Probability Review

©Prakash Ishwar

Department of Electrical and Computer Engineering
Boston University

September 6, 2015

Probability Space $(\Omega, \mathcal{F}, P)$

- **Sample space Ω:** nonempty set of **outcomes** ω of an experiment.
- **Family of events $\mathcal{F}$:** a collection of subsets of Ω that is:
  - A.1 **Nonempty:** $\mathcal{F} \neq \emptyset$, i.e., $\exists A \subseteq \Omega$ such that (s.t.) $A \in \mathcal{F}$,
  - A.2 **Closed under complementation:** $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$,
  - A.3 **Closed under countable union:** $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

  Members of $\mathcal{F}$ are called **events**.
- **Probability measure P:** real-valued function on $\mathcal{F}$, i.e., $P : \mathcal{F} \to \mathbb{R}$, satisfying the **probability axioms**:
  - P.1 **Nonnegativity:** $\forall A \in \mathcal{F}, P(A) \geq 0$,
  - P.2 **Normalization:** $P(\Omega) = 1$ ($\mathcal{F}$ contains Ω)
  - P.3 **Countable additivity:** If $A_1, A_2, \ldots$ are **mutually exclusive**, i.e., **pairwise disjoint**, events in $\mathcal{F}$, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Probability measure examples

Example 1 Let $\Omega = \{\omega_1, \omega_2, \ldots\}$ be any countable set, e.g., the set of natural (counting) numbers $\mathbb{N}$ or the set of all rational numbers $\mathbb{Q}$. Let $p(\omega_1), p(\omega_2), \ldots$, be any sequence of nonnegative numbers that sum to one, e.g., $p(\omega_i) = 2^{-i}$. For any event A, if we define $P(A) := \sum_{\omega \in A} p(\omega)$, then P is a valid probability measure.

Example 2 Let $\Omega = \mathbb{R}^n$, the n-dimensional real Euclidean space. Let $f(x)$ be any nonnegative integrable function of $x \in \mathbb{R}^n$ that integrates to one, e.g., $f(x) = 1$ for all $x \in B$ and zero otherwise, where B is any unit-volume subset of $\mathbb{R}^n$. For any event A, if we define $P(A) := \int_{\omega \in A} f(x)dx$, then P is a valid probability measure.

Indicator function: A function that takes the value 1 over a set B and the value 0 outside B is called the indicator function of the set B and is denoted by $1_B(x)$ or $1(x \in B)$ or $I_{\{x \in B\}}$.

Properties of probability measure P

- $P(A^c) = 1 - P(A)$.
- $P(\emptyset) = 0$.
- **Monotonicity:** if $A \subseteq B$ then $P(A) \leq P(B)$.
- **Unions:**
  - $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$.
  - $P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 A_2) - P(A_1 A_3) - P(A_2 A_3) + P(A_1 A_2 A_3)$.
  - General case (via inclusion-exclusion principle): $P(\cup_i A_i) = \sum_i P(A_i) - \sum_{i \neq j} P(A_i A_j) + \sum_{i \neq j \neq k} P(A_i A_j A_k) - \ldots - (-1)^n P(\cap_i A_i)$.
  - **Union bound:**
  $$P(\cup_i A_i) \leq \sum_i P(A_i).$$
- **Continuity:**
  - If $A_1 \subseteq A_2 \subseteq \ldots$ then $\lim_{j \to \infty} P(A_j) = P(\lim_{j \to \infty} \cup_{i=1}^{j} A_i)$.
  - If $A_1 \supseteq A_2 \supseteq \ldots$ then $\lim_{j \to \infty} P(A_j) = P(\lim_{j \to \infty} \cap_{i=1}^{j} A_i)$.

Conditional probability

- If A and B are events and $P(B) > 0$ then the **conditional probability** of A given B is $P(A|B) := P(A \cap B)/P(B)$.

- Let $(\Omega, \mathcal{F}, P)$ be a probability space and B an event with $P(B) > 0$. For each event $A \in \mathcal{F}$, define $P_B(A) := P(A|B)$. Then $(\Omega, \mathcal{F}, P_B)$ is a probability space.

- **Bayes' rule:** $P(B|A) = P(A|B)P(B)/P(A)$ if $P(A), P(B) > 0$.

