mean is

\[
E(Y(t)) = \int_0^t E(X(\tau)) \, d\tau = 0
\]

To obtain the autocorrelation function and average power for this case, we can specialize the previous results to

\[
R_{YX}(t_1, t_2) = \int_0^{t_1} \frac{N}{2} \delta(t_2 - \tau) \, d\tau
\]

\[
= \begin{cases} 
  \frac{N}{2}, & \text{for } t_2 \leq t_1 \\
  0, & \text{otherwise}
\end{cases}
\]

\[
E(Y^2(t)) = R_Y(t, t) = \frac{N}{2} t
\]

Note that the average power grows linearly with \( t \) (as for the random walk)

- If in addition \( X(t) \) is a GRP, then \( Y(t) \) is also a GRP and is referred to as the **Wiener process**
Consider a linear time invariant (LTI) system with real impulse response \( h(t) \) and transfer function \( H(f) = \mathcal{F}(h(t)) \), driven by WSS process \( X(t), -\infty < t < \infty \)

\[ X(t) \rightarrow h(t) \rightarrow Y(t) \]

We want to characterize its output \( Y(t) = X(t) * h(t) = \int_{-\infty}^{\infty} X(\tau)h(t-\tau) \, d\tau \)

It turns out (not surprisingly) that if the system is stable, i.e.,

\[ \left| \int_{-\infty}^{\infty} h(t) \, dt \right| = |H(0)| < \infty, \]

then \( X(t) \) and \( Y(t) \) are jointly WSS, which means that:

- \( X(t) \) and \( Y(t) \) are WSS, and
- Their crosscorrelation function \( R_{XY}(t_1, t_2) \) is time invariant, i.e.,

\[ R_{XY}(t_1, t_2) = E(\alpha \cos(\omega t + \Theta) \sin(\omega t + \Theta)) = R_{XY}(t_1 + \tau, t_2 + \tau) \quad \text{for all } \tau \]

Relabel \( R_{XY}(t_1, t_2) \) for jointly WSS \( X(t), Y(t) \) as \( R_{XY}(\tau) \), where \( \tau = t_1 - t_2 \)

\[ R_{XY}(\tau) = R_{XY}(t_2 + \tau, t_2) = R_{XY}(t_2 + (t_1 - t_2), t_2) = R_{XY}(t_1, t_2) \]

Again \( R_{XY}(\tau) \) is not necessarily even. However,

\[ R_{XY}(\tau) = R_{YX}(-\tau) \]

Example: Let \( \Theta \sim U[0, 2\pi] \). Consider two processes

\[ X(t) = \alpha \cos(\omega t + \Theta) \quad \text{and} \quad Y(t) = \alpha \sin(\omega t + \Theta) \]

These processes are jointly WSS, since each is WSS (in fact SSS) and

\[ R_{XY}(t_1, t_2) = E \left[ \alpha^2 \cos(\omega t_1 + \Theta) \sin(\omega t_2 + \Theta) \right] \]

\[ = \frac{\alpha^2}{4\pi} \int_{0}^{2\pi} \left[ \sin(\omega(t_1 + t_2) + 2\theta) - \sin(\omega(t_1 - t_2)) \right] d\theta \]

\[ = -\frac{\alpha^2}{2} \sin(\omega(t_1 - t_2)) \]

- We define the cross power spectral density for jointly WSS processes \( X(t), Y(t) \) as

\[ S_{XY}(f) = \mathcal{F}(R_{XY}(\tau)) \]
Example: Let $Y(t) = X(t) + Z(t)$, where $X(t)$ and $Z(t)$ are zero mean uncorrelated WSS processes. Show that $Y(t)$ and $X(t)$ are jointly WSS, and find $R_{XY}(\tau)$ (in terms of $R_X$ and $R_Z$) and $S_{XY}(f)$ (in terms of $S_X$ and $S_Z$)

Solution: First, we show that $Y(t)$ is WSS, since it is zero mean and

$$R_Y(t_1, t_2) = E[(X(t_1) + Z(t_1))(X(t_2) + Z(t_2))]$$

$$= E(X(t_1)X(t_2)) + E(Z(t_1)Z(t_2))$$

$(X(t), Z(t)$ zero mean, uncorrelated)$

$$= R_X(\tau) + R_Z(\tau)$

Taking the Fourier transform of both sides, $S_Y(f) = S_X(f) + S_Z(f)$

To show that $Y(t)$ and $X(t)$ are jointly WSS, we need to show that their crosscorrelation function is time invariant

$$R_{XY}(t_1, t_2) = E[X(t_1)(X(t_2) + Z(t_2))]$$

$$= E(X(t_1)X(t_2)) + E(X(t_1)Z(t_2))$$

$(X(t), Z(t)$ zero mean, uncorrelated)$

$$= R_X(t_1, t_2) + 0$$

Taking the Fourier transform, $S_{XY}(f) = S_X(f)$

---

Output Mean, Autocorrelation, and PSD

Theorem: Let $X(t), t \in \mathbb{R}$, be a WSS process input to a stable LTI system with real impulse response $h(t)$ and transfer function $H(f)$. Then the input $X(t)$ and output $Y(t)$ are jointly WSS with:

