Review of Probability Basics

Terminology and Notation

"Universal outcome set" $S$: The set of all possible outcomes of a specified random trial ("experiment")

"Outcome": An element of the set $S$.

$x \in A$ "$x$ is an element of the set $A$"

$A \subseteq B$ "$A$ is a subset of $B$"

$A \cup B = \{ x | x \in A \text{ or } x \in B \}$ "the union of $A$ and $B$"

$A \cap B = \{ x | x \in A \text{ and } x \in B \}$ "the intersection of $A$ and $B$"

$A - B = \{ x | x \in A \text{ but } x \notin B \}$ "the set difference $A$ minus $B$"

$\emptyset$ "The empty set"

$\overline{A} = S - A$ "the complement of $A$"

"Events": A collection $\mathcal{F}$ of subsets of $S$ that satisfies the sigma field (sigma-field) conditions:

(i) $\emptyset \in \mathcal{F}$

(ii) If $A \in \mathcal{F}$, then $\overline{A} \in \mathcal{F}$.

(iii) If $A_1, A_2, \ldots \in \mathcal{F}$, then $A_1 \cup A_2 \cup \ldots \in \mathcal{F}$.
Def: Events $A_1, A_2, \ldots$ are exhaustive if $A_1 \cup A_2 \cup \ldots = S$. They are mutually exclusive if $A_i \cap A_j = \emptyset$ for $i \neq j$.

**Probability Measure**

A probability measure assigns a real number to each event $i.e.$ $P: \mathcal{F} \rightarrow \mathbb{R}$. It must satisfy:

(a) $P(A) \geq 0$ for all events $A$
(b) $P(S) = 1$
(c) If $A_1, A_2, \ldots$ are mutually exclusive events, $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$

**Generation of $\sigma$-Fields**

Beginning with a collection $\mathcal{B}$ of subsets of $S$, with $S \in \mathcal{B}$, form every set possible by applying $\sigma$-field properties (ii) and (iii) to the original collection $\mathcal{B}$. The new collection $\mathcal{F}_\mathcal{B}$ will be a $\sigma$-field and is said to be "generated by $\mathcal{B}$."
If $S$ is a topological space (i.e., with open sets), the Borel $\sigma$-field is the $\sigma$-field generated by the open sets of $S$.

**Conditional Probability**

The conditional probability of event $A$ given event $B$ is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

**Bayes' Rule**

From the definition, $P(A \cap B) = P(A|B)P(B)$. But also $P(A \cap B) = P(B|A)P(A)$. So

$$P(B|A)P(A) = P(A|B)P(B)$$

or

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

**Statistical Independence**

Events $A$ and $B$ are independent if

$$P(A \cap B) = P(A)P(B)$$
\[ P(A | B) = P(A) \]

Independent events satisfy

\[ \frac{P(A \cap B)}{P(B)} = P(A) , \]

so

\[ P(A \cap B) = P(A)P(B) \]

**Total Probability Principle**

Let \( A_1, \ldots, A_n \) be mutually exclusive and exhaustive events in \( S \). If \( B \) is any event, then \((B \cap A_1), \ldots, (B \cap A_n)\) are mutually exclusive events with

\[ B = (B \cap A_1) \cup \ldots \cup (B \cap A_n) \]

Thus

\[ P(B) = P(B \cap A_1) + \ldots + P(B \cap A_n) , \]

or

\[ P(B) = P(B | A_1)P(A_1) + \ldots + P(B | A_n)P(A_n) \]
Probability of Union

Note that $A = (A \setminus B) \cup (A \cap B)$ with $A \setminus B$ and $A \cap B$ mutually exclusive. So
$P(A \setminus B) = P(A) - P(A \cap B)$. Similarly,
$P(B \setminus A) = P(B) - P(A \cap B)$. Finally,

$$A \cup B = (A \setminus B) \cup (B \setminus A) \cup (A \cap B)$$

where the three events on the right-hand side are mutually exclusive. Thus

$$P(A \cup B) = P(A \setminus B) + P(B \setminus A) + P(A \cap B)$$

$$= [P(A) - P(A \cap B)] + [P(B) - P(A \cap B)] + P(A \cap B)$$

$$= P(A) + P(B) - P(A \cap B)$$
Review of One RV

Random Variable

A (real) random variable (RV) is a function $x : S \to \mathbb{R}$ where $S$ is an outcome set with probability measure $P$. A RV must satisfy the "measurability condition": $x^{-1}(A)$ must be an event in $S$ for every Borel set $A$ in $\mathbb{R}$.

Note: The measurability condition is satisfied by all RVs typically encountered in engineering.

Cumulative Distribution

The cumulative distribution function (CDF) of a RV $x$ is defined by

$$F_x(x) = P\{s \in S | x(s) \leq x\}$$

The CDF obviously satisfies:

1. (Monotonicity) If $x' > x$, $F_x(x') > F_x(x)$
2. (Lower bound) $F_x(x) \geq 0$ and $\lim_{x \to -\infty} F_x(x) = 0$
3. (Upper bound) $F_x(x) \leq 1$ and $\lim_{x \to \infty} F_x(x) = 1$
Probability Density

The probability density function (PDF) of a RV $X$ is defined by

$$f_X(x) = \frac{d}{dx} F_X(x)$$

Existence of the PDF requires that the CDF is differentiable in some sense. Formally, the probability measure $P$ on $S$ must have a Radon-Nikodym derivative with respect to Lebesgue measure on $\mathbb{R}$. Typically in engineering, $\frac{d}{dx}$ is an ordinary derivative or a derivative that accommodates jump discontinuities in $F_X$ by allowing $f_X$ to contain Dirac delta symbols.

The PDF satisfies:

(a) (Positivity) $f_X(x) \geq 0$
(b) (Unit integral) $\int_{-\infty}^{\infty} f_X(x) \, dx = 1$

Further, the CDF is recovered from the PDF by

$$F_X(x) = \int_{-\infty}^{x} f_X(u) \, du$$
Probability \( x \) Takes a Value in a Set

From the definition of CDF,

\[
P(\{ x \mid a \leq x \leq b \}) = F_x(b) - F_x(a) = \int_a^b f_x(x) \, dx
\]

More generally, for \( A \) a Borel set in \( \mathbb{R} \),

\[
P(\{ x \in A \}) = \int_A f_x(x) \, dx
\]

**Mean**

The mean or expected value of \( x \) is

\[
\mu_x = E[x] = \int_{-\infty}^{\infty} x f_x(x) \, dx
\]

Note that the mean of \( x \) may not exist.

**Expectation Rule**

If \( y = g(x) \), then

\[
\mu_y = E[g(x)] = \int_{-\infty}^{\infty} g(x) f_x(x) \, dx
\]
This important theorem will be proven later in this class for well-behaved transformations \( g \). In particular, it implies that \( E[\cdot] \) is linear in the sense that

\[
E[ax + b] = \int_{-\infty}^{\infty} (ax + b) f_x(x) \, dx
= a \int_{-\infty}^{\infty} x f_x(x) \, dx + b \int_{-\infty}^{\infty} f_x(x) \, dx = aE[x] + b
\]

**Variance**

The variance of \( x \) is \( \text{var}(x) = E[(x - \mu_x)^2] \).

The expectation rule gives

\[
\text{var}(x) = E\left[ x^2 - 2\mu_x x + \mu_x^2 \right]
= E[x^2] - 2\mu_x E[x] + \mu_x^2
= E[x^2] - \mu_x^2
\]

**Moments**

For \( k = 1, 2, \ldots \), the \( k^{th} \) moment of \( x \) is
\[ E[x^k] \]. The \( k^{th} \) central moment of \( x \) is \( E[(x - \mu_x)^k] \). Note that the first moment of \( x \) is its mean and the second central moment of \( x \) is its variance.

**Characteristic Function**

The characteristic function of \( x \) is

\[
\Phi_X(\omega) = E[e^{-i\omega x}] = \int_{-\infty}^{\infty} e^{-i\omega x} f_x(x) \, dx
\]

i.e., it is the Fourier transform of the PDF. Note that

\[
\Phi_X(0) = \int_{-\infty}^{\infty} f_x(x) \, dx = 1
\]

and, by inverse Fourier transform,

\[
f_x(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi_X(\omega) e^{i\omega x} \, d\omega.
\]

Further,

\[
\frac{d}{d\omega} \Phi_X(\omega) = \frac{d}{d\omega} \int_{-\infty}^{\infty} e^{-i\omega x} f_x(x) \, dx = -i \int_{-\infty}^{\infty} x e^{-i\omega x} f_x(x) \, dx
\]

so that

\[
\frac{d}{d\omega} \Phi_X(0) = -i E[x]
\]

Similarly,

\[
E[x^k] = \left(\frac{-i}{\omega}\right)^k \left\{ \begin{array}{c} d \left[ \Phi_X(\omega) \right] \\
\frac{d}{d\omega} \Phi_X(\omega) \end{array} \right\}_{\omega=0}, \quad k=1, 2, \ldots
\]
Moments

Mean

The first moment (mean) of a RV,

$$\mu_x = E[x] = \int_{-\infty}^{\infty} x f_x(x) \, dx$$

may be interpreted physically as the center of mass of the PDF $f_x$. In other words, if one had a long bar of material whose density as a function of position $x$ along the bar is given by $f_x(x)$, then $\mu_x$ is the point at which the bar would balance on a fulcrum.

Variance

The second central moment (variance) of $x$,

$$\sigma_x^2 = \text{var}(x) = E[(x - \mu_x)^2] = E[x^2] - \mu_x^2$$

can be interpreted as the moment of inertia around the mean.
Skewness

The skewness of a RV $x$ with variance $\sigma_x^2$ and mean $\mu_x$ is

$$\gamma_x = \frac{E[(x-\mu_x)^3]}{\sigma_x^3}$$

Because

$$(x-\mu_x)^3 = x^3 - 3x^2\mu_x + 3x\mu_x^2 - \mu_x^3,$$

$$\gamma_x = \frac{E[x^3] - 3\mu_x E[x^2] + 2\mu_x^3}{\sigma_x^3}$$

Roughly speaking, skewness measures the asymmetry of a PDF around its mean.

- Positive skew
- Negative skew

Kurtosis

The (excess) kurtosis of a RV $x$
with mean $\mu_x$ and variance $\sigma_x^2$ is

$$K_x = \frac{E[(x-\mu_x)^4]}{\sigma_x^4} - 3$$

$$= E[x^4] - 4\mu_x E[x^3] + 6\mu_x^2 E[x^2] - 3\mu_x^4 - 3$$

**Markov and Chebyshev Inequalities**

Then (Markov inequality) Suppose a RV $x$ is non-negative. Then

$$P(\{x \geq a\}) \leq \frac{E[x]}{a}$$

**Proof:**

$$E[x] = \int_{-\infty}^{\infty} x f_x(x) \, dx = \int_{-\infty}^{a} x f_x(x) \, dx + \int_{a}^{\infty} x f_x(x) \, dx$$

$$\geq \int_{a}^{\infty} a f_x(x) \, dx = a P(\{x \geq a\})$$

Then (Chebyshev inequality): If a RV $x$ has mean $\mu_x$ and variance $\sigma_x^2$, then

$$P(\{|x - \mu_x| \geq \varepsilon\}) \leq \frac{\sigma_x^2}{\varepsilon^2}$$
Prop: \( \sigma_x^2 = \int_{-\infty}^{\infty} (x - \mu_x)^2 f_x(x) \, dx \leq \int_{|x - \mu_x| > \epsilon} (x - \mu_x)^2 f_x(x) \, dx \)

\[ \leq \int_{|x - \mu_x| > \epsilon} \epsilon^2 f_x(x) \, dx = \epsilon^2 \mathbb{P}(|x - \mu_x| > \epsilon) \]

**Examples of Common PDFs**

**Uniform RV, \( X \sim U[a, b] \)**

Parameters: \( a < b \)

\( f_x(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{else} \end{cases} \)

\[
E[X] = \int_a^b x \left( \frac{1}{b-a} \right) \, dx = \frac{b^2 - a^2}{2(b-a)} = \frac{a + b}{2}
\]

\[
E[X^2] = \int_a^b x^2 \left( \frac{1}{b-a} \right) \, dx = \frac{b^3 - a^3}{3(b-a)} = \frac{b^2 + ab + a^2}{3}
\]

\[ \text{var}(X) = \frac{(b-a)^2}{12} \]
Normal (Gaussian) RV: \( x \sim N(\mu, \sigma^2) \)

Parameters: \( \mu \in \mathbb{R}, \sigma^2 > 0 \)

\[
f_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)
\]

Note: Define \( I = \int_{-\infty}^{\infty} e^{-ax^2} \, dx \). Then with \( a > 0 \),

\[
I^2 = \int_{-\infty}^{\infty} e^{-ax^2} \, dx \cdot \int_{-\infty}^{\infty} e^{-ay^2} \, dy = \int_{0}^{\infty} \int_{0}^{\infty} e^{-ar^2} \, rdr \, d\theta
\]

\[
= \frac{\pi}{2} \left( -\frac{e^{-a\theta^2}}{2a}\right)_{\theta=0}^{\theta=\infty} = \frac{\pi}{4a} \implies I = \sqrt{\frac{\pi}{a}}
\]

In particular, \( \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \, dx = \sqrt{2\pi\sigma^2} \)

so that \( \int_{-\infty}^{\infty} f_x(x) \, dx = 1 \).