- **Law of total probability:** Let $B_1, \ldots, B_n$ form a **partition** of Ω meaning that they are mutually exclusive and $\Omega = \cup_{i=1}^n B_i$. If for each i, $P(B_i) > 0$, then for any event A

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i).$$

Independence and conditional independence of events

- Events $A_1, \ldots, A_n$ are independent if for all $\mathcal{I} \subseteq \{1, \ldots, n\}$,

$$P(\cap_{i \in \mathcal{I}} A_i) = \prod_{i \in \mathcal{I}} P(A_i),$$

  and we write $\perp\!\!\!\perp (A_1, \ldots, A_n)$. This requires a total of $2^n - n - 1$ equations to hold (for all the $n+1$ subsets $\mathcal{I}$ of size 0 or 1, the equations are trivially satisfied).
  Note: If $A \perp\!\!\!\perp B$ and $P(B) > 0$ then $P(A|B) = P(A)$.
- Events $A_1, \ldots, A_n$ are **conditionally independent** given an event B with $P(B) > 0$, if for all $\mathcal{I} \subseteq \{1, \ldots, n\}$,

$$P_B(\cap_{i \in \mathcal{I}} A_i) = \prod_{i \in \mathcal{I}} P_B(A_i), \text{ i.e., } P(\cap_{i \in \mathcal{I}} A_i | B) = \prod_{i \in \mathcal{I}} P(A_i | B)$$

  and we write $\perp\!\!\!\perp (A_1, \ldots, A_n) \mid B$.

Pairwise independence of events

- Events $A_1, \dots, A_n$ are **pairwise independent** if for all $i, j \in \{1, \dots, n\}$, $i \neq j$,

$$P(A_i \cap A_j) = P(A_i) \cdot P(A_j),$$

  and we write $A_i \perp\!\!\!\perp A_j$. This requires a total of $n(n-1)/2$ equations to hold.

- Events $A_1, \dots, A_n$ are **conditionally pairwise independent** given an event B with $P(B) > 0$, if for all $i, j \in \{1, \dots, n\}$, $i \neq j$,

$$P(A_i \cap A_j | B) = P(A_i | B) \cdot P(A_j | B).$$

- Independence (respectively conditional independence) implies pairwise (resp. conditional pairwise) independence but the reverse assertions do not, in general, hold.
- Notation: $ABC \equiv A \cap B \cap C$

A key property of independent events

- Suppose $\perp\!\!\!\perp (A_1, A_2, \ldots, A_n)$. Let $n = n_1 + n_2 + \ldots + n_k$ where $n_1, \ldots, n_k \in \mathbb{N}$. Suppose B_1 is defined by Boolean operations (intersections, complements, and unions) of the first n_1 events $A_1, \ldots, A_{n_1}$, B_2 is defined by Boolean operations on the next n_2 events $A_{n_1+1}, \ldots, A_{n_1+n_2}$, and so on. Then $B_1, B_2, \ldots, B_k$ are independent.

One random variable (RV)

- Informally, a random variable is a function mapping outcomes to real numbers.

- Formally, let $(\Omega, \mathcal{F}, P)$ be a probability space. A random variable X is function from Ω to $\mathbb{R}$ such that for all $x \in \mathbb{R}$, the set of outcomes $X^{-1}((-\infty, x]) := \{\omega : X(\omega) \leq x\}$ is a valid event, i.e., $X^{-1}((-\infty, x]) \in \mathcal{F}$.

- **Cumulative Distribution Function** (CDF) of random variable X:

$$\forall x \in \mathbb{R}, \quad F_X(x) := P(\{\omega : X(\omega) \leq x\}) = P(X \leq x).$$

- **Properties of CDF:**
  - F.1 F is nondecreasing.
  - F.2 $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$.
  - F.3 F is right continuous: $\lim_{x\downarrow a} F(x) = F(a)$.

- For $a < b$, $F_X(b) - F_X(a) = P(a < X \leq b)$.

Discrete random variable

- A random variable X is discrete (or simple) if there is a countable subset $\{x_1, x_2, \ldots\}$ of real numbers s.t. $P(X \in \{x_i : i \in \mathbb{N}\}) = 1$.

- The **probability mass function** (pmf) of a discrete RV X, denoted $p_X(x)$, is defined for $x \in \mathbb{R}$ as $p_X(x) := P(X = x)$.

- If X has only finitely many mass points in any finite interval, then F_X is a piecewise constant function.

- If X is discrete then $\forall x \in \mathbb{R}$, $F_X(x) = \sum_{y: y \leq x} p_X(y)$.