1. $E(Y(t)) = H(0)E(X(t))$
2. $R_{YX}(\tau) = h(\tau) \ast R_X(\tau)$
3. $R_Y(\tau) = h(\tau) \ast R_X(\tau) \ast h(-\tau)$

$$R_X(\tau) \xrightarrow{h(\tau)} R_{YX}(\tau) \xrightarrow{h(-\tau)} R_Y(\tau)$$

4. $S_{XY}(f) = H(f)S_X(f)$
5. $S_Y(f) = |H(f)|^2S_X(f)$

$$S_X(f) \xrightarrow{H(f)} S_{YX}(f) \xrightarrow{H(-f)} S_Y(f)$$
Remark: For a discrete time WSS process $X(n)$ and a stable LTI system $h(n)$, $X(n)$ and the output process $Y(n)$ are jointly WSS and we can similarly find $R_Y(n), \ldots$

Proof: Note that here the LTI system is in steady state

1. To find the mean of $Y(t)$, consider
   \[
   E(Y(t)) = E\left( \int_{-\infty}^{\infty} X(\tau) h(t-\tau) \, d\tau \right) \\
   = \int_{-\infty}^{\infty} E(X(\tau)) h(t-\tau) \, d\tau \\
   = E(X(t)) \int_{-\infty}^{\infty} h(t-\tau) \, d\tau = E(X(t)) H(0)
   \]

2. To find the crosscorrelation function between $Y(t)$ and $X(t)$, consider
   \[
   R_{YX}(\tau) = E\left( Y(t+\tau)X(t) \right) \\
   = E\left( \int_{-\infty}^{\infty} h(\alpha) X(t+\tau-\alpha) X(t) \, d\alpha \right)
   \]
   \[
   = \int_{-\infty}^{\infty} h(\alpha) R_X(\tau-\alpha) \, d\alpha \\
   = h(\tau) * R_X(\tau)
   \]

3. To find the autocorrelation function of $Y(t)$, consider
   \[
   R_Y(\tau) = E(Y(t+\tau)Y(t)) \\
   = E\left( Y(t+\tau) \int_{-\infty}^{\infty} h(\alpha) X(t-\alpha) \, d\alpha \right) \\
   = \int_{-\infty}^{\infty} h(\alpha) R_{YX}(\tau+\alpha) \, d\alpha \\
   = R_{YX}(\tau) * h(-\tau)
   \]

4. Follows by taking the Fourier transform of $R_{YX}(\tau)$

5. Follows by taking the Fourier transform of $R_Y(\tau)$
$S_X(f)$ is the Power Spectral Density

- We can use the above results to show that $S_X(f)$ is indeed the power spectral density of $X(t)$; i.e., the average power in any frequency band $[f_1, f_2]$ is

$$2 \int_{f_1}^{f_2} S_X(f) \, df$$

- To show this we pass $X(t)$ through an ideal band-pass filter

$$X(t) \rightarrow h(t) \rightarrow Y(t)$$

- Now the average power of $X(t)$ in the band $[f_1, f_2]$ is

$$E(Y^2(t)) = \int_{-\infty}^{\infty} S_Y(f) \, df$$

$$= \int_{-\infty}^{\infty} |H(f)|^2 S_X(f) \, df$$

$$= \int_{-f_2}^{-f_1} S_X(f) \, df + \int_{f_1}^{f_2} S_X(f) \, df$$

$$= 2 \int_{f_1}^{f_2} S_X(f) \, df$$

- This also shows that $S_X(f) \geq 0$ for all $f$
KT/C Noise

- The noise in a resistor $R$ (in ohms) due to thermal noise is modeled as a WGN voltage source $V(t)$ in series with $R$. The psd of $V(t)$ is $S_V(f) = 2kTR V^2/Hz$ for all $f$, where $k$ is Boltzmann’s constant and $T$ is the temperature in degrees K.