Mean: Direct integration gives \( E[x] = \mu \)

Variance: Integration by parts gives \( \text{var}(x) = \sigma^2 \)
Exponential RV

Parameter: \( \lambda > 0 \)

\[
f_X(x) = \begin{cases} 
\lambda e^{-\lambda x} & x > 0 \\
0 & x < 0 
\end{cases}
\]

Characteristic function

\[
\tilde{F}_X(\omega) = \int_{-\infty}^{\infty} f_X(x)e^{-i\omega x} \, dx = \int_{0}^{\infty} \lambda e^{-\lambda x} \, dx
\]

\[
= \frac{\lambda}{\lambda + i\omega}
\]

\[
\tilde{F}_X(0) = 1 = \int_{-\infty}^{\infty} f_X(x) \, dx
\]

Mean

\[
\tilde{F}_X'(\omega) = \frac{-i\lambda}{(\lambda + i\omega)^2} \Rightarrow E[X] = \frac{\tilde{F}_X'(0)}{-i} = \frac{1}{\lambda}
\]

Variance

\[
\tilde{F}_X''(\omega) = \frac{2(-i)^2 \lambda}{(\lambda + i\omega)^3} \Rightarrow E[X^2] = \frac{\tilde{F}_X''(0)}{(-1)^2} = \frac{2}{\lambda^2}
\]

\[
\therefore \text{var}(X) = \frac{1}{\lambda^2}
\]
Higher Moments

\[ \Phi^{(n)}_{X}(w) = \frac{n! (-i)^n \Lambda}{(\Lambda + i\omega)^{n+1}} \Rightarrow E[x^n] = \frac{n!}{\Lambda^n} \]

\[ \gamma_{X} = \Lambda^3 \left[ \frac{3}{\Lambda^3} - \frac{3}{\Lambda} \frac{1}{\Lambda^2} + \frac{2}{\Lambda^3} \right] = 2 \]

\[ K_X = 6 \]
Exponential RV (cont'd)

Note that

\[ P \left\{ \left( \frac{x^2}{\mu_x} \right)^2 = \lambda \int_{\lambda}^\infty e^{-\lambda x} \, dx = e^{-1} \approx 0.37 \right\} \]

Example (Markov Inequality): With \( x \) exponential with \( \lambda = 1 \),

\[ P \left\{ x \geq \frac{3}{2} \right\} = \int_{\frac{3}{2}}^\infty e^{-x} \, dx = e^{-3/2} \approx 0.22 \]

Using Markov's inequality,

\[ P \left\{ x \geq \frac{3}{2} \right\} \leq \frac{E[x^2]}{\left( \frac{1}{2} \right)^2} = \frac{2}{3} \approx 0.67 \]

Gamma RV: \( x \sim \Gamma(\alpha, \beta) \)

Parameters: \( \alpha > 0 \) "shape", \( \beta > 0 \) "rate"

PDF:

\[ f_X(x) = \begin{cases} \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} e^{-x/\beta} & x > 0 \\ 0 & x < 0 \end{cases} \]
In this expression, $\Gamma$ denotes the gamma function, defined for any complex number $t$ that is not a non-positive integer by

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} \, dx$$

Note that

$$\Gamma(1) = \int_0^\infty x^0 e^{-x} \, dx = 1$$

and for $k = 1, 2, 3, \ldots$,

$$\Gamma(k+1) = \int_0^\infty x^k e^{-x} \, dx = -x^k e^{-x}\bigg|_0^\infty + \int_0^\infty k x^{k-1} e^{-x} \, dx$$

$$= k \Gamma(k)$$

So, $\Gamma(2) = \Gamma(1+1) = 1$, $\Gamma(1) = 1$, $\Gamma(3) = 2 \Gamma(2) = 2$, ...

$\Gamma(n) = (n-1)!$.

Note that

$$\int_0^\infty f(x) \, dx = \frac{1}{\Gamma(a)} \int_0^\infty \beta^{\alpha-1} x^{\alpha-1} e^{-\beta x} \, dx \quad \{u = \beta x\}$$

$$= \frac{1}{\Gamma(a)} \int_0^\infty u^{\alpha-1} e^{-u} \, du = \frac{\Gamma(a)}{\Gamma(a)} = 1.$$
Similar calculations yield

\[ E[x^2] = \frac{\alpha^2}{\beta} \quad \text{var}(x) = \frac{\alpha}{\beta^2} \]

**Gaussian CDF**

Let \( x \sim N[0,1] \). Then

\[ F_x(u) = \int_{-\infty}^{u} f_x(x) \, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{x^2}{2}} \, dx = \Phi(u) \]

No closed-form expression for \( \Phi(u) \) exists, but it can be evaluated numerically and is tabulated in many references. One often sees

\[ Q(u) = 1 - \Phi(u) \]

If \( x \sim N[\mu, \sigma^2] \),

\[ P\{x \leq u\} = \frac{1}{\sqrt{2\pi} \sigma^2} \int_{-\infty}^{\frac{u - \mu}{\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx \quad \left\{ \begin{array}{c}
\mathcal{N} = \frac{x-\mu}{\sigma} \\
\frac{d\mathcal{N}}{d\mathcal{V}} = \frac{d\mathcal{V}}{d\mathcal{X}}
\end{array} \right. \]

\[ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\frac{u-\mu}{\sigma}} e^{-\frac{v^2}{2}} \, dv = \Phi\left( \frac{u-\mu}{\sigma} \right) \]
Example: $x \sim N[2, 9]$

$$P(\{|x-2| \geq 1\})$$

$$= 2\Phi\left(\frac{1-2}{3}\right) = 2\Phi\left(-\frac{1}{3}\right)$$
Bernoulli Trials

A Bernoulli trial is a random experiment with exactly two possible outcomes, e.g.,

\[ S = \{0, 1\} \quad S = \{\text{head, tail}\} \]
\[ S = \{\text{success, failure}\} \]

It is completely specified by the probability of one outcome, say \( P(\text{success}) = p \). The other outcome then has probability \( 1-p \).

An n-fold Bernoulli trial consists of a sequence of n independent Bernoulli trials. The outcomes are ordered n-tuples of individual outcomes; e.g., with \( n=3 \),

\[ S = \{(000), (001), (010), (011), (100), (101), (110), (111)\} \]

Note that \( S \) contains \( 2^n \) outcomes. Of these

\[ 1 = \frac{n!}{0!} \quad \text{contains } n \text{ ones} \]
\[ n = \frac{n!}{1!} \quad \text{contain } n-1 \text{ ones} \]
\[ \binom{n}{k} = \frac{n!}{(n-k)!k!} \quad \text{contain } n-k \text{ ones} \]
The quantity
\[
\binom{n}{k} = \frac{n!}{(n-k)!k!}, \quad 0 \leq k \leq n
\]

is called a binomial coefficient because
\[
(a + b)^n = \binom{n}{0}a^n b^0 + \binom{n}{1}a^{n-1}b + \ldots + \binom{n}{n}a^0 b^n
\]

Taking \(a = b = 1\) gives
\[
2^n = \sum_{k=0}^{n} \binom{n}{k}
\]

These numbers also appear in Pascal's triangle:

\[
\begin{array}{cccccccc}
  & & & & & & 1 & \\
  & & & & 1 & 1 & \\
  & & 1 & 2 & 1 & \\
  & 1 & 3 & 3 & 1 & \\
\end{array}
\]

If the probability of obtaining a 1 in a single trial is \(p\), the probability of obtaining exactly \(k\) ones in \(n\) trials is
\[
\binom{n}{k} p^k (1-p)^{n-k}
\]
The probability of obtaining at least \( k \) ones in \( n \) trials is thus
\[
\sum_{m=k}^{n} \binom{n}{m} p^m (1-p)^{n-m}
\]

**Binary Hypothesis Testing**

Consider two hypotheses \( H_0 \) and \( H_1 \) that are mutually exclusive and exhaustive; \( P(H_0) + P(H_1) = 1 \). These typically represent alternative states:

- \( H_1 = \{ \text{signal present} \} \)
- \( H_0 = \{ \text{signal absent} \} \)
- \( H_1 = \{ \text{transmitted bit = 1} \} \)
- \( H_0 = \{ \text{bit = 0} \} \)

etc.

\( M \) is an event representing an observed measurement or collected data.

The decisions are

- \( \rightarrow H_1 = \{ \text{decide in favor of } H_1 \} \)
- \( \rightarrow H_0 = \{ \text{decide in favor of } H_0 \} \)
The Bayesian Decision Rule decides in favor of $H_1$ when its conditional probability given the data $M$ is greater than the conditional probability of $H_0$ given $M$, i.e.,

$$\rightarrow H_1 = \{ P(H_1|M) > P(H_0|M) \}$$

Applying Bayes' rule,

$$\rightarrow H_1 = \left\{ \frac{P(M|H_1)P(H_1)}{P(M)} > \frac{P(M|H_0)P(H_0)}{P(M)} \right\}$$

$$= \left\{ \frac{P(M|H_1)}{P(M|H_0)} > \frac{P(H_0)}{P(H_1)} \right\}$$

"likelihood ratio"

"priors"

Consider the case where the data consists of realization of one RV $x$:

$$M = \{ \text{$x$ takes on a value near $x$} \}$$

$$= \{ x \leq x < x+4 \}$$

Under $H_1$, $x$ has conditional PDF $f_{X}(x|H_1)$.
and

\[ P(M|H_i) = \int_{x}^{x+\Delta} f_x(x|H_i) \, dx \]

For small \( \Delta \),

\[ P(M|H_i) \approx \Delta f_x(x|H_i) \]

Similarly, \( P(M|H_0) \approx \Delta f_x(x|H_0) \) and the BDR becomes

\[ \rightarrow H_1 = \left\{ \frac{f_x(x|H_i)}{f_x(x|H_0)} > \frac{P(H_0)}{P(H_1)} \right\} \]

**Example:**

\[ f_x(x|H_i) = \begin{cases} 4x^2 & 0 \leq x \leq 1 \\ 0 & \text{else} \end{cases} \]

\[ f_x(x|H_0) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{else} \end{cases} \]

With \( P(H_0) = P(H_1) = \frac{1}{2} \),

\[ \rightarrow H_1 = \left\{ \frac{4x^2}{1} > 1 \right\} = \left\{ x > \sqrt{\frac{1}{4}} \right\} \approx \left\{ x > 0.63 \right\} \]

With \( H_0 \) more probable a priori, say \( P(H_0) = \frac{3}{4} \),

\[ \rightarrow H_1 = \left\{ 4x^2 > 3 \right\} \approx \left\{ x > 0.91 \right\} \]
"Probability of correct detection"

\[ P_d = P(\rightarrow H_1 | H_1) = \int_{\rightarrow H_1} f_x(x | H_1) \, dx \]

is desirable to have near 1.

"Probability of false alarm"

\[ P_f = P(\rightarrow H_1 | H_0) = \int_{\rightarrow H_1} f_x(x | H_0) \, dx \]

is desirable to have near zero. There is usually some compromise necessary between these criteria; e.g., one can increase \( P_d \) by expanding \( \rightarrow H_1 \), but this will also increase \( P_f \).

With priors given, one can calculate the "Probability of error"

\[ P_e = P(\rightarrow H_1 | H_0) P(H_0) + P(\rightarrow H_0 | H_1) P(H_1) \]

\[ = P_f P(H_0) + (1 - P_d) P(H_1) \]
Conditioning a RV on an Event

Let \( A \) be an event defined in terms of a RV \( X \); e.g., \( A = \{0 < X < 1\} \) or \( A = \{ |X| < 5 \} \). The conditional PDF of \( X \) given \( A \) is

\[
\begin{align*}
  f_X(x|A) &= \begin{cases} 
    \frac{f_X(x)}{P(A)} & x \in A \\
    0 & x \notin A
  \end{cases}
\end{align*}
\]

Note that \( f_X(x|A) \geq 0 \) for all \( x \) and

\[
\int_{-\infty}^{\infty} f_X(x|A) \, dx = \frac{1}{P(A)} \int_{A} f_X(x) \, dx = 1,
\]

so \( f_X(x|A) \) is a PDF.

Conditional Moments

Moments are defined in the usual way with respect to the conditional PDF:

\[
E[X|A] = \int_{-\infty}^{\infty} x \cdot f_X(x|A) \, dx
\]

And the expectation rule applies:

\[
E[g(X)|A] = \int_{-\infty}^{\infty} g(x) \cdot f_X(x|A) \, dx
\]
So, \( E[x^n | A] = \int_{-\infty}^{\infty} x^n f_x(x | A) \, dx \) and

\[ \text{var}(x | A) = E[x^2 | A] - (E[x | A])^2 \]

Example:
\[ f_x(x) = \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & \text{else} \end{cases} \]

\[ A = \{ x > \frac{1}{2} \} \Rightarrow P(A) = \frac{3}{4} \]

\[ f_x(x | A) = \begin{cases} \frac{3}{4} & \frac{1}{2} \leq x \leq 1 \\ 0 & \text{else} \end{cases} \]

\[ E[x | A] = \frac{8}{3} \int_{\frac{1}{2}}^{1} x^2 \, dx = \frac{8}{9} x^3 \bigg|_{\frac{1}{2}}^{1} = \frac{8}{9} \left( 1 - \frac{1}{8} \right) = \frac{7}{9} \]

\[ E[x^2 | A] = \frac{8}{3} \int_{\frac{1}{2}}^{1} x^2 \, dx = \frac{2}{3} x^4 \bigg|_{\frac{1}{2}}^{1} = \frac{2}{3} \left( 1 - \frac{1}{16} \right) = \frac{5}{8} \]

\[ \text{var}(x | A) = \frac{5}{8} - \frac{49}{81} = \frac{13}{648} \]

**Transformation of one RV**

Suppose \( g(x) \) is a monotonically increasing function. Define a new RV \( y = g(x) \).

Then

\[ F_y(y) = P(\{y \leq y_0\}) = P(\{g(x) \leq y_0\}) = P(\{x \leq g^{-1}(y_0)\}) = F_x(g^{-1}(y_0)) \]
where $g^{-1}$ is defined so that

$$g(g^{-1}(y)) = y$$

Note that

$$\frac{d}{dy} g(g^{-1}(x)) = g'(g^{-1}(y))(g^{-1})'(y) = 1$$

so

$$(g^{-1})'(y) = \frac{1}{g'(g^{-1}(y))}$$

Thus,

$$f_y(y) = \frac{d}{dy} f_y(y) = \frac{d}{dy} F_y(g^{-1}(y))$$

$$= f_x(g^{-1}(y))(g^{-1})'(y)$$

$$= \left. \frac{f_x(x)}{g'(x)} \right|_{x=g^{-1}(y)}$$

Example: $x \sim U[-1,1]$, $y = g(x) = x^2$

$g^{-1}(y) = y^{\frac{1}{2}}$; $g'(x) = 2x$ \Rightarrow $f_y(y) = \left. \frac{\int_{-1}^{1} f_x(x) \, dx}{2x} \right|_{x=y^{\frac{1}{2}}} = \frac{1}{2} y^{\frac{1}{2}}$
for $-1 \leq y \leq 1$ and zero otherwise. Note that

$$\int_{-\infty}^{\infty} y f_y(y) dy = \frac{1}{6} \int_{-1}^{1} y^{2/3} dy = \frac{1}{2} y^{5/3} \bigg|_{-1}^{1} = 1$$

With this PDF for $y$, direct calculation yields

$$E[y^2] = \int_{-\infty}^{\infty} y^2 f_y(y) dy = \frac{1}{6} \int_{-1}^{1} y^{4/3} dy = \frac{1}{7}$$

Using the expectation rule with $f_x$ gives

$$E[y^2] = E[x^6] = \frac{1}{2} \int x^6 dx = \frac{1}{7}$$

If $g$ is monotonically decreasing, a similar calculation to the one above gives

$$f_y(y) = \frac{f_x(g(y))}{|g'(x)|} \bigg|_{x=g(y)}$$

If $g$ is monotonic on a sequence of disjoint intervals $A_k = [a_k, b_k]$ then

$$f_x(x) = \sum_{k} f_x(x) |A_k| p(A_k)$$
where each $f_{X}(x|A_{k})$ is monotonic. The expression (8) may be applied to each of these conditional PDFs individually, which can be combined according to (**) to obtain $f_{Y}(y)$.