- **Example 3** A Geometric random variable with parameter $q \in (0, 1]$, denoted Geom(q), has pmf $p_X(i) = q(1-q)^{i-1}, i \in \mathbb{N}$. This models the number of independent coin flips until the first heads appears with $P(\text{Heads}) = q$.

Continuous random variable

- A random variable X is called continuous if its CDF can be expressed as the integral of a nonnegative function:

$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy.$$

- f_X is called the **probability density function** (pdf) of X and if f_X is continuous at x,

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

- For any subset A of $\mathbb{R}$,

$$P(X \in A) = \int_A f_X(x)dx.$$

- **Example 4** A **standard** Gaussian (or Normal) random variable, denoted $\mathcal{N}(0,1)$, has the pdf $f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\}$.

Additional remarks on discrete and continuous RVs

- The **probability distribution** of X (or induced by X), denoted P_X, is defined as $P_X(A) := P(X \in A)$ where A is a subset of $\mathbb{R}$.

- A discrete random variable X may be viewed as having a *generalized* pdf made up of Diracs (impulse functions):

$$f_X(x) = \sum_{y : p_X(y) > 0} p_X(y)\delta(x - y)$$

- A random variable which is neither discrete nor continuous is called mixed. Example, a random variable X with pdf

$$f_X(x) = 0.2\delta(x - 3) + 0.8\mathcal{N}(0, 1)(x).$$

Function of a random variable

- Let X be an RV on a probability space $(\Omega, \mathcal{F}, P)$. Then X is a function from Ω to $\mathbb{R}$.

- Suppose g is a function from $\mathbb{R}$ to $\mathbb{R}$ such that

$$g^{-1}((-\infty, y]) = \{x : g(x) \leq y\}$$

  is an event for all y.

- Then $Y(\omega) = g(X(\omega))$ is a function from Ω to $\mathbb{R}$ and therefore an RV.

- The CDF of $Y = g(X)$ is given by

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \in g^{-1}((-\infty, y])).$$

Function of a random variable (cont.)

Example 5 Let U be a continuous RV with pdf $f_U(u) = 1_{[0,1]}(u)$. This RV is uniformly distributed over $[0,1]$ and is denoted Unif$(0,1)$. It's CDF is given by

$$F_U(u) = \int_{-\infty}^{u} 1_{[0,1]}(x)dx = \begin{cases} 0 & u < 0 \\ u & u \in [0,1) \\ 1 & u \geq 1 \end{cases}$$

Let $X = g(U)$ with $g(u) = u^2$. Then

$$\begin{aligned} F_X(x) &= P(X \leq x) = P(g(U) \leq x) = P(U^2 \leq x) \\ &= P(U \leq \sqrt{x}) \text{ since } U \geq 0 \\ &= F_U(\sqrt{x}) \\ &= \begin{cases} 0 & x < 0 \\ \sqrt{x} & x \in [0,1) \\ 1 & x \geq 1 \end{cases} \end{aligned}$$

Differentiate to get $f_X(x) = \frac{1}{2\sqrt{x}} 1_{(0,1]}(x)$.

Function of a random variable (cont.)

- If
  (i) X is continuous,
  (ii) g has a continuous derivative g' (which is therefore bounded),
  (iii) the inverse image set $g^{-1}(y) := \{x : g(x) = y\}$ is **countable** for all y in the range of g and
  (iv) $\inf_{x \neq \tilde{x} \in g^{-1}(y)} |x - \tilde{x}| > 0$,

  then $Y = g(X)$ is continuous with pdf

$$f_Y(y) = \sum_{x:g(x)=y} \frac{f_X(x)}{|g'(x)|}.$$

- **Intuition:** Conservation of probability: If $g^{-1}(y) = \{x_1, x_2, \ldots\}$ then for all sufficiently small Δy

$$\begin{aligned} f_Y(y)\Delta y &\approx P(Y \in (y - \Delta y, y + \Delta y)) \\ &= P(X \in \cup_i (x_i - \Delta x_i, x_i + \Delta x_i)) \\ &\approx \sum_i f_X(x_i)\Delta x_i \end{aligned}$$

  where the ratios $\Delta y / \Delta x_i \to |g'(x_i)|$ as $\Delta y, \Delta x_i \to 0$.