- Now let’s find the average output noise power for an RC circuit

First we find the transfer function for the circuit

$$H(f) = \frac{1}{1 + i2\pi f RC} \Rightarrow |H(f)|^2 = \frac{1}{1 + (2\pi f RC)^2}$$

Now we write the output psd in terms of the input psd as

$$S_{V_0} = S_V(f)|H(f)|^2 = 2kTR \frac{1}{1 + (2\pi f RC)^2}, -\infty < f < \infty$$

Thus the average output power is

$$E(V_0^2(t)) = \int_{-\infty}^{\infty} S_{V_0}(f) df = \frac{2kTR}{2\pi RC} \int_{-\infty}^{\infty} \frac{1}{1 + (2\pi f RC)^2} d(2\pi f RC)$$

$$= \frac{kT}{\pi C} \int_{-\infty}^{\infty} \frac{1}{1 + x^2} dx$$

$$= \frac{kT}{\pi C} \arctan x \bigg|_{-\infty}^{+\infty} = \frac{kT}{\pi C} \pi = \frac{kT}{C},$$

which is independent of $R$!
Autoregressive Moving Average Process

- Let \( X_n, -\infty < n < \infty \), be a discrete time white noise process with average power \( N \).

The autoregressive moving average (ARMA) process of order \((p, q)\), \( Y_n, -\infty < n < \infty \), is defined as

\[
Y_n = -\sum_{k=1}^{p} \alpha_k Y_{n-k} + \sum_{l=0}^{q} \beta_l X_{n-l}
\]

where \( \beta_0 = 1, \alpha_1, \ldots, \alpha_p \) and \( \beta_1, \ldots, \beta_q \) are fixed parameters.

- This process can be viewed as the output of an LTI system with transfer function

\[
H(f) = \frac{1 + \sum_{l=1}^{q} \beta_l e^{-i2\pi fl}}{1 + \sum_{k=1}^{p} \alpha_k e^{-i2\pi fk}}, \quad |f| < \frac{1}{2}
\]

Therefore, the PSD of \( Y_n \) is \( S_Y(f) = |H(f)|^2 N \) for \( |f| < 1/2 \).

- Moving average (MA) process of order \( q \): Let \( \alpha_1 = \cdots = \alpha_p = 0 \), then \( Y_n \) is simply a weighted sum of the \( q + 1 \) most recent \( X_n \) samples with weights \((1, \beta_1, \ldots, \beta_q)\), i.e.,

\[
Y_n = \sum_{l=0}^{q} \beta_l X_{n-l}, \quad \text{and the transfer function of the LTI system is}
\]

\[
H(f) = 1 + \sum_{l=1}^{q} \beta_l e^{-i2\pi fl}, \quad |f| < \frac{1}{2}
\]

- Communication channel with intersymbol interference: The \( X_n \) process represents the transmitted information symbols and \( 1, \beta_1, \ldots, \beta_q \) are the coefficients of the channel impulse response.

The process \( Y_n \) is the interference-impaired received symbols.

- Autoregressive (AR) process of order \( p \): Let \( \beta_1 = \cdots = \beta_q = 0 \). Then

\[
Y_n = -\sum_{k=1}^{p} \alpha_k Y_{n-k} + X_n, \quad \text{and the transfer function of the LTI system is}
\]

\[
H(f) = \frac{1}{1 + \sum_{k=1}^{p} \alpha_k e^{-i2\pi fk}}, \quad |f| < \frac{1}{2}
\]
Modeling the human speech generation process: The process $X_n$ is generated by the vocal cords. The vocal tract is modeled as a series of coupled lossless acoustic tubes parameterized by $(\alpha_1, \ldots, \alpha_p)$.

The process $Y_n$ is the uttered speech signal after it passes through the vocal tract.

For $p = 1$, we obtain the first-order autoregressive process

$$Y_n = -\alpha_1 Y_{n-1} + X_n,$$

$$H(f) = \frac{1}{1 + \alpha_1 e^{-i2\pi f}}, \quad |f| < \frac{1}{2}$$

$$h(n) = (-\alpha_1)^n u(n)$$

This transfer function is stable iff $\sum_{n=-\infty}^{\infty} |h(n)| < \infty$, i.e., iff $|\alpha_1| < 1$

If $X_n$ is Gaussian, we obtain a stationary version of the Gauss–Markov process discussed in Lecture Notes 6 with $\alpha = -\alpha_1$