**Linear Transformation of a Normal RV**

Let $x \sim N[\mu, \sigma^2]$ and set $y = g(x) = ax + b$ with $a \neq 0$. Then $g^{-1}(y) = \frac{y - b}{a}$ and $g'(x) = a$, and

$$f_{Y}(y) = \frac{f_{X}(x)}{|g'(x)|} \bigg|_{x = \frac{y - b}{a}}$$

$$= \frac{1}{|a|} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} (\frac{y - b}{a} - \mu)^2 \right\}$$

$$= \frac{1}{\sqrt{2\pi}(a\sigma^2)} \exp \left\{ -\frac{1}{2(a\sigma^2)} (y-(b\mu+b))^2 \right\}$$

So, $y \sim N[\mu a + b, a^2\sigma^2]$; i.e., a linear transformation on a normal RV produces another normal RV.

Even without calculating $f_{Y}$, it is obvious that $E[y] = \mu a + b$ and $\text{var}(y) = a^2\sigma^2$. Hence knowing $y$ is normal gives its PDF.
Discrete-Type Random Variables

Suppose \( X \) is a RV that takes on values in a discrete set \( R = \{ r_k, k \in \mathbb{Z}\} \). Then the CDF \( F_X \) of \( X \) is a piecewise-constant function with

\[
F_X(x) = \sum_{r_k \leq x} P(\{ X = r_k \})
\]

\( F_X \) has jump discontinuities at \( x = r_k \):

Such a CDF has a formal derivative in terms of Dirac delta symbols:

\[
f_X(x) = \sum_{r_k} P(\{ X = r_k \}) \delta(x - r_k)
\]

Note that

\[
\int_{-\infty}^{x} f_X(u) \, du = \sum_{r_k \leq x} P(\{ X = r_k \}) \int_{-\infty}^{x} \delta(u - r_k) \, du
\]

\[
= \sum_{r_k \leq x} P(\{ X = r_k \}) = F_X(x)
\]
and

\[
\int_{-\infty}^{\infty} f_x(x) \, dx = \sum_{r_k} P(\{x = r_k\}) = 1
\]

Example: \( x \) = outcome of a die roll

\( R = \{1, 2, 3, 4, 5, 6\} \quad r_k = k, \quad k = 1, \ldots, 6 \)

\[
P(\{x = r_k\}) = \frac{1}{6} \quad k = 1, \ldots, 6
\]

If \( x \) takes integer values, the discrete function

\[
p(k) = P(\{x = k\})
\]
is known as the probability mass function. It is clearly equivalent to the PDF given above.

Example (cont'd):

\[ p(k) \]

\[ 0 \ 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \rightarrow k \]

\[ E[x] = \sum_{k=-\infty}^{\infty} k \cdot p(k) = \frac{7}{2}, \text{ as above.} \]

**Mixed-Type RVs**

In certain situations, the PDF of a RV may contain both delta symbols and real values. Such a RV is said to be of mixed type.

Example: \( x \sim N[0, 1], y = x U(x) \) with \( U \) a unit step function. The PDF of \( y \) is

\[ f_y(y) = \begin{cases} \frac{1}{2} \delta(y) & y < 0 \\ \frac{1}{2} \delta(y) & y = 0 \\ f_x(x) & y > 0 \end{cases} \]

The delta captures \( P(x < 0) \); every such \( x \) gives \( y = 0 \).
Multiple RVs

The joint CDF of RVs $x_1, \ldots, x_n$ is

$$F_X(x_1, \ldots, x_n) = P(x_1 \leq x_1 \cap \ldots \cap x_n \leq x_n)$$

The subscript $X$ represents the vector containing the $x_k$ as its components; i.e.,

$$X = \left[ \begin{array}{c} x_1 \\ \vdots \\ x_n \end{array} \right]$$

Note that $0 \leq F_X(x_1, \ldots, x_n) \leq 1$. Also, if $A > 0$,

$$F_X(x_1, \ldots, x_{k-1}, x_{k-1} + A, x_{k+1}, \ldots, x_n) \geq F_X(x_1, \ldots, x_n)$$

and

$$\lim_{x_1 \to -\infty, x_n \to -\infty} F_X(x_1, \ldots, x_n) = 0$$

$$\lim_{x_1 \to +\infty, x_n \to +\infty} F_X(x_1, \ldots, x_n) = 1$$

These properties are obvious from the definition.
If \( F_x \) is differentiable, the joint PDF of \( x_1, \ldots, x_n \) is

\[
f_{x}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{x_i}(x_i)
\]

Note that \( \prod \) implies \( f_x(x_1, \ldots, x_n) > 0 \). Also,

\[
F_x(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{x_i}(u_1, \ldots, u_n) \, du_1 \cdots du_n
\]

so that \( \prod \frac{d}{dx} \) implies

\[
\int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{x_i}(x_1, \ldots, x_n) \, dx_1 \cdots dx_n = 1
\]

**Marginal PDFs**

The CDF of \( x_1 \) can be obtained from the joint CDF of \( x_1, \ldots, x_n \) by allowing \( x_2, \ldots, x_n \) to take on any possible value, i.e.,

\[
F_{x_1}(x_1) = \lim_{x_2 \to \infty} F_x(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{x_i}(u_1, \ldots, u_n) \, du_1 \cdots du_n
\]
So
\[ f_{X_1}(x_1) = \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} f_{X_1, \ldots, X_n}(x_1, \ldots, x_n) \, dx_2 \ldots \, dx_n \right) \, dx_1. \]

Similarly, the joint PDF of any subset of the \( X_k \) can be obtained by integrating \( f_X \) from \(-\infty\) to \( \infty \) with respect to each undesired variable. This process is known as marginalization of \( f_X \).

**Example:**
\[ f_{X_1, X_2}(x_1, x_2) = \begin{cases} 1 & 0 \leq x_1 \leq 2, \quad 0 \leq x_2 \leq 1, \quad x_1 \leq x_2 \\ 0 & \text{otherwise} \end{cases} \]

\[ f_{X_2}(x_2) = \int_{-\infty}^{\infty} f_{X_1, X_2}(x_1, x_2) \, dx_1 = \int_{x_2}^{\infty} 1 \, dx_1 = 2x_2, \]

for \( 0 \leq x_2 \leq 1 \) and \( 0 \) otherwise. From this, it is straightforward to calculate \( E[X_2] \), \( \text{var}(X_2) \), etc.
Multiple RVs, cont'd

It is convenient to introduce some abbreviated notation:

(a) \( \mathbf{X} \) denotes the \( n \)-tuple \((X_1, \ldots, X_n)\) of RVs and also the random vector

\[
\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}
\]

(b) \( X \) denotes the corresponding \( n \)-tuple or vector of deterministic variables

(c) \[
\int_A g(X) \, dX = \int_{(x_1, \ldots, x_n) \in A} g(x_1, \ldots, x_n) \, dx_1 \ldots \, dx_n
\]

With this notation, the joint CDF of RVs \( X_1, \ldots, X_n \) is \( F_X(X) \) and the joint PDF is \( f_X(X) \).

Probability of an Event in \( \mathbf{X} \)

\[
P(A) = P(\{ (x_1, \ldots, x_n) \in A \}) = \int_A f_X(X) \, dX
\]
Conditioning on an Event

The conditional PDF of $X$ given $X \in A$ is

$$f_X(x | A) = \begin{cases} \frac{f_X(x)}{P(A)} & x \in A \\ 0 & x \notin A \end{cases}$$

Example: $f_X(x_1, x_2) = \begin{cases} 6x_1^2x_2 & 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$

$$A = \{x_2 \geq x_1^2\}$$

$$P(A) = \int_0^1 \int_0^{x_1^2} 6x_1^2x_2 \, dx_1 \, dx_2 = \frac{1}{3} \int_0^1 x_1^3 \, dx_1 = 2 \int_0^1 x_2^{5/2} \, dx_2 = \frac{4}{5}$$

Thus,

$$f_X(x_1, x_2 | A) = \begin{cases} \frac{15}{2} x_1^2 x_2 & (x_1, x_2) \in A \\ 0 & \text{otherwise} \end{cases}$$

Independent RVs

RVs $x_1, \ldots, x_n$ are statistically independent if

$$f_X(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$
Examples: Independence can be checked by marginalization of the joint PDF.

(a) \[ f_{X_1}(x_1) = \int_0^1 f_{X_1, X_2}(x_1, x_2) \, dx_2 = 3 \int_0^1 \frac{x_1^2}{2} \, dx_2 = 3x_1, \quad 0 \leq x_1 \leq 1 \]

(b) \[ f_{X_1}(x_1) = \int_0^2 f_{X_1, X_2}(x_1, x_2) \, dx_2 = 30 \int_0^1 \frac{x_1^2}{2} \, dx_2 = 30x_1^2(1-x_1)^2, \quad 0 \leq x_1 \leq 1 \]

Since \( f_{X_1}(x_1) f_{X_2}(x_2) = f_{X}(x_1, x_2) \), \( X_1 \) and \( X_2 \) are statistically independent.

(b) \[ f_{X}(x_1, x_2) = \begin{cases} 15x_1^2 x_2 & 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 2, x_1 + x_2 \leq 2 \\ 0 & \text{otherwise} \end{cases} \]

Because \( f_{X_1}(x_1) f_{X_2}(x_2) \neq f_{X}(x_1, x_2) \), \( X_1 \) and \( X_2 \) are not independent RVs.
Expectation Rule

If \( z = g(x_1, \ldots, x_n) \), then

\[
E[z] = \int_{\mathbb{R}^n} g(x) f_x(x) \, dx
\]

In particular, this implies the following:

(a) \( E[x_k^l] = \int_{\mathbb{R}^n} x_k^l f_x(x) \, dx \) are the moments of \( x_k \).

(b) \( E[a_1x_1 + \cdots + a_nx_n] = \int_{\mathbb{R}^n} (a_1x_1 + \cdots + a_nx_n) f_x(x) \, dx \)

\[
= a_1 E[x_1] + \cdots + a_n E[x_n]
\]

(c) Denoting \( E[\mathbf{x}] = \begin{bmatrix} E[x_1] \\ \vdots \\ E[x_n] \end{bmatrix} = \begin{bmatrix} \mu_{x_1} \\ \vdots \\ \mu_{x_n} \end{bmatrix} \),

if

\[
A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}
\]

is an \( m \times n \) matrix, then
\[ E[AX] = \begin{bmatrix} E[a_{11}X_1 + \ldots + a_{1n}X_n] \\ \vdots \\ E[a_{m1}X_1 + \ldots + a_{mn}X_n] \end{bmatrix} = \begin{bmatrix} a_{11}\mu_1 + \ldots + a_{1n}\mu_n \\ \vdots \\ a_{m1}\mu_1 + \ldots + a_{mn}\mu_n \end{bmatrix} = \mathbf{A} \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \mathbf{A}\mathbf{\mu} \]

Similarly, \[ E[XAX^T] = E[X^T]\mathbf{A}^T. \]
Beta Distribution

For $a > 0$ and $b > 0$, the beta function is defined by

$$B(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} \, dx = \frac{\Gamma(a) \Gamma(b)}{\Gamma(a+b)}$$

A RV $x$ is beta distributed with parameters $a > 0$ and $b > 0$ if

$$f_x(x) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}$$

for $0 \leq x \leq 1$ and zero otherwise. This is expressed by the notation $X \sim \beta(a, b)$.

Note that, for $X \sim \beta(a, b)$,

$$\int_0^\infty f_x(x) \, dx = \frac{1}{B(a, b)} \int_0^1 x^{a-1} (1-x)^{b-1} \, dx = \frac{B(a, b)}{B(a, b)} = 1$$

and

$$E[X] = \frac{1}{B(a, b)} \int_0^1 x^a (1-x)^{b-1} \, dx = \frac{B(a+1, b)}{B(a, b)} = \frac{\Gamma(a+1) \Gamma(b)}{\Gamma(a+b+1) \Gamma(a) \Gamma(b)} = \frac{a}{a+b}$$
Covariance of two RVs

The covariance of \( x_1 \) and \( x_2 \) is

\[
\text{cov}(x_1, x_2) = E[(x_1 - \mu_1)(x_2 - \mu_2)]
\]

where \( \mu_1 = E[x_1] \) and \( \mu_2 = E[x_2] \). Note that

\[
\text{cov}(x_1, x_2) = E[x_1 x_2 - x_1 \mu_2 - x_2 \mu_1 + \mu_1 \mu_2] = E[x_1 x_2] - \mu_1 \mu_2
\]

RVs \( x_1 \) and \( x_2 \) are uncorrelated if \( \text{cov}(x_1, x_2) = 0 \); i.e., if \( E[x_1 x_2] = \mu_1 \mu_2 \).

Theorem: Independent RVs are uncorrelated.

Proof: \( E[xy] = \iint_{-\infty}^{\infty} xy f_{xy}(x, y) \, dx \, dy \) (Expectation)

\[
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_x(x) f_y(y) \, dx \, dy \quad (\text{Independence})
\]

\[
= \int_{-\infty}^{\infty} f_x(x) \, dx \int_{-\infty}^{\infty} f_y(y) \, dy = E[x] E[y]
\]
Suppose $x_1$ and $x_2$ are independent and $Z = X_1 + X_2$. Then

$$F_Z(z) = P(Z \leq z) = P(X_1 + X_2 \leq z)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{z-x_1} f_{X_1}(x_1) f_{X_2}(x_2) \, dx_2 \, dx_1$$

Using the fundamental theorem of calculus,

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \int_{-\infty}^{\infty} \left[ \frac{d}{dz} \int_{-\infty}^{z-x_1} f_{X_1}(x_1) \, dx_1 \right] f_{X_2}(z-x_1) \, dx_1$$

$$= \int_{-\infty}^{\infty} f_{X_1}(x_1) f_{X_2}(z-x_1) \, dx_1 = [f_{X_1} * f_{X_2}](z)$$

i.e., the convolution of $f_{X_1}$ and $f_{X_2}$.