Application: generating RVs with a specified CDF

- **Given:** Target CDF F and $U \sim \text{Unif}(0,1)$.
- **Find:** function g so that $X := g(U)$ has specified CDF F.
- **Solution:** Define $g(u) := \min\{x : F(x) \geq u\}$.
- **Intuition:** If F is strictly increasing and continuous, it is invertible and $g(u) = F^{-1}(u)$. Then

$$\begin{aligned} F_X(x) &= P(X \leq x) = P(F^{-1}(U) \leq x) \\ &= P(F(F^{-1}(U)) \leq F(x)) = P(U \leq F(x)) \\ &= F(x). \end{aligned}$$

- F is not invertible at u if its graph is either flat at u or u is within a jump. In either case the solution works (verify!).
- This technique is used in computer simulations of random systems.

Expectation of a random variable

- Expectation, expected-value, mean, or mean value of a random variable X, denoted $E[X]$ or μ_X, on a probability space $(\Omega, \mathcal{F}, P)$ with CDF F_X and probability distribution P_X is defined as:

$$
\begin{aligned}
E[X] &= \int_\Omega X(\omega) P(d\omega) \\
&= \int_{-\infty}^{\infty} x P_X(dx) \\
&= \int_{-\infty}^{\infty} x dF_X(x) \\
&= \sum_{x>0} x p_X(x) + \sum_{x<0} x p_X(x) \\
&\quad \text{(for } X \text{ discrete, at least one sum finite)} \\
&= \int_{x>0} x f_X(x) dx + \int_{x<0} x f_X(x) dx \\
&\quad \text{(for } X \text{ continuous, at least one integral finite)}
\end{aligned}
$$

Properties of expectation

1. **Linearity:** If $E[X], E[Y]$, and $E[X]+E[Y]$ are well defined, then $E[X+Y]$ is well defined and $E[X+Y]=E[X]+E[Y]$. Also, for all $a \in \mathbb{R}$, $E[aX]=aE[X]$.
2. **Order preservation:** If $P(X \geq Y)=1$ and $E[Y]$ is well defined then $E[X]$ is well defined and $E[X] \geq E[Y]$.
3. **Expectation via CDF:**
$$E[X]=\int_0^\infty (1-F_X(x))dx - \int_{-\infty}^0 F_X(x)dx$$
whenever at least one of the two integrals is finite.
4. If $Y=g(X)$,
$$\begin{aligned} E[Y]=E[g(X)] &= \int_\Omega g(X(\omega))P(d\omega) = \int_{-\infty}^\infty g(x)dF_X(x) \\ &= \int_{-\infty}^\infty g(x)f_X(x)dx \quad (X \text{ continuous}) \\ &= \sum_x g(x)p_X(x) \quad (X \text{ discrete}). \end{aligned}$$

Quantities defined via expectation

- $E[1_A(X)] = P(X \in A)$ for any subset A.
- **Variance:** If $\mu_X := E[X]$ is finite, the variance of X, denoted $\mathrm{var}(X)$, is defined by

$$\begin{aligned} \mathrm{var}(X) &:= E[(X - E[X])^2] \\ &= E[X^2 - 2X\mu_X + \mu_X^2] \\ &= E[X^2] - (E[X])^2 \quad \text{(linearity of expectation).} \end{aligned}$$

- **Standard Deviation:** The standard deviation of X, denoted σ_X, is defined by $\sigma_X = +\sqrt{\mathrm{var}(X)}$.
- **Markov inequality:** If Y is a nonnegative RV then for any $c > 0$,

$$P(Y \geq c) \leq \frac{E[Y]}{c}.$$

  Proof: $c1_{[c,\infty)}(Y) \leq Y$ and take expectations on both sides.
- **Chebychev inequality:** If X has finite mean μ_X and variance σ_X^2 then for any $d > 0$,

$$P(|X - \mu_X| \geq d) \leq \frac{\sigma_X^2}{d^2}.$$

Characteristic function

- The characteristic function of an RV X, denoted $\Phi_X(v)$ is defined by
$$\Phi_X(v) = E[e^{jvX}], \quad v \in \mathbb{R},\ j = \sqrt{-1}.$$
If X has pdf f_X then
$$\Phi_X(v) = \int_{-\infty}^{\infty} \exp(jvx) f_X(x) dx,$$
which is 2π times the inverse Fourier transform (in radians) of f_X.
- Two RVs have the same probability distribution if, and only if (iff) they have the same characteristic function.
- If $E[X^k]$, the k**-th moment of** X, exists and is finite for an integer $k \geq 1$, then the derivatives of Φ_X up to order k exist and are continuous, and
$$\Phi_X^{(k)}(0) = \frac{d^k}{dv^k}\Phi(v)\bigg|_{v=0} = j^k E[X^k].$$