---

**Sampling Theorem for Bandlimited WSS Processes**

- Recall the *Nyquist sampling theorem* for bandlimited deterministic signals:
  - Let $x(t)$ be a signal with Fourier transform $X(f)$ such that $X(f) = 0$ for $f \not\in [-B, B]$
  - We sample the signal at rate $1/T$ to obtain the sampled signal
    $$y_n = x(nT) \quad \text{for } n = \ldots, -2, -1, 0, 1, 2, \ldots$$
    The Fourier transform of the sequence $y_n$,
    $$Y(f) = \sum_{n=-\infty}^{\infty} X(f - n/T),$$
    is periodic with period $1/T$
  - To recover the signal, we pass $y_n$ through an ideal low pass filter of bandwidth $1/T$. The Fourier transform of the reconstructed signal is
    $$\hat{X}(f) = Y(f) \cdot \mathcal{I}(fT)$$
  - Hence if the sampling rate $1/T \geq 2B$, $\hat{X}(f) = X(f)$ and the signal can be reconstructed *perfectly* from its samples
It turns out that a similar result holds for sampling of bandlimited WSS random processes.

**Sampling theorem for WSS processes:**

- Let \( X(t) \) be a continuous time WSS process with zero mean and autocorrelation function \( R_X(\tau) \) and PSD \( S_X(f) = 0 \) for \( f \notin [-B, B] \)
- We sample \( X(t) \) at rate \( 1/T \) to obtain the sampled (discrete time) process \( Y_n = X(nT) \)

Hence \( Y_n \) is WSS with zero mean and autocorrelation function

\[
R_Y(n) = R_X(nT)
\]

The PSD of \( Y_n \),

\[
S_Y(f) = \sum_{n=-\infty}^{\infty} S_X(f-n/T),
\]

is periodic with period \( 1/T \)

- As for the deterministic signal case, to reconstruct the RP \( X(t) \), we pass \( Y_n \) through an ideal low pass filter.
  The resulting reconstruction process \( \hat{X}(t) \) is WSS with PSD

\[
S_{\hat{X}}(f) = |S_Y(f)| \cap |fT|^2
\]

- Hence if the sampling rate \( 1/T \geq 2B \), \( S_{\hat{X}}(f) = S_X(f) \)
- We show that this implies that the reconstruction process \( \hat{X}(t) = X(t) \) for every \( t \) with probability one. Consider

\[
\mathbb{E}[(X(t) - \hat{X}(t))^2] = 0 \quad \text{for every } t
\]

- Proof: We know that if \( 1/T \geq 2B \), \( S_{\hat{X}}(f) = S_X(f) \), which implies that 

\[
R_{\hat{X}}(\tau) = R_X(\tau)
\]

Moreover,

\[
R_{\hat{X}X}(\tau) = \text{sinc} \left( \frac{\tau}{T} \right) * R_X(\tau)
\]

Now, consider

\[
\mathbb{E}[(X(t) - \hat{X}(t))^2] = R_X(0) + R_{\hat{X}}(0) - 2R_{\hat{X}X}(0) = 2R_X(0) - 2R_X(0) = 0
\]

Hence, \( \hat{X}(t) = X(t) \) w.p.1 for every \( t \)
• Let $X(t)$ and $Y(t)$ be zero mean jointly WSS processes with known autocorrelation and crosscorrelation functions $R_X(\tau), R_Y(\tau),$ and $R_{XY}(\tau)$

• We observe the random process $Y(\alpha)$ for $t - a \leq \alpha \leq t + b \ ( -a \leq b)$ and wish to find the MMSE linear estimate of the signal $X(t)$, i.e., $\hat{X}(t)$ such that the MSE $= E \left[ (X(t) - \hat{X}(t))^2 \right]$ is minimized

• The linear estimate is of the form

\[
\hat{X}(t) = \int_{-b}^{a} h(\tau)Y(t - \tau) \, d\tau
\]

• By the orthogonality principle, the MMSE linear estimate must satisfy

\[(X(t) - \hat{X}(t)) \perp Y(t - \tau), \quad -b \leq \tau \leq a\]

or

\[E \left[ (X(t) - \hat{X}(t))Y(t - \tau) \right] = 0, \quad -b \leq \tau \leq a\]

Thus, for $-b \leq \tau \leq a$, we must have

\[
R_{XY}(\tau) = E \left[ X(t)Y(t - \tau) \right] = E \left[ \hat{X}(t)Y(t - \tau) \right]
\]

\[= E \left( \int_{-b}^{a} h(\alpha)Y(t - \alpha)Y(t - \tau) \, d\alpha \right)
\]

\[= \int_{-b}^{a} h(\alpha)R_Y(\tau - \alpha) \, d\alpha
\]