Example:

$f_{X_1}(x_1) = \begin{cases} 3/4 & 0 \leq x_1 \leq 1, 0 \leq x_2 \leq x_1 \\ 0 & \text{otherwise} \end{cases}$

$f_{X_2}(x_1, x_2) = \begin{cases} 3/4 & 1 \leq x_1 \leq 2, 0 \leq x_2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$

$$f_{X_1}(x_1) = \int_{\sqrt{x_1}}^{2} \frac{3}{4} \, dx_2 = \frac{3}{4} (2 - \sqrt{x_1}) \quad 0 \leq x_1 \leq 1$$
So,
\[ E[X_{1,1}] = \frac{3}{4} \int_{0}^{1} x_2^2 \, dx_2 = \frac{9}{20} \]

Similarly,
\[
 f_{X_1}(x_1) = \begin{cases} 
 \frac{3}{4}, & 1 \leq x_1 \leq 2 \\
 \frac{1}{8}, & 0 \leq x_1 < 1 
\end{cases}
\]

and
\[
 E[X_1] = \int_{0}^{1} \frac{3}{4} x_1^3 \, dx_1 + \int_{0}^{2} \frac{3}{4} x_1 \, dx_1 = \frac{3}{16} + \frac{9}{8} = \frac{21}{16}
\]

Finally,
\[
 E[X_1X_2] = \int_{0}^{1} \int_{0}^{\frac{3}{4} x_1} x_2 \, dx_1 \, dx_2 = \frac{3}{4} \int_{0}^{1} x_2^2 \, dx_2
\]
\[
 = \frac{3}{8} \int_{0}^{1} x_2 (4 - x_2) \, dx_2 = \frac{5}{8}
\]

and
\[
 \text{cov}(X_1, X_2) = \frac{5}{8} - \left( \frac{9}{20} \right) \left( \frac{21}{16} \right) = \frac{11}{320}.
\]

Not surprisingly, these RVs are positively correlated.

Now, define \( A = \{ x_1 > 1 \} \). Then \( P(A) = \frac{3}{4} \) and
\[
 f_{X_1}(x_1 | A) = \begin{cases} 
 1, & 1 \leq x_1 \leq 2, \ 0 \leq x_1 < 1 \\
 0, & \text{otherwise}
\end{cases}
\]

Thus, \( f_{X_1}(x_1 | A) \sim U[1,2] \), \( f_{X_2}(x_2 | A) \sim U[0,1] \), and
\[ f_{x_1, x_2}(x_1, x_2 | A) = f_{x_2}(x_2 | A) f_{x_1}(x_1 | A) \]
i.e., \( x_1 \) and \( x_2 \) are conditionally independent given \( A \). Consequently,

\[ \text{cov}(x_1, x_2 | A) = 0 \]

which implies \( E[x_1 x_2 | A] = E[x_1 | A] E[x_2 | A] = \frac{3}{4} \)
Linear Algebra on $\mathbb{C}^n$ (review)

Notation:

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}, \text{ } mxn \text{ matrix } \alpha_{ij} \in \mathbb{C}$$

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \text{ } n \text{-vector } x_i \in \mathbb{C}$$

$$A^+ = \begin{bmatrix} a_{11}^* & \cdots & a_{1m}^* \\ \vdots & \ddots & \vdots \\ a_{m1}^* & \cdots & a_{mn}^* \end{bmatrix}, \text{ } nxm \text{ matrix, Hermitian transpose of } A.$$  
$$a_{ij}^* = \text{complex conjugate of } a_{ij}$$

$$\|X\| = \sqrt{X^+X} \geq 0, \text{ norm of } X, \text{ zero only if } X \text{ is the zero vector}$$

Positive Definite Matrices

An $nxn$ matrix $S$ is positive definite if

$$X^+SX > 0$$

for every non-zero vector $X$. 
S is non-negative definite if $X^+SXX > 0$ for every vector $X$.

Note: If $A$ is $m \times n$ and $S = A^+A$, then

$X^+SXX = X^+A^+AX = (AX)^+(AX) = ||AX||^2 > 0$

Eigenvalues and Eigenvectors

If $S$ is $n \times n$ and $E$ is an $n$-vector such that

$SE = \lambda E$

for some scalar $\lambda$, $E$ is an eigenvector of $S$ and $\lambda$ is its associated eigenvalue.

Theorem: If $S$ is positive definite (non-negative definite) every eigenvalue of $S$ is positive (non-negative).

Proof: $SE = \lambda E$ implies $E^+SE = E^+\lambda E = \lambda E^+E$,
so

$E^+SE = \lambda ||E||^2$ and $\lambda = \frac{E^+SE}{||E||^2} > 0$.

(Similarly for non-negative definite $S$.)
Remarks about eigenvalues and eigenvectors:

a) If \( \lambda_1, \ldots, \lambda_n \) are eigenvalues of \( S \),
\[
\det S = \lambda_1 \cdots \lambda_n \quad \text{tr} S = \lambda_1 + \cdots + \lambda_n
\]

b) If \( E_1 \) and \( E_2 \) are eigenvectors of \( S \) corresponding to distinct eigenvalues, the \( E_1^T E_2 = 0 \) (i.e., \( E_1 \) and \( E_2 \) are orthogonal).

c) A positive definite matrix has positive determinant and is thus invertible.

d) If \( \lambda_1, \ldots, \lambda_n \) are eigenvalues of an \( n \times n \) matrix \( S \) with corresponding orthonormal eigenvectors \( E_1, \ldots, E_n \), then
\[
S = \lambda_1 E_1 E_1^T + \cdots + \lambda_n E_n E_n^T
\]

e) If \( S \) has \( l \) non-zero eigenvalues, \( r = n - l \) is called the rank of \( S \).
Example: \( S = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \)

Note that:

\[
\begin{bmatrix} x_1^* & x_2^* \end{bmatrix} S \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1^* x_1^2 \\ x_2^* x_2^2 \end{bmatrix} \begin{bmatrix} 3x_1 + x_2 \\ x_1 + 3x_2 \end{bmatrix} = 3|x_1|^2 + 2\text{Re}(x_1 x_2^*) + 3|x_2|^2 > 0
\]

The eigenvalues are roots of the characteristic polynomial, i.e.,

\[
0 = \det(\lambda I - S) = \det\begin{bmatrix} \lambda - 3 & -1 \\ -1 & \lambda - 2 \end{bmatrix} = \lambda^2 - 6\lambda + 8
\]

\[
= (\lambda - 4)(\lambda - 2)
\]

So \( \lambda_1 = 4, \lambda_2 = 2 \). Note \( \lambda_1 + \lambda_2 = 6 = \text{tr} S \) and \( \lambda_1 \lambda_2 = 8 = \det S \).

\( E_1 = \begin{pmatrix} a \\ b \end{pmatrix} \) satisfies \( SE_1 = \lambda_1 E_1 = 4 \begin{pmatrix} a \\ b \end{pmatrix} \), so \( 3a + b = 4a \Rightarrow a = b \). Choosing \( \|E_1\| = 1 \),

\[
E_1 = \begin{pmatrix} \sqrt{2} \\ 1 \end{pmatrix}, \quad E_2 = \begin{pmatrix} \sqrt{2} \\ -1 \end{pmatrix}
\]

where \( E_2 \) is obvious from the orthogonality condition. Finally

\[
\lambda_1 E_1 E_1^* + \lambda_2 E_2 E_2^* = 4 \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + 2 \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} = S.
\]
Unitary Matrices

A \( n \times n \) matrix \( U \) is unitary if \( U^*U = I \), where

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

is the \( n \times n \) identity matrix. Note that \( U^* = U^t \) so \( UU^* = I \).

Writing

\[
U = \begin{bmatrix}
U_1 & \cdots & U_n
\end{bmatrix}
\]

gives

\[
U^*U = \begin{bmatrix}
U_1^* & \cdots & U_n^*
\end{bmatrix}
\begin{bmatrix}
U_1 & \cdots & U_n
\end{bmatrix}
= \begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix}
\]

so that \( U_i^*U_j = 1 \) if \( i=j \) and \( U_i^*U_j = 0 \) if \( i \neq j \)

i.e., the columns (and rows) of \( U \) are orthonormal vectors.

If \( \lambda \) is an eigenvalue of \( U \) with corresponding eigenvector \( E \),

\[
UE = \lambda E \Rightarrow U^*UE = U^t \lambda E \Rightarrow U^*E = \frac{\lambda}{\lambda} E
\]

i.e., \( E \) is an eigenvector of \( U^t \) with eigenvalue \( \frac{1}{\lambda} \).
Covariance Matrix

The covariance matrix of a random vector \( \mathbf{X} \) is the \( n \times n \) matrix

\[
\mathbf{C}_\mathbf{X} = \mathbf{E}\left[ (\mathbf{X} - \mathbf{M}_\mathbf{X})(\mathbf{X} - \mathbf{M}_\mathbf{X})^T \right]
\]

\[
= \begin{bmatrix}
\text{var}(X_1) & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_n) \\
\vdots & \ddots & \vdots & \vdots \\
\text{cov}(X_n, X_1) & \cdots & \text{var}(X_n)
\end{bmatrix}
\]

where \( \mathbf{M}_\mathbf{X} = \left( \begin{array}{c} \mu_{X_1} \\ \vdots \\ \mu_{X_n} \end{array} \right) = \mathbf{E}[\mathbf{X}]. \)

Note that \( \text{cov}(X_i, X_j) = \text{cov}(X_j, X_i) \) implies \( \mathbf{C}_\mathbf{X} \) is symmetric. Because \( \mathbf{C}_\mathbf{X} \) has the form \( \mathbf{A} \mathbf{A}^T \), it is non-negative definite.

Example: Suppose

\[
\mathbf{C}_\mathbf{X} = \begin{bmatrix} 3 & -1 \\ -1 & 2 \end{bmatrix}, \quad \mathbf{M}_\mathbf{X} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}
\]

which has eigenvalues \( \lambda_1 = 4 \) and \( \lambda_2 = 2 \) and corresponding eigenvectors:

\[
\mathbf{E}_1 = \frac{\sqrt{2}}{2} \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \mathbf{E}_2 = \frac{\sqrt{2}}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix}
\]
Plotting these eigenvectors weighted by their corresponding eigenvalues, centered at $M_X$, gives the axes of the covariance ellipse. Note the angle of the ellipse of these negatively correlated variables.

Note that $C_X = E[XX^T] - M_X M_X^T$.

**Linear Transformations**

If $A$ is a (deterministic) $m \times n$ matrix and $X$ a random $n$-vector, then $\tilde{Y} = AX$ is a random $m$-vector. Recall that

$$E[\tilde{Y}] = E[AX] = A E[X]$$

The covariance matrix of $\tilde{Y}$ is

$$C_Y = E[(\tilde{Y} - M_{\tilde{Y}})(\tilde{Y} - M_{\tilde{Y}})^T]$$

$$= E[(AX - AM_X)(AX - AM_X)^T]$$

$$= AE[(X - M_X)(X - M_X)^T]A^T$$

$$= AC_X A^T$$
**Multivariate Normal Distribution**

A random vector \( \mathbf{X} \) is multivariate normal (Gaussian) if its PDF is

\[
f_{\mathbf{X}}(\mathbf{X}) = \frac{1}{(2\pi)^{n/2}\det C} \exp \left\{ -\frac{1}{2} (\mathbf{X} - \mathbf{M})^T C^{-1} (\mathbf{X} - \mathbf{M}) \right\}
\]

for a positive definite matrix \( C \) and a deterministic vector \( \mathbf{M} \).

**Notes:**

1. Direct integration confirms \( \int_{\mathbb{R}^n} f_{\mathbf{X}}(\mathbf{X}) \, d\mathbf{X} = 1 \).

2. Application of the expectation rule shows

\[
E[\mathbf{X}] = \int_{\mathbb{R}^n} \mathbf{X} f_{\mathbf{X}}(\mathbf{X}) \, d\mathbf{X} = \mathbf{M}
\]

\[
C_{\mathbf{X}} = \int_{\mathbb{R}^n} (\mathbf{X} - \mathbf{M})(\mathbf{X} - \mathbf{M})^T f_{\mathbf{X}}(\mathbf{X}) \, d\mathbf{X} = C
\]

3. With parameters \( \mathbf{M} \) and \( C \), one denotes

\( \mathbf{X} \sim N[\mathbf{M}, C] \)
Uncorrelated Normal Random Vectors

A random vector \( \mathbf{X} \) is uncorrelated if \( \mathbf{C}_\mathbf{X} \) is a diagonal matrix.

If \( \mathbf{X} \) is Gaussian and uncorrelated, then

\[
\mathbf{C}_\mathbf{X} = \begin{bmatrix}
\sigma_1^2 & 0 \\
0 & \sigma_2^2 \\
\cdots & \cdots \\
0 & \sigma_n^2
\end{bmatrix}
\]

and \( \det \mathbf{C}_\mathbf{X} = \sigma_1^2 \cdots \sigma_n^2 \),

\[
\mathbf{C}_\mathbf{X} = \begin{bmatrix}
\sigma_1^{-2} & 0 \\
0 & \sigma_2^{-2} \\
\cdots & \cdots \\
0 & \sigma_n^{-2}
\end{bmatrix}
\]

So,

\[
f_{\mathbf{X}}(\mathbf{X}) = \frac{1}{(2\pi)^{n/2} \sigma_1 \cdots \sigma_n} \exp \left\{ -\frac{1}{2} \sum_{k=1}^{n} (x_k - M_k)^2 \sigma_k^{-2} \right\}
\]

\[
= \frac{1}{\sqrt{2\pi \sigma_1^2}} \exp \left\{ -\frac{(x_1 - M_1)^2}{2 \sigma_1^2} \right\} \cdots \frac{1}{\sqrt{2\pi \sigma_n^2}} \exp \left\{ -\frac{(x_n - M_n)^2}{2 \sigma_n^2} \right\}
\]

\[
= f_{x_1}(x_1) \cdots f_{x_n}(x_n)
\]

i.e., the components of an uncorrelated Gaussian random vector are independent Gaussian RVs.
Decorrelation and Whitening

Suppose \( \mathbf{U}_1, \ldots, \mathbf{U}_n \) are orthonormal eigenvectors of \( \mathbf{C}_X \). Then the matrix

\[
\mathbf{U}^T = \begin{bmatrix} \mathbf{U}_1 & \cdots & \mathbf{U}_n \end{bmatrix}
\]

is unitary and the random vector \( \mathbf{Y} = \mathbf{UX} \) has covariance matrix

\[
\mathbf{C}_Y = \mathbf{UC}_X \mathbf{U}^T = \mathbf{U} \begin{bmatrix} \lambda_1 \mathbf{U}_1 & \cdots & \lambda_n \mathbf{U}_n \end{bmatrix}
\]

\[
= \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}
\]

where \( \lambda_1, \ldots, \lambda_n \) are the (non-negative) eigenvalues corresponding to \( \mathbf{U}_1, \ldots, \mathbf{U}_n \) respectively.