Frequently used distributions

- **Discrete RVs:** Bernoulli, Binomial, Geometric, Poisson, etc.
- **Continous RVs:** Uniform, Exponential, Rayleigh, Gamma, Gaussian, etc.
- **Gaussian (Normal):** with mean μ and variance σ^2 is denoted by $\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma \geq 0$. Characteristic function:

$$\Phi_X(v) = \exp\left(jv\mu - \frac{1}{2}v^2\sigma^2\right).$$

$$f_X(x) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) & \sigma > 0, X \text{ continuous} \\ \delta(x-\mu) & \sigma = 0, X \text{ discrete.} \end{cases}$$

The so called Q-function is defined as the **tail probability**:

$$Q(x) := \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = 1 - F_X(x), \quad X \sim \mathcal{N}(0,1).$$

Jointly distributed random variables

- Let $X_1, X_2, \ldots, X_m$ be RVs on the same probability space $(\Omega, \mathcal{F}, P)$. The **joint CDF** is a function on $\mathbb{R}^m$ defined by

  $$F_{X_1 X_2 \ldots X_m}(x_1, x_2, \ldots, x_m) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_m \leq x_m)$$

- Notation: comma $\equiv \cap$, i.e., logical AND.
  $P(X_1 \leq x_1, X_2 \leq x_2) = P(\{X_1 \leq x_1\} \cap \{X_2 \leq x_2\})$.

- $$F_{X_1 X_2}(x_1, +\infty) = \lim_{x_2 \to \infty} F_{X_1 X_2}(x_1, x_2) = F_{X_1}(x_1),$$

  where $F_{X_1}(x_1)$ is the **marginal CDF** of X_1.

- $$F_{X_1 X_2}(x_1, -\infty) := \lim_{x_2 \to -\infty} F_{X_1 X_2}(x_1, x_2) = 0.$$

Jointly distributed random variables (cont.)

- The RVs are **jointly continuous** if there exists a function $f_{X_1 X_2 \dots X_m}$ called the **joint pdf** such that

$$F_{X_1 \dots X_m}(x_1, \dots, x_m) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_m} f_{X_1 \dots X_m}(u_1, \dots, u_m) du_1 \dots du_m.$$

- If X_1, X_2 are jointly continuous, then

$$\begin{aligned} F_{X_1}(x_1) &= F_{X_1 X_2}(x_1, +\infty) = \int_{-\infty}^{x_1} \left[\int_{-\infty}^{\infty} f_{X_1 X_2}(u_1, u_2) du_2 \right] du_1 \\ &= \int_{-\infty}^{x_1} f_{X_1}(u_1) du_1. \end{aligned}$$

- $f_{X_1}, f_{X_2}, \dots, f_{X_m}$ are called the **marginal pdfs** and can be obtained by integrating out other coordinates of the joint pdf.
- If $f_{X_1 \dots X_m}$ is continuous at $(x_1, \dots, x_m)$ then

$$\frac{\partial^m}{\partial x_1 \dots \partial x_m} F_{X_1 \dots X_m}(x_1, \dots, x_m) = f_{X_1 \dots X_m}(x_1, \dots, x_m).$$

Jointly distributed random variables (cont.)

- If $X_1, \ldots, X_m$ are each discrete RVs, then they have a **joint pmf** $p_{X_1 X_2 \ldots X_m}$ defined by

  $$p_{X_1 \ldots X_m}(x_1, \ldots, x_m) = P(\{X_1 = x_1\} \cap \{X_2 = x_2\} \cap \ldots \cap \{X_m = x_m\})$$

  or in short $P(X_1 = x_1, \ldots, X_m = x_m)$,
- For any subset A of $\mathbb{R}^m$,

  $$P((X_1, \ldots, X_m) \in A) = \sum_{(u_1, \ldots, u_m) \in A} p_{X_1 \ldots X_m}(u_1, \ldots, u_m)$$

- The **marginal pmf**s can be obtained by summing out other coordinates of the joint pmf, e.g.,

  $$p_{X_1}(x_1) = \sum_{u_2} p_{X_1 X_2}(x_1, u_2)$$

- **Joint characteristic function:**

  $$\Phi_{X_1 \ldots X_m}(v_1, \ldots, v_m) := E[e^{j(v_1 X_1 + \ldots + v_m X_m)}].$$