So, to find $h(\alpha)$ we need to solve an infinite set of integral equations

• Solving these equations analytically is not possible in general. However, it can be done for two important special cases:
  o **Infinite smoothing**: when $a, b \to \infty$
  o **Filtering**: when $a \to \infty$ and $b = 0$ (Wiener–Hopf equations)
Infinite Smoothing Filter

- When \( a, b \to \infty \), the integral equations for the MMSE linear estimate become
  \[
  R_{XY}(\tau) = \int_{-\infty}^{\infty} h(\alpha) R_Y(\tau - \alpha) d\alpha, \quad -\infty < \tau < +\infty
  \]
  In other words,
  \[
  R_{XY}(\tau) = h(\tau) * R_Y(\tau)
  \]

- The Fourier transform convolution theorem gives the transfer function for the optimal infinite smoothing filter:
  \[
  S_{XY}(f) = H(f)S_Y(f) \implies H(f) = \frac{S_{XY}(f)}{S_Y(f)}
  \]

- The minimum MSE is
  \[
  \text{MSE} = E[(X(t) - \hat{X}(t))^2]
  = E[(X(t) - \hat{X}(t))X(t)] - E[(X(t) - \hat{X}(t))\hat{X}(t)]
  = E[(X(t) - \hat{X}(t))X(t)] \quad \text{(by orthogonality)}
  = E[(X(t))^2] - E[X(t)\hat{X}(t)]
  \]
  To evaluate the second term, consider
  \[
  R_{X\hat{X}}(\tau) = E(X(t + \tau)\hat{X}(t))
  = E\left(X(t + \tau) \int_{-\infty}^{\infty} h(\alpha)Y(t - \alpha) d\alpha\right)
  = \int_{-\infty}^{\infty} h(\alpha) R_{XY}(\tau + \alpha) d\alpha = R_{XY}(\tau) * h(-\tau)
  \]
  Therefore
  \[
  E(X(t)\hat{X}(t)) = R_{X\hat{X}}(0) = \int_{-\infty}^{\infty} H(-f)S_{XY}(f) df = \int_{-\infty}^{\infty} \frac{|S_{XY}(f)|^2}{S_Y(f)} df,
  \]
and the minimum MSE is
\[
E [(X(t) - \hat{X}(t))^2] = E [(X(t)^2) - E (X(t))\hat{X}(t)]
\]
\[
= \int_{-\infty}^{\infty} S_X(f)\,df - \int_{-\infty}^{\infty} \frac{|S_{XY}(f)|^2}{S_Y(f)}\,df
\]
\[
= \int_{-\infty}^{\infty} \left( S_X(f) - \frac{|S_{XY}(f)|^2}{S_Y(f)} \right)\,df
\]

- Example (Additive White Noise Channel): Let \( X(t) \) and \( Z(t) \) be zero mean uncorrelated WSS processes with

\[
S_X(f) = \begin{cases} \frac{P}{2} & |f| \leq B \\ 0 & \text{otherwise} \end{cases}
\]

\[
S_Z(f) = \frac{N}{2} \quad \text{for all } f
\]

Here the signal \( X \) is bandlimited white noise, and \( Z \) is white noise.

Find the optimal infinite smoothing filter for estimating \( X(t) \) given

\[ Y(\tau) = X(\tau) + Z(\tau), \quad -\infty < \tau < +\infty \]

and the MSE for the estimate produced by this filter.

The power spectral densities of \( X \) and \( Z \) are shown below.
• The transfer function of the optimal infinite smoothing filter is given by

\[ H(f) = \frac{S_{XY}(f)}{S_Y(f)} \]

\[ = \frac{S_X(f)}{S_X(f) + S_Z(f)} \]

\[ = \begin{cases} \frac{P}{P+N} & |f| \leq B \\ 0 & \text{otherwise} \end{cases} \]

The MMSE is given by

\[ \text{MSE} = \int_{-\infty}^{\infty} S_X(f) \, df - \int_{-\infty}^{\infty} \frac{|S_{XY}(f)|^2}{S_Y(f)} \, df \]

\[ = \int_{-B}^{+B} \frac{P}{2} \, df - \int_{-B}^{+B} \frac{(P/2)^2}{P/2 + N/2} \, df \]

\[ = PB - \frac{P^2}{4} \left( \frac{P+N}{2} \right) 2B \]

\[ = \frac{NPB}{N+P} \]

---

**Spectral Factorization**

• It can be shown that the power spectral density \( S_X(f) \) of WSS process \( X(t) \) has a square root, i.e., a transfer function \( H(f) \) such that

\[ S_X(f) = H(f)H^*(f) = |H(f)|^2 \]

This is similar to the square root of a covariance (correlation) matrix for a random vector discussed in Lecture notes 4.