Note that \( \mathbf{Y} \) is an uncorrelated random vector.

A random vector \( \mathbf{Y} \) is white if \( \mathbf{C}_Y \) is a constant times \( \mathbf{I} \); under some circumstances one also requires

\[
\mathbb{E}[\mathbf{Y}] = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix},
\]
With the notation above, define

$$V^T = \left[ \frac{1}{\sqrt{n_1}} u_1 \cdots \frac{1}{\sqrt{n_k}} u_n \right].$$

If $Y \sim VX$, then

$$C_X = V C_x V^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

so $Y$ is white.
Complex Random Vectors

Complex Vectors

A complex n-vector $z \in \mathbb{C}^n$ has the form

$$z = x + iy = \begin{bmatrix}
  x_1 + iy_1 \\
  \\
  x_n + iy_n
\end{bmatrix}$$

where $x$ and $y$ are real n-vectors (i.e., $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$) with $i^2 = -1$.

Notes:
(a) $x$ is called the real part of $z$.
(b) $y$ is the imaginary part of $z$.

The hermitian transpose of $z$ is

$$z^\dagger = [z_1^* \ldots z_n^*] = [x_1 - iy_1 \ldots x_n - iy_n]$$

where $z_k^*$ denotes the complex conjugate of the complex scalar $z_k$.

The (euclidean) norm of $z$ is

$$\|z\| = (z^\dagger z)^{1/2} = \sqrt{\sum_{k=1}^{n} x_k^2 + y_k^2} \geq 0$$
Complex Matrices

The Hermitian transpose of a complex matrix $A = [a_{ij}]$ is

$$A^\dagger = [a_{*j}^\dagger]$$

i.e., $A^\dagger$ is the transpose of the matrix whose elements are the complex conjugates of those of $A$.

A complex $n \times n$ matrix $A$ is positive definite if

$$Z^\dagger A Z \geq 0$$

for all complex $n$-vectors $Z$. Non-negative definite matrices are defined analogously.

Note: If $R = A^\dagger A$, then

$$Z^\dagger R Z = Z^\dagger A^\dagger A Z = (AZ)^\dagger (AZ)$$

$$= ||AZ||^2 \geq 0$$

so $A^\dagger A$ is non-negative definite.
A \( n \times n \) complex matrix \( A \) is hermitian if \( A^\dagger = A \).

**Complex Random Vectors**

A complex random vector has the form

\[
\mathbf{Z} = \begin{bmatrix} X + iy_1 \\ \\ X_n + iy_n \end{bmatrix}
\]

where the real part \( X \) and imaginary part \( Y \) are real random vectors.

The mean of \( Z \) is \( \mathbf{M}_Z = E[Z] = E[X] + iE[Y] \)

\[
= M_X + iM_Y.
\]

The covariance matrix of \( \mathbf{Z} \) is the hermitian non-negative definite matrix

\[
C_Z = E[(Z - \mathbf{M}_Z)(Z - \mathbf{M}_Z)^\dagger]
\]

Note that the CDF of \( Z \) is taken as the joint CDF of \( X \) and \( Y \); i.e.,

\[
F_Z(Z) = F_{XY}(X,Y) = P\{x_1 \leq x_1, \ldots, x_n \leq x_n, y_1 \leq y_1, \ldots, y_n \leq y_n\}
\]
The PDF of $\mathbf{z}$ is thus

$$f_{\mathbf{z}}(\mathbf{z}) = \frac{2^n}{\partial y_1 \partial x_1 \ldots \partial y_k \partial x_k} F_{\mathbf{z}}(\mathbf{z})$$

provided the derivative exists in some suitable sense.

**Proper Complex Random Vectors**

Consider the covariance matrix of a complex random vector $\mathbf{z} = \mathbf{x} + i \mathbf{y}$,

$$C_{\mathbf{z}} = E \left[ (\mathbf{z} - M_{\mathbf{z}})(\mathbf{z} - M_{\mathbf{z}})^\dagger \right]$$

$$= E \left[ \left( (\mathbf{x} - M_{\mathbf{x}}) + i(\mathbf{y} - M_{\mathbf{y}}) \right) \left( \left( \mathbf{x} - M_{\mathbf{x}} \right)^\dagger - i \left( \mathbf{y} - M_{\mathbf{y}} \right)^\dagger \right) \right]$$

$$= E \left[ (\mathbf{x} - M_{\mathbf{x}})(\mathbf{x} - M_{\mathbf{x}})^\dagger \right] + E \left[ (\mathbf{y} - M_{\mathbf{y}})(\mathbf{y} - M_{\mathbf{y}})^\dagger \right]$$

$$+ i E \left[ (\mathbf{y} - M_{\mathbf{y}})(\mathbf{x} - M_{\mathbf{x}})^\dagger \right] - i E \left[ (\mathbf{x} - M_{\mathbf{x}})(\mathbf{y} - M_{\mathbf{y}})^\dagger \right]$$

$$= C_{\mathbf{x}} + C_{\mathbf{y}} + i(C_{yx} - C_{xy})$$

$\mathbf{z}$ is proper if $C_{xy} = C_{yx}$ and $C_{yx} = -C_{xy}$, in which case

$$C_{\mathbf{z}} = 2 \left( C_{\mathbf{x}} + i C_{yx} \right)$$
Note that for $Z_i$ proper,

\[
E \left[ (Z_i - M_Z)(Z_i - M_Z)^T \right] = E \left[ \left\{ (X_i - M_X) + i(Y_i - M_Y) \right\} \left\{ (X_i - M_X)^T + i(Y_i - M_Y)^T \right\} \right] = C_{X_i} - C_{Y_i} + i(C_{XY} + C_{YX}) = 0
\]

Complex Normal Random Vectors

A proper complex random $n$-vector $Z$ is complex normal if

\[
f_{\mathbf{Z}}(\mathbf{Z}) = \frac{1}{(2\pi)^{n/2} \det C} \exp \left\{ - (\mathbf{Z} - \mathbf{M})^C (\mathbf{Z} - \mathbf{M}) \right\}
\]

for some positive definite Hermitian matrix $C$ and some complex $n$-vector $M$.

Notes:

(a) Direct calculation reveals that $E[Z_i] = M$ and $C_{Z_i} = C$ for $Z_i$ having this PDF.

(b) This distribution is indicated by the notation $Z_i \sim \text{CN}[M, C]$ or $Z_i \sim \text{CN}[M,C]$. 

(c) Proper complex normal random vectors (i.e., having the above PDF) are also called circular complex normal random vectors.
Parameter Estimation

Consider a family of PDFs indexed by a parameter \( \theta \in \mathbb{R}^m \) and denoted \( \{ f_x(x|\theta) \} \). For the moment, \( x \) will be assumed to be a (scalar) RV.

**Example:** The family is unit-variance normal PDFs indexed by their means \( \theta = \mu_x \in \mathbb{R} \)

\[
f_x(x|\theta) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} (x-\theta)^2 \right\}
\]

Suppose \( x_1, \ldots, x_n \) are iid RVs drawn from \( f_x(x|\theta) \) for a fixed value of \( \theta \). An estimator for the parameter \( \theta \) from data \( x_1, \ldots, x_n \) is a function

\[
g: \mathbb{R}^n \rightarrow \mathbb{R}^m
\]

One writes

\[
\hat{\theta} = g(x_1, \ldots, x_n) = g(\bar{x})
\]

Note that the estimate \( \hat{\theta} \) is a RV because it is a function of the RVs \( x_1, \ldots, x_n \); each realization of \( \bar{x} \) give a realization of \( \hat{\theta} \).
Bias

The bias of an estimator $\hat{\theta}$ of $\theta$ is

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

An estimator is unbiased if $\beta(\hat{\theta}) = 0$; i.e., if $E[\hat{\theta}] = \theta$.

Example (Sample mean estimator). Suppose $x_1, \ldots, x_n$ are iid with mean $\theta = \mu$. Define an estimator for $\theta$ by

$$\hat{\theta} = g(x_1, \ldots, x_n) = \frac{1}{n}(x_1 + \ldots + x_n)$$

Then $E[\hat{\theta}] = \frac{1}{n}E[x_1 + \ldots + x_n] = \frac{1}{n}(n\mu) = \mu = \theta$. So $\hat{\theta}$ is an unbiased estimator of the mean.

Estimator Variance

A second measure of estimator quality is its variance:

$$\text{var}(\hat{\theta})$$

For an unbiased estimator, small variance
ensures that the probability that a realization of \( \hat{\Theta} \) will be near \( \Theta \) is high.

PDFs of two unbiased estimators

\[ \begin{array}{c}
\Theta \quad \text{Small variance} \\
\hat{\Theta} \quad \text{Large variance}
\end{array} \]

Example (cont'd): For the sample mean estimator

\[ \hat{\Theta} = \frac{1}{n} (X_1 + \cdots + X_n), \]

\[ \text{var} (\hat{\Theta}) = \frac{1}{n^2} \text{var} (X_1 + \cdots + X_n) = \frac{1}{n} \text{var} (X) \]

Note that the variance of this unbiased estimator of \( \mu_x \) decreases as the number \( n \) of independent data increases. An estimator with this property is called consistent.

Standard Estimators

1. Sample mean \( \hat{\mu}_x = \frac{1}{n} \sum_{k=1}^{n} X_k \)

- Unbiased estimator for \( \mu_x = E[X] \).

- \( \text{var} (\hat{\mu}_x) = \frac{1}{n} \text{var} (X) \)
2. Sample Variance Estimator

\[ \hat{\sigma}^2 = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \hat{\mu})^2 \]

- Bias:

\[ E[\hat{\sigma}^2] = \frac{1}{n-1} \sum \left[ \frac{x_k^2}{n} - 2 \frac{x_k}{n} \hat{\mu} + \hat{\mu}^2 \right] \]

\[ = \frac{1}{n-1} \left\{ n \frac{E[x_k^2]}{n} - 2 \sum \frac{E[x_k \hat{\mu}]}{n} \hat{\mu} + n E[\hat{\mu}^2] \right\} \]

\[ = \frac{1}{n-1} \left\{ n \frac{\text{var}(x) + \mu^2}{n} - 2 \sum \frac{\frac{\sum x_j}{n} \hat{\mu}}{n} \hat{\mu} + n \frac{\text{var}(\hat{\mu}) + \mu^2}{n} \right\} \]

\[ = \frac{1}{n-1} \left\{ n \text{var}(x) + n \mu^2 - 2 \frac{n(n-1)}{n} \mu^2 - 2 \frac{n \text{var}(x) + n \mu^2}{n} \right\} \]

\[ = \frac{1}{n-1} \left( n-2+1 \right) \text{var}(x) + \frac{1}{n-1} \left( n-2 \right) \left( n-1 \right) - 2 + n \mu^2 \]

\[ \text{var}(\hat{\sigma}^2) = \frac{\text{var}(\hat{\mu}) + \mu^2}{n} \]

- Variance: In general, \( \text{var}(\hat{\sigma}^2) \) is not possible to calculate without additional assumptions on \( f_x \). If \( x \) is Gaussian,

\[ \text{var}(\hat{\sigma}^2) = \frac{2 (\text{var}(x))^2}{n-1} \]

and \( \hat{\sigma}^2 \) is a consistent estimator of \( \text{var}(x) \).
3. Sample Covariance Estimator

With data \( z = (x_1, y_1), \ldots, (x_n, y_n) \) independently drawn from a joint PDF \( f_{xy} \), the sample covariance

\[
\hat{c} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \hat{\mu}_x)(y_k - \hat{\mu}_y)
\]

is an unbiased estimator of \( \text{cov}(x, y) \).
Maximum-likelihood Estimation

Given a family \([f(x|\theta)]\) of probability densities parameterized by \(\theta \in \mathbb{R}^m\) and data \(x_1, \ldots, x_n\) drawn from one of these PDFs, the goal of estimation is to determine a value for \(\theta\) that is most compatible with the data in some sense. One such sense is maximum likelihood.

Suppose \(X = (x_1, \ldots, x_n)^T\) with each \(x_k\) having PDF \(f_{x_k}(x_k|\theta)\). The likelihood function is

\[
l(\theta, X) = f_X(X|\theta)
\]

The log-likelihood function is

\[
L(\theta, X) = \log l(\theta, X) = \log \left(f_X(X|\theta)\right)
\]

Note that log is monotonic, so maximizing \(l(\theta, X)\) is equivalent to maximizing \(L(\theta, X)\).
A maximum-likelihood (ML) estimator for $\theta$ from data $X$ has the form

$$\hat{\theta}_{ML} = \arg\max_{\theta} l(\theta, X) = \arg\max_{\theta} L(\theta, X)$$

Note that $\hat{\theta}_{ML}$ may not be unique because it is possible that more than one value of $\theta$ maximizes $l(\theta, X)$.

Example: Suppose $X_1, \ldots, X_n$ are iid normal RVs with variance $\sigma^2$ and unknown mean $\theta = \mu$,

$$l(\theta, X) = f_X(X|\theta) = f_{X_1}(X_1|\theta) \cdots f_{X_n}(X_n|\theta)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(X_1-\mu)^2}{2\sigma^2} \right\} \cdots \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(X_n-\mu)^2}{2\sigma^2} \right\}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}^n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\}$$

So,

$$L(\mu, X) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i^2 - 2\mu x_i + \mu^2)$$

and

$$\frac{d}{d\mu} L(\mu, X) = \frac{1}{2\sigma^2} \sum (2\mu - 2\mu) = 0$$

when $n\mu - \sum x_k = 0$. Thus $\hat{\mu}_{ML} = \frac{1}{n} \sum_{k=1}^n x_k$. 
Example: Now suppose \( x_1, \ldots, x_n \) are iid normal RVs with mean \( \mu \) and unknown variance \( \sigma^2 \).

\[
L(\Theta, \bar{X}) = L(\sigma^2, \bar{X}) = -\frac{n}{2} \log(2\pi\Theta) - \frac{2}{\Theta} \sum \left(X_k - \mu\right)^2
\]

So,

\[
\frac{d}{d\Theta} L(\Theta, \bar{X}) = -\frac{n}{2} \frac{1}{2\pi\Theta} 2\pi + \frac{1}{2\Theta^2} \sum \left(X_k - \mu\right)^2
\]

\[
= -\frac{n}{2\Theta} + \frac{1}{2\Theta^2} \sum \left(X_k - \mu\right)^2
\]

Assuming \( \Theta \neq 0 \), \( \frac{d}{d\Theta} L(\Theta, \bar{X}) = 0 \Rightarrow \hat{\Theta}_{ML} = \frac{1}{n} \sum_{k=1}^{n} \left(X_k - \mu\right)^2
\]

Example (Tale of two estimators): Take \( x_1, \ldots, x_n \) to be iid \( \sim U[0, b] \). The parameter to be estimated is \( \Theta = b \).

\[
l(\Theta, \bar{X}) = l(b, \bar{X}) = f_{X}(\bar{X}|b) = \begin{cases} \frac{n}{b} & 0 \leq x_k \leq b \\ 0 & \text{else} \end{cases}
\]

This is maximized by taking \( b \) to be as small as possible while still satisfying \( x_k \leq b \) for all \( k=1, \ldots, n \); i.e.,

\[
\hat{\Theta}_{ML} = \max(x_1, \ldots, x_n)
\]
The CDF of $\hat{\theta}_{ML}$ is

$$F_{\hat{\theta}}(v) = P(\sum_{i=1}^{n} x_i \leq v) = F_X(v)$$

Thus,

$$f_{\hat{\theta}}(v) = n F_X(v) f_X(v) = n \left(\frac{v}{b}\right)^{n-1} \left(\frac{1}{b}\right) = n \frac{v^{n-1}}{b^n}$$

and

$$E[\hat{\theta}_{ML}] = \int_0^b n \frac{v^n}{b^n} = \frac{n}{n+1} b$$

So, $\hat{\theta}_{ML}$ is biased with

$$\beta(\hat{\theta}_{ML}) = \left(\frac{n}{n+1} - 1\right) b = \frac{-1}{n+1} b$$

Note that $\beta \to 0$ as $n \to \infty$, so $\hat{\theta}_{ML}$ is asymptotically unbiased.