Independence via CDF, expectation, ch.fn., pmf, and pdf

- Random variables $X_1, \ldots, X_m$ are independent, denoted by $\perp\!\!\!\perp (X_1, \ldots, X_m)$, if for **all** subsets $B_1, \ldots, B_m$ of $\mathbb{R}$, the events $A_1 := \{X \in B_1\}, \ldots, A_m := \{X \in B_m\}$ are independent.
- $\Leftrightarrow$ the joint CDF is separable, i.e., it factorizes into the product of all the marginal CDFs:
$$F_{X_1 \ldots X_m}(x_1, \ldots, x_m) = F_{X_1}(x_1) \cdots F_{X_m}(x_m).$$
- $\Leftrightarrow$ for all functions $g_1, \ldots, g_m$, from $\mathbb{R}$ to $\mathbb{R}$,
$$E\left[g_1(X_1) \cdots g_m(X_m)\right] = E\left[g_1(X_1)\right] \cdots E\left[g_1(X_m)\right].$$
- $\Leftrightarrow$ the joint characteristic function is separable:
$$\Phi_{X_1 \ldots X_m}(v_1, \ldots, v_m) = \Phi_{X_1}(v_1) \cdots \Phi_{X_m}(v_m).$$
- $\Leftrightarrow$ (if $X_1, \ldots, X_m$ are each discrete):
$$p_{X_1 \ldots X_m}(x_1, \ldots, x_m) = p_{X_1}(x_1) \cdots p_{X_m}(x_m).$$
- $\Leftrightarrow$ (if $X_1, \ldots, X_m$ are jointly continuous):
$$f_{X_1 \ldots X_m}(x_1, \ldots, x_m) = f_{X_1}(x_1) \cdots f_{X_m}(x_m).$$

Conditional densities

- Let X and Y be jointly continuous random variables with joint pdf $f_{XY}(x, y)$. For all y s.t. $f_Y(y) > 0$ the conditional density of X given Y is defined by

$$f_{X|Y}(x|y) := \frac{f_{XY}(x, y)}{f_Y(y)}.$$

  Note: If X, Y and jointly continuous and independent, then $f_{X|Y}(x|y) = f_X(x)$ for all $y : f_Y(y) > 0$.

- If y is fixed and $f_Y(y) > 0$, then as a function of x, $f_{X|Y}(x|y)$ is itself a pdf.

- If $f_Y(y) > 0$ and A is a subset,

$$P(X \in A|Y = y) := \int_A f_{X|Y}(x|y)dx.$$

Conditional expectation

- The expectation of the conditional pdf is called the conditional expectation (or conditional mean) of X given $Y = y$:

$$E[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx.$$

- $g(y) := E[X|Y = y]$ is a number; a deterministic function of y.
- Substituting $y = Y$ in g makes $g(Y) = E[X|Y]$ a random variable. It is in fact a function of the random variable Y. (Note: a function of a random variable is a random variable).
- Discrete RVs: $E[\psi(X)|Y = y] = \sum_x \psi(x) p_{X|Y}(x|y)$.
- Continuous RVs: $E[\psi(X)|Y = y] = \int_{-\infty}^{\infty} \psi(x) f_{X|Y}(x|y) dx$.

Law of iterated expectations

- **Law of iterated expectations, total expectation, tower rule, or smoothing rule:** If $g(Y) = E[X|Y]$, then

$$E[g(Y)] = E[E[X|Y]] = E[X].$$

  The inner expectation $E[X|Y]$ in a conditional expectation with respect to (w.r.t.) $f_{X|Y}$. The outer expectation in $E[E[X|Y]]$ is w.r.t. f_Y.

- Similarly, if X, Y, Z have a joint pdf then

$$E[X|Z] = E[E[X|Y, Z]|Z]$$

  where the inner expectation is w.r.t. $f_{X|YZ}$ and the outer w.r.t. $f_{Y|Z}$.

Conditional independence and Markov chain

- RVs X and Z are conditionally independent given RV Y if for all y with $p_Y(y) > 0$ (discrete RV) or $f_Y(y) > 0$ (continuous RV):

$$p_{XZ|Y}(x, z|y) = p_{X|Y}(x|y)p_{Z|Y}(z|y) \quad \text{(discrete RVs)}$$
$$f_{XZ|Y}(x, z|y) = f_{X|Y}(x|y)f_{Z|Y}(z|y) \quad \text{(continuous RVs)}$$

  and we say $X - Y - Z$ is a Markov chain.
  Note: $X - Y - Z \Leftrightarrow Z - Y - X$.