• As for the random vector case, the square root of a PSD, \( H(f) \), and its inverse \( 1/H(f) \) can be used for coloring and whitening of WSS processes, e.g.,

  ○ Coloring:

  \[
  X(t) \quad \xrightarrow{H(f)} \quad Y(t) \\
  S_X(f) = 1 \quad \quad \quad S_Y(f) = S(f)
  \]
Whitening:

\[ Y(t) \xrightarrow{1/H(f)} X(t) \]

\[ S_Y(f) = S(f) \quad S_X(f) = 1 \]

Here \( X(t) \) is the innovation process of \( Y(t) \)

- It turns out that under certain conditions, the PSD \( S(f) \) of a WSS process has a causal square root, that is, \( S^+(f) \) such that \( S(f) = S^+(f)S^-(f) \), where \( S^-(f) = (S^+(f))^* \) is an anticausal filter (note the similarity to the square root for correlation matrix via Cholesky decomposition)

- In particular, if \( S(f) \) is a rational PSD for a continuous time WSS process, i.e.,

\[ S(f) = c \frac{(2\pi if + a_1)(2\pi if + a_2)\ldots(2\pi if + a_m)}{(2\pi if + b_1)(2\pi if + b_2)\ldots(2\pi if + b_n)}, \]

then it can be factorized into product of causal and anticausal square roots

Proof: Since \( S(f) \) is real and nonnegative, \( S^*(f) = S(f) \), if the denominator has factor \((2\pi if + b)\), \( \Re(b) > 0 \), then it must have factor \((-2\pi if + b^*)\).

Similarly, if numerator has factor \((2\pi if + a)\), \( \Re(a) > 0 \), then it must have factor \((-2\pi if + a^*)\)

Then we can express any rational PSD as \( S(f) = S^+(f)S^-(f) \), where \( S^+(f) \) is a causal square root that consists of the \( f \) factors and \( S^-(f) \) is an anti-causal square root consisting of the \(-f \) factors

- Example: Consider the PSD

\[ S(f) = \frac{4\pi^2f^2 + 3}{4\pi^2f^2 + 1} \]

The causal square root of \( S(f) \) is

\[ S^+(f) = \frac{i2\pi f + \sqrt{3}}{i2\pi f + 1} \quad \text{and} \quad S^-(f) = \frac{-i2\pi f + \sqrt{3}}{-i2\pi f + 1} \]

Remark: For a discrete time WSS process a rational PSD is of the form

\[ S(f) = c \frac{(a_1 - e^{-i2\pi f})(a_1^* - e^{i2\pi f})\ldots(a_m - e^{-i2\pi f})(a_m^* - e^{i2\pi f})}{(b_1 - e^{-i2\pi f})(b_1^* - e^{i2\pi f})\ldots(b_m - e^{-i2\pi f})(b_m^* - e^{i2\pi f})} \]

and can be expressed also as \( S(f) = S^+(f)S^-(f) \) (the \( e^{-i2\pi f} \) terms are causal and the \( e^{i2\pi f} \) terms are noncausal)
• Example: Consider the PSD for a discrete time process

\[ S(f) = \frac{3}{5 - 4 \cos(2\pi f)} \]

The causal square root is

\[ S^+(f) = \frac{\sqrt{3}}{2 - e^{-i2\pi f}} \text{ and } S^-(f) = \frac{\sqrt{3}}{2 - e^{i2\pi f}} \]

**Spectral factorization theorem**: In general, a PSD \( S(f) \) has a causal square root if it satisfies the Paley-Wiener condition

\[ \int_{-\infty}^{\infty} \log \frac{S(f)}{1 + 4\pi^2 f} df > -\infty \text{ for continuous time process} \]

\[ \int_{-1/2}^{1/2} \log S(f) df > -\infty \text{ for discrete time process} \]

• These condition are not always satisfied. For example they are not satisfied for bandlimited processes

• Remark: We assume throughout that \( F^{-1}[S^+(f)] \) is real; hence \( S^-(f) = S^+(-f) \)

### Wiener Filter

• Again let \( X(t) \) and \( Y(t) \) be jointly WSS random processes. consider the linear estimation of process \( X(t) \) from observations \( Y(\alpha), t - a \leq \alpha \leq t + b \)