To get $\text{var}(\hat{\theta}_{ML})$, consider

$$E[\hat{\theta}^2] = \frac{n}{b^n} \int_0^b v^{n+1} dv = \frac{n}{n+2} b^2$$

So

$$\text{var}(\hat{\theta}_{ML}) = \frac{n}{(n+1)(n+2)} b^2$$
Compare this to the linear estimator
\[ \hat{\theta} = \frac{2}{n} \sum_{k=1}^{n} x_k, \]
for which
\[ E[\hat{\theta}] = \frac{2}{n} n \left( \frac{b}{2} \right) = b. \]
This estimator of \( b \) is unbiased. Its variance is four times that of the sample mean estimator; i.e.,
\[ \text{var}(\hat{\theta}) = 4 \frac{\text{var}(x)}{n} = 4 \frac{b^2}{12n} = \frac{1}{3n} b^2. \]
While biased, \( \hat{\theta}_{ML} \) has variance that decreases quadratically in \( n \). The variance of \( \hat{\theta} \) decreases only linearly with \( n \).
Bayesian Estimation

In Bayesian estimation, the parameter $\Theta$ is taken to be a RV $\Theta$ with a known PDF $f_\Theta(\Theta)$, often called the "prior" PDF.

The Bayesian estimator for $\Theta$ is

$$\hat{\Theta}_B = E[\Theta|X]$$

Calculation of this expectation is via Bayes' rule:

$$f_\Theta(\Theta|X) = \frac{f_X(X|\Theta)f_\Theta(\Theta)}{f_X(X)}$$

Both factors in the numerator (i.e., the likelihood function and the prior PDF of $\Theta$) are assumed known. The denominator is obtained by marginalization, i.e.,

$$f_X(X) = \int f_{X,\Theta}(X, \Theta) d\Theta$$

where
\[ f_{x|\theta}(x, \theta) = f_x(x|\theta)f_\theta(\theta) \]

**Example:** The probability of a coin toss yielding \( H \) is \( p \).

**Data:** \( E = \{ n \text{ out of } n \text{ independent tosses yield } H \} \)

**Prior:** Take \( p \sim U[0, 1] \),

\[ f_p(p|E) = \frac{P(E|p)f_p(p)}{P(E)} \]

with \( P(E|p) = p^n \) and \( f_p(p) = \frac{1}{p} \) for \( 0 \leq p \leq 1 \) otherwise. So,

\[ f_p(p|E) = \frac{p^n}{P(E)} \quad 0 \leq p \leq 1 \]

For proper normalization,

\[ 1 = \int_0^1 f_p(p|E) \, dp = \frac{1}{P(E)} \int_0^1 p^n \, dp = \frac{1}{(n+1)P(E)} \]

and hence \( P(E) = \frac{1}{n+1} \) and \( f_p(p|E) = (n+1)p^n \), \( 0 \leq p \leq 1 \).

The Bayesian estimator is thus

\[ \hat{p} = E[p|E] = \int_0^1 p f_p(p|E) \, dp = \frac{n+1}{n+2} \]

Note that \( \hat{p} \to 1 \) as \( n \to \infty \), as expected. Also, it is possible to compute \( P(E) \) directly if desired.
\[ P(E) = \int_{0}^{1} P(E|p) f_{\theta}(p) \, dp = \int_{0}^{1} p^{n} \, dp = \frac{1}{n+1}. \]

Example: Let \( x_1, \ldots, x_n \) be i.i.d. \( N[\mu, 1] \) RVs. The parameter to be estimated is \( \theta = \mu \). Take the prior PDF of \( \theta \) to be \( N[0, \sigma^2] \). Then

\[ f_{x}(x|\theta) = \frac{1}{(2\pi)^{n/2}} \exp \left\{ -\frac{1}{2} \sum_{k=1}^{n} (x_k - \theta)^2 \right\} \]

and

\[ f_{\theta}(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2} \frac{\theta^2}{\sigma^2} \right\} \]

Denoting

\[ C = \frac{1}{(2\pi)^{n/2}} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \frac{1}{f_{x}(x)}, \]

\[ f_{\theta}(\theta|x) = C \exp \left\{ -\frac{1}{2} \sum_{k=1}^{n} (x_k - \theta)^2 \right\} \exp \left\{ -\frac{1}{2} \frac{\theta^2}{\sigma^2} \right\} \]

\[ = C \exp \left\{ -\frac{1}{2} \left[ \sum_{k=1}^{n} x_k^2 - 2\theta \sum_{k=1}^{n} x_k + n\theta^2 + \frac{\theta^2}{\sigma^2} \right] \right\} \]

\[ = C \exp \left\{ -\frac{1}{2} \sum_{k=1}^{n} x_k^2 \right\} \exp \left\{ -\frac{1}{2} \left[ (n+1)\theta^2 - 2\theta \sum_{k=1}^{n} x_k \right] \right\} \]

\[ = C \exp \left\{ -\frac{1}{2} \sum_{k=1}^{n} x_k^2 \right\} \exp \left\{ -\frac{1}{2} \left( n+1 \right) \frac{\theta^2}{\sigma^2} \right\} \]
Completing the square in $\theta$,

$$f_{\theta}(\theta|x) = C \exp \left\{ -\frac{1}{2} \sum x_k^2 \right\} \exp \left\{ \frac{1}{2} \left( \frac{\sum x_k^n}{(n+1)^2} \right) \right\}$$

$$\times \exp \left\{ -\frac{1}{2} \left( \frac{1}{(n+1)^2} \right) \right\} \left[ \theta - \frac{\sum x_k}{(n+1)^2} \right]^2$$

This is a normal PDF with mean $\frac{1}{(n+1)^2} \sum x_k$, hence

$$\hat{\theta} = \frac{1}{n+\frac{1}{\rho^2}} \sum_{k=1}^n x_k$$

Note that as $\rho^2 \to \infty$, the prior becomes uninformative and

$$\hat{\theta} \to \frac{1}{n} \sum_{k=1}^n x_k$$

i.e., the sample mean. As $\rho^2 \to 0$, the prior dominates the data and $\hat{\theta} \to 0$ regardless of $x$.

It is helpful to observe that $f_{\theta}(x)$ and several other normalizing values never had to be calculated. Because $f_{\theta}(\theta|x)$ was clearly Gaussian in this example, only the exponential containing $\theta$ was needed to deduce the mean.
Detecting a Known Signal in Additive Gaussian Noise

Consider the detection problem

\[ H_0 : X = N \]
\[ H_1 : \tilde{X} = S + N \]

\( S \in \mathbb{R}^n \) known
\( N \sim \mathcal{N}[0, \sigma^2 I] \)
with \( \sigma^2 \) known

With \( P(h_i) \) known, the Bayesian decision rule gives

\[ \rightarrow H_1 : \left\{ X \mid \frac{f_X(X|H_1)}{f_X(X|H_0)} > \frac{P(H_0)}{P(H_1)} \right\} \]

More generally, the Neyman-Pearson lemma holds that a decision rule of the form

\[ \rightarrow H_1 : \left\{ X \mid \frac{f_X(X|H_1)}{f_X(X|H_0)} > T \right\} \]

will maximize \( P_d \) for any fixed \( P_f \). Thus, the likelihood ratio

\[ \Lambda(X) = \frac{f_X(X|H_1)}{f_X(X|H_0)} \]
is a crucial quantity in signal detection.

For the problem under consideration,

\[ f_{x|H_0}(x) = C \exp \left\{ -\frac{1}{2\sigma^2} x^T x \right\} \]

\[ f_{x|H_1}(x) = C \exp \left\{ -\frac{1}{2\sigma^2} (x - s)^T (x - s) \right\} \]

with \( C = (2\pi)^{-n/2} \). Thus

\[ \Lambda(x) = \exp \left\{ -\frac{1}{2\sigma^2} (x^T s + s^T x - s^T s) \right\} \]

Note that \( x^T s = s x^T \) is the inner product of \( x \) with \( s \) and \( s^T s = \| s \|^2 \) is the energy of the signal \( s \). So,

\[ \rightarrow H_1 = \left\{ \Lambda(x) > T \right\} = \left\{ \exp \left\{ -\frac{1}{2\sigma^2} x^T s - \frac{1}{2\sigma^2} \| s \|^2 \right\} > T \right\} \]

\[ = \left\{ \frac{1}{\sigma^2} x^T s - \frac{1}{\sigma^2} \| s \|^2 > \log T \right\} \]

\[ = \left\{ x^T s > \sigma^2 \log T - \frac{1}{2} \| s \|^2 \right\} = \left\{ x^T s > T' \right\} \]

where \( T' = \sigma^2 \log T + \frac{1}{2} \| s \|^2 \).
Detector Performance

The detector for a known signal vector $S$ in additive white gaussian noise of known variance $\sigma^2$ has the form

$$H_1 = \{ S^T X > T' \} = \{ \rho > T' \}$$

with the detection statistic $\rho = S^T X$.

Under $H_0$, $X \sim N$ and hence

$$\rho \sim \sum_{k=1}^{n} s_k n_k \sim N[0, \sigma^2 \sum s_k^2]$$

Under $H_1$, $X \sim S + N$ and

$$\rho \sim S^T (S + N) = \|S\|^2 + S^T N \sim N[\|S\|^2, \sigma^2 \sum s_k^2]$$

So

$$P_F = P(\rho > T' | H_0) = 1 - \Phi \left( \frac{T'}{\sigma \|S\|} \right)$$

and

$$P_d = P(\rho > T' | H_1) = 1 - \Phi \left( \frac{T' - \|S\|^2}{\sigma \|S\|^2} \right)$$
Complex Signal and Colored Noise

Now consider a known complex signal vector $s \in \mathbb{C}^n$ in additive complex Gaussian noise $\tilde{n} \sim \mathbb{CN}[0, \Sigma_n]$, with known covariance matrix $\Sigma_n$. The detection problem is

$$H_0: \tilde{X} = \tilde{n} \quad \quad H_1: \tilde{X} = s + \tilde{n}$$

for which

$$\Lambda(\tilde{X}) = \frac{\mathbf{C} \exp \left\{ - (\tilde{X} - s)^\dagger \Sigma_n^{-1} (\tilde{X} - s) \right\}}{\mathbf{C} \exp \left\{ - \tilde{X}^\dagger \Sigma_n^{-1} \tilde{X} \right\}}$$

$$= \exp \left\{ s^\dagger \Sigma_n^{-1} \tilde{X} + \tilde{X}^\dagger \Sigma_n^{-1} s - s^\dagger \Sigma_n^{-1} s \right\}$$

$$= \exp \left\{ 2 \text{Re}(s^\dagger \Sigma_n^{-1} \tilde{X}) - s^\dagger \Sigma_n^{-1} s \right\}$$

The detector thus has the form

$$\Rightarrow H_1 = \left\{ \theta > T \right\}$$

where

$$\theta = \text{Re}(s^\dagger \Sigma_n^{-1} \tilde{X})$$

which is $N[0, \frac{s^\dagger \Sigma_n^{-1} s}{2}]$ distributed under $H_0$, and

$N[s^\dagger \Sigma_n^{-1} s, \frac{s^\dagger \Sigma_n^{-1} s}{2}]$ distributed under $H_1$. 
Central Limit Theorem

Characteristic Function

Recall that, for a RV $z$,

$$\mathcal{F}_z(w) = \mathbb{E}[e^{-iwz}] = \int_{-\infty}^{\infty} f_z(z) e^{-iwz} \, dz$$

Thus,

$$\frac{d}{dw} \mathcal{F}_z(w) = -i \int_{-\infty}^{\infty} zf_z(z) e^{iwz} \, dz$$

so that $E[z] = \frac{1}{(-i)} \frac{d}{dw} \mathcal{F}_z(0)$. Similarly,

$$\frac{d^2}{dw^2} \mathcal{F}_z(w) = (-i)^2 \int_{-\infty}^{\infty} z^2 f_z(z) e^{-iwz} \, dz$$

and $E[z^2] = \frac{1}{(-i)^2} \frac{d^2}{dw^2} \mathcal{F}_z(z)$.

If $z \sim N[0, \sigma^2]$,

$$\mathcal{F}_z(w) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2\sigma^2} + i\sigma w} \, dz = \exp\left\{ -\frac{1}{2} \sigma^2 w^2 \right\}$$

where the integral is evaluated by completing the square in $z$. 
Central Limit Derivation

Suppose $x_1, \ldots, x_n$ are iid with mean $\mu$ and variance $\sigma^2$. Define

$$y = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} (x_k - \mu)$$

Note that $E[y] = 0$ and $\text{var}(y) = \left(\frac{1}{\sqrt{n}}\right)^2 \sum_{k=1}^{n} \sigma^2 = \sigma^2$.

The characteristic function of $y$ is

$$\Phi_y(w) = E\left[\exp\left\{-iw \frac{1}{\sqrt{n}} \sum_{k=1}^{n} (x_k - \mu)\right\}\right]$$

$$= \prod_{k=1}^{n} \Phi\left[\exp\left\{-iw \frac{1}{\sqrt{n}} (x_k - \mu)\right\}\right]$$

$$= \left[\Phi_{x}(\frac{1}{\sqrt{n}} w)\right]^n$$

where $\Phi_x(w) = E[\exp\{-iw(x - \mu)\}]$, i.e., the characteristic function of $x - \mu$.