• When \( a \to \infty \) and \( b = 0 \), the equations for the MMSE linear estimate, called **Wiener–Hopf equations**, are

\[ R_{XY}(\tau) = \int_{-\infty}^{\infty} h(\alpha) R_Y(\tau - \alpha) d\alpha, \quad 0 \leq \tau < \infty \]

\[ = \int_{-\infty}^{\infty} h(\alpha) R_Y(\tau - \alpha) d\alpha, \quad 0 \leq \tau < \infty \]

where \( h(t) \) is a **causal** impulse response

• Notation: A real-valued function \( h(t) \) can be expressed as

\[ h(t) = [h(t)]_+ + [h(t)]_- \]

where \([h(t)]_+ = h(t)\) for \( t \geq 0 \) and \([h(t)]_- = 0\) for \( t < 0 \) is the **positive** (causal) part of \( h(t) \), and \([h(t)]_- = h(t) - [h(t)]_+\) is the **negative** (anticausal) part
Taking the Fourier transform, we have
\[ H(f) = [H(f)]_+ + [H(f)]_-, \]
where \([H(f)]_+\) and \([H(f)]_-\), are the FT of the positive and negative parts of \(h(t)\), respectively.

Example: Let
\[ S(f) = \frac{4\pi^2 f^2 + 3}{4\pi^2 f^2 + 1} \]
We can write
\[ S(f) = \frac{i2\pi f + 2}{i2\pi f + 1} + \frac{1}{-i2\pi f + 1} \]
The first term is \([S(f)]_+\) and the second is \([S(f)]_-\). The corresponding impulse responses are
\[ [R(t)]_+ = \delta(t) + e^{-t}u(t) \]
\[ [R(t)]_- = e^{t}u(-t) \]

Compare to the causal square root factors

- Now, back to the linear estimation problem. First assume that the observation process \(Y(\tau)\) is white, i.e., \(R_Y(\tau) = \delta(\tau)\), then the Wiener–Hopf equations reduce to
  \[ R_{XY}(\tau) = h(\tau), \quad 0 \leq \tau < \infty, \]
i.e., \(h(\tau) = [R_{XY}(\tau)]_+\)
and the corresponding transfer function is
\[ H(f) = \int_0^\infty R_{XY}(\tau)e^{-2\pi if\tau} d\tau, \]
i.e., \(H(f) = [S_{XY}(f)]_+\)
- For general \(S_Y(f)\) with causal square root \(S_Y^+(f)\), we first whiten the process to obtain \(\tilde{Y}(\tau)\) with \(R_{\tilde{Y}}(\tau) = \delta(\tau)\), then convolve with \([R_{X\tilde{Y}}(\tau)]_+\)

\[
\begin{array}{c}
Y(t) \quad \frac{1}{S_Y^+(f)} \quad \tilde{Y}(t) \quad [S_{X\tilde{Y}}(f)]_+ \quad \tilde{X}(t)
\end{array}
\]
Now to find $R_{XY}(\tau)$, let $g(t) = \mathcal{F}^{-1}[1/S^+_Y(f)]$ and consider

\[
R_{XY}(\tau) = E(X(t + \tau)\tilde{Y}(\tau)) \\
= E(X(t + \tau)Y(t) \ast g(t)) \\
= R_{XY}(\tau) \ast g(-\tau)
\]

Taking the Fourier Transform we have

\[
S_{XY}(f) = \frac{S_{XY}(f)}{S_Y(f)}
\]

Hence,

\[
[S_{XY}(f)]_+ = \left[\frac{S_{XY}(f)}{S_Y(f)}\right]_+
\]

The transfer function of the Wiener filter is then given by

\[
H(f) = \frac{1}{S^+_Y(f)} \left[\frac{S_{XY}(f)}{S_Y(f)}\right]_+
\]

To find the MMSE, we follow similar steps to the infinite smoothing case to obtain

\[
\text{MSE} = \int_{-\infty}^{\infty} \left( S_X(f) - \left|\left[\frac{S_{XY}(f)}{S_Y(f)}\right]_+\right|^2 \right) df
\]

Example: Consider a continuous-time RP $X(t)$ with

\[
S_X(f) = \frac{2}{1 + 4\pi^2 f^2}
\]

and the noisy observation $Y(t) = X(t) + Z(t)$, where $Z(t)$ is white noise uncorrelated with $X(t)$ with $S_Z(f) = 1$