Assume

$$\Phi(w) = \Phi(0) + \Phi'(0) w + \frac{\Phi''(0)}{2} w^2 + \Theta(w)|w|^3$$

with $|\Theta(w)| < B$ for all $w$. Note that $E[x - \mu] = 0$ and $E[(x - \mu)^2] = \sigma^2$, so

$$\Phi(0) = 1 \quad \Phi'(0) = 0 \quad \Phi''(0) = -\sigma^2$$
Thus,
\[ \mathbb{E}(w) = 1 - \frac{1}{2} \sigma^2 w^2 + \theta(w) |w|^3 \]
and
\[ \mathbb{E}_Y(w) = \left[ \mathbb{E}(\frac{1}{n}w) \right]^n = \left( 1 - \frac{1}{2n} \sigma^2 w^2 + \theta\left(\frac{w}{\sqrt{n}}\right)n^{-3/2} |w|^3 \right)^n \]
Recall that
\[ \lim_{n \to \infty} \left( 1 + \frac{v}{n} \right)^n = e^v \]
and set \( v = -\frac{1}{2} \sigma^2 w^2 + \theta\left(\frac{w}{\sqrt{n}}\right)n^{-3/2} |w|^3 \). Note that
\[ \lim_{n \to \infty} v = -\frac{1}{2} \sigma^2 w^2 \]
so
\[ \lim_{n \to \infty} \mathbb{E}_Y(w) = \lim_{n \to \infty} \left( 1 + \frac{v}{n} \right)^n = \exp\left\{ -\frac{1}{2} \sigma^2 w^2 \right\} \]
i.e., \( \mathbb{E}_Y \to \mathbb{E}_z \) with \( z \sim N[0, \sigma^2] \) as \( n \to \infty \).

**Theorem (Central Limit):** If \( x_1, x_2, \ldots \) are iid with mean \( \mu \) and variance \( \sigma^2 \) and the assumption on the characteristic function \( \mathbb{E}^i \) of \( x - \mu \) holds, then
\[ \sqrt{n} \sum_{k=1}^{n} (x_k - \mu) \]
is asymptotically \( N[0, \sigma^2] \) distributed as \( n \to \infty \).
Remarks: 1. The assumptions made above can be weakened, but the derivation becomes more difficult; e.g.,

\[ y = n^{-\frac{1}{2}} \sum_{K=1}^{n} \left[ x_k - E[x_k] \right] \]

approaches gaussian as \( n \to \infty \) provided \( 0 < A \leq \text{var}(x_k) \leq B < \infty \) for all \( K \) and that \( E[(x_k - E[x_k])^3] \) is bounded by some \( C < \infty \) for all \( K \). This uses independence, but the \( x_k \) need not be identically distributed.

2. If \( y = \sum_{K=1}^{n} x_k \) with the \( x_k \) iid, then \( y \) is approximately \( N[\mu, n\sigma^2] \) for large \( n \).

3. The sample mean \( \mu = \frac{1}{n} \sum_{K=1}^{n} x_k \) is approximately \( N[\mu, \frac{\sigma^2}{n}] \) for large \( n \).

Example: \( z = \sum_{K=1}^{20} x_k \) with \( x_k \) iid and \( U[-1,1] \).

Then \( z \) is approximately \( N[0, \frac{20}{3}] \) so that

\[ P(z > 4) \approx 1 - \Phi\left( \frac{2\sqrt{3}}{\sqrt{20}} \right) \approx 0.061 \]
Spectral Density

If \( x(\cdot) \) is WSS, its spectral density is

\[
S_x(\omega) = \sum_{k=-\infty}^{\infty} c_x(k) e^{-i2\pi k \omega}
\]

i.e., the (discrete-time) Fourier transform of the autocovariance function.

\( S_x(\cdot) \) has the following properties:

1. \[
S_x(\omega + 2\pi) = \sum_{k} c_x(k) e^{-i(\omega + 2\pi)k} = \sum_{k} c_x(k) e^{-i\omega k} e^{-i2\pi k} = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_x(\eta) e^{i\omega \eta} d\eta
\]

   i.e., \( S_x(\cdot) \) is periodic with period \( 2\pi \).

2. \[
S_x(-\omega) = \sum_{k} c_x(k) e^{i\omega k} = \sum_{k} c_x(-k) e^{i\omega k} = \sum_{j} c_x(j) e^{-i\omega j} = S_x(\omega)
\]

   i.e., \( S_x(\cdot) \) is even.

3. \( S_x(\omega) \geq 0 \) for all \( \omega \) (to be shown later).
Example (White process): Take \( x_0, x_1, \ldots \) to be iid with mean \( \mu \) and variance \( \sigma^2 \), and \( x(k) = x_k \). Then
\[
\mu_x(k) = E[x(k)] = E[x_k] = \mu
\]
for all \( k \). And,
\[
c_x(k_1, k_2) = \text{cov}(x(k_1), x(k_2)) = \begin{cases} 
\sigma^2 & k_1 = k_2 \\
0 & k_1 \neq k_2 
\end{cases}
\]
i.e., \( c_x(k_1, k_2) = \sigma^2 \) if \( k_1 = 0 \), \( c_x(k_1, k_2) = 0 \) if \( k_1 \neq 0 \). So this process is also WSS.

Notes:
(1) If \( x(\cdot) \) is WSS, it is common to write
\[
c_x(K) = c_x(k_1, k_2)
\]
for any \( k_1 \) and \( k_2 \) with \( k_2 - k_1 = K \).

(2) \( c_x(k_1, k_1) = \text{cov}(x(k_1), x(k_2)) = \text{cov}(x(k_1), x(k_1)) = c_x(k_1, k_1) \)

In particular, if \( x(\cdot) \) is WSS, then
\[
c_x(K) = c_x(-K).
\]
Discrete-time Stochastic Processes

Consider an infinite sequence of RVs \( x_{-1}, x_0, x_1, \ldots \) indexed by the integers \( \mathbb{Z} \). A function \( x(\cdot) \) on \( \mathbb{Z} \) that assigns \( k \) to \( x(k) = x_k \) for each \( k \) is called a discrete-time stochastic process.

Notes: 1. It is typical to think of \( k \) as representing time and \( x(\cdot) \) as a discrete-time random signal. In various applications, \( k \) may represent other quantities, such as spatial position, frequency, etc.

2. The RVs \( x_k \) may be real or complex (or even random vectors), and they can be of continuous, discrete, or mixed type. In what follows, the default assumption will be that they are real RVs.

3. Realizing each of the \( x_k \) gives a deterministic function \( x(\cdot) \) called a realization or sample path of \( x(\cdot) \).

Example: The \( x_k \) are iid binary RVs. A sample path looks like:

\[
\begin{array}{c}
\cdots \quad x(-1) \quad x(0) \quad x(1) \cdots
\end{array}
\]
Mean and Autocovariance

The mean \( \mu_x(t) \) of a process \( x(t) \) has values

\[
\mu_x(k) = E[x(k)] = E[x_k]
\]

For integers \( k_1 \) and \( k_2 \), the autocovariance function of the process \( x(t) \) has values

\[
c_x(k_1, k_2) = \text{cov}(x_{k_1}, x_{k_2})
\]

Example: With the RVs \( \ldots, x_{-1}, x_0, x_1, \ldots \) iid an \( \sim U[a, b] \),

\[
\mu_x(k) = E[x_k] = \frac{b+a}{2}
\]

and

\[
c_x(k_1, k_2) = \text{cov}(x_{k_1}, x_{k_2}) = \begin{cases} \frac{(b-a)^2}{12} & k_1 = k_2 \\ 0 & \text{otherwise} \end{cases}
\]

With the \( y_k \) independent and \( \sim U[0, k_1] \),

\[
\mu_{y}(k) = \frac{k_1}{2}
\]

and

\[
c_{y}(k_1, k_2) = \begin{cases} \frac{1k_1^2}{12} & k_1 = k_2 = k \\ 0 & \text{otherwise} \end{cases}
\]
Wide-sense Stationarity

A discrete-time process $x(t)$ is wide-sense stationary (WSS) if

(a) $\mu_x(k)$ is constant, and
(b) $c_x(k_1, k_2)$ depends only on $k = k_2 - k_1$.

Condition (b) can be interpreted to mean that the degree to which two RVs in the sequence depends on how far apart they are but not where they occur in an absolute sense; e.g., $\text{cov}(x(0), x(3)) = \text{cov}(x(147), x(150))$ because the separation of time indices in both cases is $k = 3$.

Example (constant process): Define $\mathbf{x}(k) \sim \mathcal{N}[0, 1]$ for all $k$. Then

$\mu_x(k) = E[x(k)] = E[x] = 0$

for all $k$, so condition (a) is satisfied. Also,

$c_x(k_1, k_2) = \text{cov}(x(k_1), x(k_2)) = \text{cov}(x, x) = 1$

for every value of $k = k_2 - k_1$, so $x(t)$ is WSS.
Filtering of WSS Processes

Consider the situation in which a discrete-time WSS process \( x(\cdot) \) is input to a linear time-invariant system with impulse response \( h(\cdot) \). The output process \( y(\cdot) \) has values

\[
y(k) = [x * h](k) = \sum_{n=-\infty}^{\infty} h(n) x(k-n)
\]

i.e., the convolution of \( x(\cdot) \) and \( h(\cdot) \).

Pictorially:

\[
\begin{array}{c}
x(\cdot) \\
\downarrow \\
h(\cdot) \\
\uparrow \\
y(\cdot) = [x * h](\cdot)
\end{array}
\]

Mean

In general,

\[
\mu_y(k) = E[y(k)] = E\left[\sum_n h(n) x(k-n)\right]
\]

\[
= \sum_n h(n) E[x(k-n)]
\]

\[
= \sum_n h(n) \mu_x(k-n) = [\mu_x * h](k)
\]
When \( x(\cdot) \) is WSS, \( \mu_x(k) = \mu_x \) is constant.

So

\[
\mu_y(k) = \mu_x \sum_n h(n)
\]

and \( \mu_y \) is also constant.

Autocovariance

The autocovariance function of \( y(\cdot) \) is given by

\[
c_y(k_1, k_2) = \text{cov}(y(k_1), y(k_2))
\]

\[
= E \left[ (y(k_1) - \mu_y(k_1))(y(k_2) - \mu_y(k_2)) \right]
\]

\[
= E \left[ \sum_n h(n)(x(k_1 - n) - \mu_x(k_1 - n)) \sum_m h(m)(x(k_2 - m) - \mu_x(k_2 - m)) \right]
\]

\[
= \sum_n \sum_m h(n)h(m)C_x(k_1 - n, k_2 - m)
\]

If \( x(\cdot) \) is WSS, \( C_x(k_1 - n, k_2 - m) = C_x(k - n + m) \) with \( k = k_2 - k_1 \). So, in this case, \( c_y(k_1, k_2) \) depends only on \( k_1 \) thus \( y(\cdot) \) is WSS with

\[
c_y(k) = \sum_n \sum_m h(n)h(m)C_x(k - n + m)
\]
Spectral Density

With \( x(\cdot) \) WSS, \( y(\cdot) = [x^* h](\cdot) \) is also WSS with spectral density

\[
S_y(w) = \sum_{k=-\infty}^{\infty} C_y(k) e^{-i\omega k}
\]

\[
= \sum_{k} \sum_{n} h(n) h(m) C_x(k-n+m) e^{-i\omega k}
\]

\[
= \sum_{n} \sum_{m} h(n) h(m) \left( \sum_{k} C_x(k-n+m) e^{-i\omega k} \right)
\]

\[
= \sum_{n} \sum_{m} h(n) h(m) \sum_{l} C_x(l) e^{-i\omega(l+n-m)}
\]

\[
= \sum_{n} \sum_{m} h(n) h(m) e^{-i\omega n} e^{i\omega m} S_x(w)
\]

\[
= H(w) H^*(w) S_x(w) = |H(w)|^2 S_x(w)
\]

where \( H(w) = \sum_{k=-\infty}^{\infty} h(k) e^{-i\omega k} \) is the frequency response of the LTI system.

Example: The RVs \( x_k \) are iid and \( N[0, \sigma^2] \) for \( k \in \mathbb{Z} \).
Define \( x(k) = x_k \). Then \( E[x(k)] = 0 \) for all \( k \) and

\[
C_x(k) = \begin{cases} \sigma^2 & k = 0 \\ 0 & k \neq 0 \end{cases} = S(k)
\]
Thus,
\[ S_x(w) = \sum_k \sigma_k^2 S(k) e^{-i\omega k} = \sigma^2 \]
for all \( w \) (\( x(\cdot) \) is a white process). If a LTI system has impulse response
\[ h(k) = \begin{cases} 1 & k = 0, 1 \\ 0 & \text{otherwise} \end{cases} \]
then
\[ H(w) = \sum_k h(k) e^{-i\omega k} = 1 + e^{-i\omega} \]
\[ = (1 + \cos(\omega)) - i \sin(\omega) \]
So,
\[ |H(w)|^2 = 2 + 2\cos(\omega) \]
If \( x(\cdot) \) is passed through this system, the output process \( y(\cdot) \) will have spectral density
\[ S_y(w) = |H(w)|^2 S_x(w) = 2\sigma^2 (1 + \cos(w)) \]
\[ \text{Hence } y(\cdot) \text{ is a low-pass process.} \]
Moving Average (MA) Processes

Consider the causal FIR system pictured:

\[ x(k) \xrightarrow{b_0} \xrightarrow{b_1} \xrightarrow{z^{-1}} \xrightarrow{\ldots} \xrightarrow{b_m} \xrightarrow{z^{-m}} y(k) \]

The corresponding system equation is

\[ y(k) = b_0 x(k) + \ldots + b_m x(k-m) \]

and the transfer function is, thus,

\[ H(z) = \frac{Y(z)}{X(z)} = b_0 + b_1 z^{-1} + \ldots + b_m z^{-m} \]

i.e., a polynomial of degree \( m \) in \( z^{-1} \).

**Def:** An \( m \)-th-order moving average (MA) process is the output of a causal FIR filter (as above) when the driving signal \( x(t) \) is a zero-mean, finite variance white process.
Notes:  

(1) If \( x(\cdot) \) has variance \( \sigma^2 \),

\[
S_y(w) = |b_0 + b_1 e^{-i\omega} + \cdots + b_m e^{-im\omega}|^2 \sigma^2
\]

(2) The filter weights \( b_0, \ldots, b_m \) are called the "MA coefficients" of \( y(\cdot) \).