To compute the Wiener filter, we first factor the PSD

\[
S_Y(f) = S_X(f) + S_Z(f) = \frac{4\pi^2 f^2 + 3}{4\pi^2 f^2 + 1}
\]

to obtain

\[
S^+_Y(f) = \frac{i2\pi f + \sqrt{3}}{i2\pi f + 1}, \\
S^-_Y(f) = -\frac{i2\pi f + \sqrt{3}}{-i2\pi f + 1}
\]
The crosspower spectral density $S_{XY}(f) = S_X(f)$, hence
\[
\frac{S_{XY}(f)}{S_Y(f)} = \frac{2}{1 + 4\pi^2 f^2} \cdot \frac{-i2\pi f + 1}{-i2\pi f + \sqrt{3}}
\]
\[
= \frac{2}{(i2\pi f + 1)(-i2\pi f + \sqrt{3})}
\]
\[
= \frac{\sqrt{3} - 1}{i2\pi f + 1} + \frac{\sqrt{3} - 1}{-i2\pi f + \sqrt{3}}
\]

The first term is causal and the second term is anticausal. Therefore,
\[
\left[ \frac{S_{XY}(f)}{S_Y(f)} \right]_+ = \frac{\sqrt{3} - 1}{i2\pi f + 1}
\]

Hence, the Wiener filter is
\[
H(f) = \frac{\sqrt{3} - 1}{i2\pi f + \sqrt{3}}
\]
\[
h(t) = (\sqrt{3} - 1) e^{-\sqrt{3}t} u(t)
\]

The MSE is
\[
\text{MSE} = \int_{-\infty}^{\infty} \frac{2\sqrt{3} - 2}{1 + 4\pi^2 f^2} \, df = \sqrt{3} - 1
\]

- Example: Consider a discrete-time RP $X(t)$ with
\[
S_X(f) = \frac{3}{5 - 4 \cos(2\pi f)}
\]
and the noisy observation $Y(t) = X(t) + Z(t)$, where $Z(t)$ is white noise, independent of $X(t)$, with $S_Z(f) = 1$.

Again we factor the PSD
\[
S_Y(f) = S_X(f) + S_Z(f) = \frac{8 - 4 \cos(2\pi f)}{5 - 4 \cos(2\pi f)}
\]
to obtain
\[
S_Y^+(f) = \sqrt{4 - 2\sqrt{3}} \cdot \frac{2 + \sqrt{3} - e^{-i2\pi f}}{2 - e^{-i2\pi f}}
\]
\[
S_Y^-(f) = \sqrt{4 - 2\sqrt{3}} \cdot \frac{2 + \sqrt{3} - i2\pi f}{2 - e^{i2\pi f}}
\]
The crosspower spectral density is $S_{XY}(f) = S_X(f)$ and
\[
\frac{S_{XY}(f)}{S_Y(f)} = \frac{3}{\sqrt{4 - 2\sqrt{3}}} \cdot \frac{1}{(2 - e^{-i2\pi f})(2 + \sqrt{3} - e^{-i2\pi f})} = \frac{\sqrt{12 - 6\sqrt{3}}}{2 - e^{-i2\pi f}} + \frac{\sqrt{3 - 3\sqrt{3}/2}}{2 + \sqrt{3} - e^{i2\pi f}}
\]
The first term is causal and the second term is anticausal. Therefore
\[
\left[\frac{S_{XY}(f)}{S_Y(f)}\right] = \frac{\sqrt{12 - 6\sqrt{3}}}{2 - e^{-i2\pi f}}
\]
Hence, the Wiener filter is
\[
H(f) = \frac{\sqrt{3}}{2 + \sqrt{3} + e^{-i2\pi f}}
\]
\[
h(n) = (2\sqrt{3} - 3) (2 - \sqrt{3})^n u(n)
\]

**Wiener Filter Versus Kalman Filter**

- Both the Kalman and Wiener filters are MMSE *linear* estimates of a process from causal observations.
- There are several differences, however:

<table>
<thead>
<tr>
<th>Kalman filter</th>
<th>Wiener filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_n$, $Y_n$ state space model</td>
<td>$X_n$, $Y_n$ jointly WSS</td>
</tr>
<tr>
<td>not necessarily WSS</td>
<td>frequency domain filter</td>
</tr>
<tr>
<td>time domain filter</td>
<td>non-recursive</td>
</tr>
<tr>
<td>recursive</td>
<td></td>
</tr>
</tbody>
</table>

- Remarks:
  - A continuous time counterpart to the Kalman filter exists and is known as the *Kalman-Bucy* filter
  - If $X_n, Y_n$ are jointly WSS and can be described via a state space model, then the Kalman filter gives a recursive way to compute the Wiener filter