**Autoregressive (AR) Processes**

Consider the recursive linear system pictured:

The corresponding system equation is

\[
y(k) + a_1 y(k-1) + \cdots + a_n y(k-n) = x(k)
\]

and the transfer function is thus

\[
\tilde{H}(z) = \frac{\tilde{Y}(z)}{\tilde{X}(z)} = \frac{1}{1 + a_1 z^{-1} + \cdots + a_n z^{-n}}
\]
Recall that the filter is stable if all poles of $\tilde{H}(z)$ lie within the unit disc in the $z$-plane.

**Def.** An $n^{th}$-order autoregressive (AR) process is the output of a stable recursive filter (as above) when the driving signal $x(t)$ is a zero-mean, finite-variance white process.

**Notes:**

1. If $x(t)$ has variance $\sigma^2$,

\[
S_y(w) = |H(w)|^2 \quad S_x(w) = \sigma^2 |\tilde{H}(e^{i\omega})|^2
\]

\[
= \frac{\sigma^2}{|1 + a_1 e^{-i\omega} + \ldots + a_p e^{-i(p-1)\omega}|^2}
\]

2. The filter weights $a_1, \ldots, a_p$ are called the "AR coefficients" of $y(t)$.
ARMA Processes

Consider a causal FIR filter with transfer function

\[ \tilde{H}_{MA}(z) = b_0 + \cdots + b_m z^{-m} \]

and a stable recursive filter with transfer function

\[ \tilde{H}_{AR}(z) = \frac{1}{1 + a_1 z^{-1} + \cdots + a_n z^{-n}} \]

Cascading these filters

\[ x(k) \rightarrow \tilde{H}_{MA}(z) \rightarrow \tilde{H}_{AR}(z) \rightarrow y(k) \]

yields a new filter

\[ x(k) \rightarrow \tilde{H}_{ARMA}(z) \rightarrow y(k) \]

with transfer function
\[ H_{\text{ARMA}}(z) = \frac{b_0 + \cdots + b_m z^{-m}}{1 + a_1 z^{-1} + \cdots + a_n z^{-n}} \]

**Def:** An autoregressive moving-average (ARMA) process is the output of a cascade filter (as above) when the driving signal is a zero-mean, finite-variance, white process.

**Remark:** If \( x(\cdot) \) has variance \( \sigma^2 \),

\[ S_y(w) = \sigma^2 \left| \frac{b_0 + b_1 e^{i\omega} + \cdots + b_m e^{-i\omega}}{1 + a_1 e^{-i\omega} + \cdots + a_n e^{-i\omega}} \right|^2 = \sigma^2 \frac{B_0 + B_1 \cos(w) + \cdots + B_m \cos(mw)}{1 + A_1 \cos(w) + \cdots + A_n \cos(nw)} \]

where the last expression is obtained by application of trigonometric identities.

**Spectral Factorization**

Given a desired spectral density for an ARMA process \( y(\cdot) \), spectral factorization of \( S_y(w) \) allows identification of a filter to produce \( y(\cdot) \) from zero-mean white noise \( x(\cdot) \).
Example: The desired spectral density for \( y(t) \) is

\[
S_y(w) = \frac{405 + 324 \cos(w)}{94 + 42 \cos(w) - 36 \cos(2w)}
\]

With \( z = e^{iw} \), \( \cos(w) = \frac{z + z^{-1}}{2} \), \( \cos(2w) = \frac{z^2 + z^{-2}}{2} \), so the filter must satisfy

\[
\hat{H}(z) \hat{H}(z^{-1}) = \frac{405 + 162z + 162z^{-1}}{94 + 21z + 21z^{-1} - 18z^2 - 18z^{-2}} = \frac{\hat{N}(z)}{\hat{D}(z)}
\]

Factoring numerator and denominator separately,

\[
\hat{N}(z) = z^{-1}(162z^2 + 405z + 162) = 162z^{-1}(z + \frac{1}{2})(z + 2)
\]

\[
= 324 (z + \frac{1}{2})(z^{-1} + \frac{1}{2})
\]

\[
\hat{D}(z) = 81 (z + \frac{2}{3})(z^{-1} + \frac{2}{3})(z - \frac{1}{2})(z^{-1} - \frac{1}{3})
\]

So,

\[
\hat{S}_y(z) = \sigma^2 |\hat{H}(z)|^2 = \sigma^2 \hat{H}(z) \hat{H}(z^{-1})
\]

\[
= \sigma^2 \frac{4(z + \frac{1}{2})(z^{-1} + \frac{1}{2})}{(z + \frac{2}{3})(z^{-1} + \frac{2}{3})(z - \frac{1}{2})(z^{-1} - \frac{1}{3})}
\]

Taking \( \sigma^2 = 1 \) and choosing factors for causality and stability,

\[
\hat{H}(z) = \frac{2(z^{-1} + \frac{1}{2})}{(z + \frac{2}{3})(z - \frac{1}{2})} = \frac{1 + 2z^{-1}}{z^2(1 + \frac{1}{3}z^{-1} - \frac{2}{3}z^{-2})}
\]
Thus the MA coefficients are $b_0 = 1$, $b_1 = 2$ and the AR coefficients are $a_1 = \frac{1}{3}$, $a_2 = -\frac{3}{9}$.

Note that

$$S_y(w) = o^2 |H(w)|^2 = \frac{1 + 2e^{-iw}}{|e^{iw}|^2 - \frac{1}{3} |e^{iw} + \frac{1}{3}|^2 |e^{iw} - \frac{1}{3}|^2}$$

$$= \frac{5 + 4 \cos w}{\left(\frac{13}{9} + \frac{4}{3} \cos w\right)\left(\frac{10}{9} \cos w - \frac{2}{3}\right)} = \frac{5 + 4 \cos w}{\frac{120}{81} + \frac{42}{81} \cos w - \frac{72}{81} \cos^2 w}$$

$$\cos^2(w) = \frac{1 + \cos 2w}{2} = \frac{\cos(w)}{2}$$

**Whitening Filter**

If a process $x(t)$ is observed to have rational spectral density $S_x$, form a filter with transfer function $\tilde{H}(z)$ satisfying

$$\tilde{H}(z) \tilde{H}(z^{-1}) = \left[\tilde{S}_x(z)\right]^{-1}$$

by spectral factorization. Filtering $x(t)$ with this filter yields a process $y$ with

$$S_y(w) = |H(w)|^2 S_x(w) = \left[S_x(w)\right]^{-1} S_y(w) = 1$$

i.e., $y(t)$ is a white process.
Continuous-time Processes

The theory of continuous-time WSS processes generally parallels the theory of discrete-time WSS processes. A synopsis follows:

Definitions:

1. A continuous-time stochastic process \( x(t) \) is a family of RVs parameterized by the real numbers; i.e., \( x(t) \) is a RV for each \( t \in \mathbb{R} \).

2. The mean of \( x(t) \) is defined by

\[
\mu_x(t) = \mathbb{E}[x(t)] \quad t \in \mathbb{R}
\]

3. The autocovariance function of \( x(t) \) is defined by

\[
c_x(t_1, t_2) = \text{cov}(x(t_1), x(t_2)) \quad t_1, t_2 \in \mathbb{R}
\]

4. The process \( x(t) \) is WSS if

(i) \( \mu_x(t) \) is constant, and
(ii) \( c_x(t_1, t_2) \) depends only on \( \tau = t_2 - t_1 \).

5. The spectral density of a WSS process \( x(t) \) is

\[
S_x(\omega) = \int_{-\infty}^{\infty} c_x(t)e^{-i\omega t} \, dt
\]
Notes: 1. For a continuous-time WSS process $x(t)$, $S_x(w) > 0$ for all $w \in \mathbb{R}$. But $S_x(t)$ will generally not be periodic.

2. A WSS process $x(t)$ is white if

$$C_x(t) = \sigma^2 \delta(t)$$

In this case $\delta$ is the Dirac delta symbol and

$$S_x(w) = \sigma^2$$ for all $w \in \mathbb{R}$.

Filtering

If $x(t)$ is passed through a LTI system with impulse response $h(t)$, the output is

$$y(t) = [h^* * x](t) = \int_{-\infty}^{\infty} h(u)x(t-u)\,du$$

So,

$$E[y(t)] = \int_{-\infty}^{\infty} h(u)E[x(t-u)]\,du = [h \ast \mu_x](t)$$

and thus, if $x(t)$ is WSS, $\mu_y(t) = \mu_x \int h(u)\,du$. In this case $\mu_y(t)$ is constant.
Also,
\[ c_y(t_1, t_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u)h(v) c_x(t_1-u, t_2-v) \, du \, dv \]

If \( x(\cdot) \) is WSS

\[ c_y(t_1, t_2) = \int_{\mathbb{R}^2} h(u)h(v) c_x(t_1-u, t_2-v+u) \, du \, dv = c_x(r) \]

with \( r = t_2 - t_1 \). So \( y(\cdot) \) is WSS. Direct calculation gives

\[ S_y(w) = |H(w)|^2 S_x(w) \]

with \( H(w) = \int_{-\infty}^{\infty} h(t)e^{-jwt} \, dt \).

**Complex Processes**

The theory developed above carries over almost without alteration to complex WSS processes, both in the discrete-time and continuous-time settings. One must be mindful that, in the complex case,

\[ \text{cov}(x(t_2), x(t_1)) = \text{cov}(x(t_1), x(t_2)^*) \]

so that, with \( x(\cdot) \) WSS, \( c_x(r) = c_x^*(r) \); i.e., \( c_x(\cdot) \) is conjugate symmetric rather than even.
Markov Processes

For a discrete-time process \( x(t) \), denote the PDF of the RV \( x(k) \) by \( f_x(x, k) \). For any \( k \), let \( k_n < k_{n-1} < \ldots < k_1 < k \). The RVs

\[
x(k), x(k_1), \ldots, x(k_n)
\]

will generally not be independent. The process \( x(t) \) is a Markov process if

\[
f_x(x, k | x(k_1) = x_1, \ldots, x(k_n) = x_n) = f_x(x, k | x(k_1) = x_1)
\]

i.e., only the most recent realized value of \( x(t) \) affects the conditional density of \( x(k) \).

Note that, for a Markov process,

\[
f_x(x, k) = \int_{-\infty}^{\infty} f_x(x, k | x, x_1, k_1) \, dx_1
\]

\[
= \int_{-\infty}^{\infty} f_x(x, k | x(k_1) = x_1) f_x(x, k_1) \, dx_1
\]

In particular, taking \( k_1 = k-1 \),

\[
f_x(x, k) = \int_{-\infty}^{\infty} f_x(x, k | x(k-1) = x_1) f_x(x, k-1) \, dx_1
\]
So, if one knows the PDF of $x(t)$ at any time, its PDF at any future time can be calculated by successive applications of this rule. The crucial element in this rule is the transition probability density

$$f_x(x, k \mid x(k-1) = x_i)$$

The Markov process $x(t)$ is homogenous if its transition PDFs are all the same (do not depend on $k$); i.e.,

$$f_x(x, k \mid x(k-1) = x_i) = g(x, x_i)$$

**Markov Chains**

Suppose each RV $x(k)$ in a discrete-time Markov process is a discrete-type RV taking on values in some countable set $R$. Since $R$ is in one-to-one correspondence with the positive integers $\mathbb{Z}_+ = \{1, 2, \ldots\}$ or a finite subset $\{1, 2, \ldots, n\}$ of $\mathbb{Z}_+$, it is typical just to label the possible values $x(k)$ can assume by positive integers.
A discrete-time, discrete-valued Markov process, as just described, is known as a Markov chain. Its PDF (or PMF) at time $k$ is simply a list of probabilities:

$$f_x(m,k) = \mathbb{P}(\xi x(k) = m) \quad m = 1, 2, \ldots$$

When the set of possible values each $x(k)$ can take on is finite, represented by $\xi_1, \ldots, \xi_n$, then the PDF of $x(t)$ at time $k$ can be written as an $n$-vector

$$p(k) = \begin{pmatrix} p_1(k) \\ \vdots \\ p_n(k) \end{pmatrix} = \begin{pmatrix} \mathbb{P}(\xi x(k) = 1) \\ \vdots \\ \mathbb{P}(\xi x(k) = n) \end{pmatrix}$$

Note that the components of $p(k)$ are non-negative and sum to one.

If $x(t)$ is a homogeneous Markov process, the transition probabilities

$$g_{m\ell} = \mathbb{P}(\xi x(k) = m_j | \xi x(k-1) = l)$$

do not depend on $k$. Define the $n \times n$ transition matrix $G = [g_{m\ell}]$ and note the elements of $G$ are non-negative. Further,
the sum of elements in any column of G must be one.

With this notation, for any k,

\[ p(k) = Gp(k-1) = \ldots = G^{r}p(k-r) = \ldots \]

In particular, for \( k \geq 0 \),

\[ p(k) = G^{k}p(0). \]

**State Transition Diagrams**

A common artifice for representing a homogeneous Markov chain is a directed graph whose vertices represent the possible states of the chain (i.e., possible values of \( x(k) \)). Each edge connecting two vertices is labeled with the probability of transitioning from the state at the start of the edge to the state at the end of the edge.

**Example (n = 2):**

\[ \begin{array}{c}
\frac{1}{3} \\
\frac{2}{3}
\end{array} \] \[ \begin{array}{c}
1 \\
0
\end{array} \] \[ \begin{array}{c}
4 \\
1
\end{array} \] \[ \begin{array}{c}
\frac{1}{3} \\
\frac{2}{3}
\end{array} \] \[ \begin{array}{c}
0 \\
1
\end{array} \]

\[ G = \begin{bmatrix}
\frac{1}{3} & 0 \\
\frac{2}{3} & 1
\end{bmatrix} \]
Suppose at $k=0$, this chain has equal probability of being in state 1 and state 2:

$$p(0) = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$$

Then,

$$p(1) = Gp(0) = \begin{bmatrix} \frac{1}{3} & 0 \\ \frac{2}{3} & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{6} \\ \frac{5}{6} \end{bmatrix}$$

$$p(2) = Gp(1) = G^2p(0) = \begin{bmatrix} \frac{1}{12} \\ \frac{11}{12} \end{bmatrix}$$

**Steady State Distribution**

In the preceding example, it is clear that

$$\lim_{k \to \infty} p(k) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

The state $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ is a steady state for this chain because

$$G \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & 0 \\ \frac{2}{3} & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

More generally, a steady state is for a
homogeneous Markov chain satisfies

\[ G \pi = \pi \]

i.e., it is an eigenvector of \( G \) with eigenvalue 1. In this example, the eigenvalues of \( G \) are \( \lambda_1 = 1 \) and \( \lambda_2 = \frac{1}{3} \), so \( \pi = [i] \) is the only steady state.