- Equivalently, $X - Y - Z$ if

$$p_{XYZ}(x, y, z)p_Y(y) = p_{XY}(x, y)p_{ZY}(z, y) \quad \text{(discrete RVs)}$$
$$f_{XYZ}(x, y, z) = f_{XY}(x, y)f_{ZY}(z, y) \quad \text{(continuous RVs)}$$

- Equivalently, $X - Y - Z$ if for all (x, y) with $p_{X,Y}(x, y) > 0$ (discrete) or $f_{X,Y}(x, y) > 0$ (continuous):

$$p_{Z|Y,X}(z|y, x) = p_{Z|Y}(z|y) \quad \text{(discrete RVs)}$$
$$f_{Z|Y,X}(z|y, x) = f_{Z|Y}(z|y) \quad \text{(continuous RVs)}$$

Cross moments of 2 random variables

- **Correlation:** $R_{XY} := E[XY]$.
- **Covariance:**
  $\text{Cov}(X,Y) := E[(X - E[X])(Y - E[Y])] = R_{XY} - \mu_X \mu_Y$. Also denoted by σ_{XY}, C_{XY}, K_{XY}, and Σ_{XY} in the literature.
- **Correlation coefficient:** $\rho_{XY} := \frac{\text{Cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$, $\text{var}(X), \text{var}(Y) > 0$.
- **Cauchy-Schwarz inequality:**

$$|E[XY]| \leq \sqrt{E[X^2]E[Y^2]}.$$

  If $E[Y^2] > 0$ then equality holds, iff $P(X = cY) = 1$ for some constant c.
- $|\rho_{XY}| \leq 1$ (follows from Cauchy-Schwarz).
- L^2 **Triangle inequality:** (follows from Cauchy-Schwarz)

$$\sqrt{E[(X+Y)^2]} \leq \sqrt{E[X^2]} + \sqrt{E[Y^2]}.$$

Cross moments of 2 random variables (cont.)

- **Orthogonal:** X and Y are called orthogonal if their **correlation** $R_{XY} = E[XY] = 0$ and we write $X \perp Y$.

- **Uncorrelated:** X and Y are called uncorrelated if their **covariance** $\text{Cov}(X, Y) = 0$.

- $X \perp\!\!\!\perp Y \Rightarrow \text{Cov}(X, Y) = 0$ but in general $X \perp\!\!\!\perp Y \not\Leftarrow \text{Cov}(X, Y) = 0$.

- Properties of Cov :
  1. $\text{var}(X) = \text{Cov}(X, X)$.
  2. $\text{Cov}(X, Y) = E[X(Y - E[Y])] = E[(X - E[X])Y]$
  3. $\text{Cov}(X + Y, U + V) =$ $\text{Cov}(X, U) + \text{Cov}(X, V) + \text{Cov}(Y, U) + \text{Cov}(Y, V)$.
  4. If $X_1, \ldots, X_m$ are (pairwise) uncorrelated each with mean μ and variance σ^2 and $S_m := \sum_{i=1}^{m} X_i$ then $E[S_m] = m\mu$, $\text{Cov}(S_m) = m\sigma^2$, and $\frac{1}{\sqrt{m\sigma^2}}(S_m - m\mu)$ has zero mean and unit variance.

Random vectors

- A random vector X of dimension m is an **ordered tuple** of m random variables on the same probability space arranged as an $m \times 1$ column vector $(X_1, \ldots, X_m)^T$, where ^{T} denotes transpose.
- The CDF of X, is the joint CDF of the m component RVs: $F_X(x) = P(X_1 \leq x_1, \ldots, X_m \leq x_m)$, $x = (x_1, \ldots, x_m)^T$.
- The expectation or mean of X is the $m \times 1$ vector $\mu_X := E[X] = (E[X_1], \ldots, E[X_m])^T$.
- Let $X = (X_1, \ldots, X_m)^T$ and $Y = (Y_1, \ldots, Y_n)^T$ be two random vectors, of dimensions m and n respectively, on the same probability space. Their joint CDF is $F_{XY}(x, y) = P(X_1 \leq x_1, \ldots, X_m \leq x_m, Y_1 \leq y_1, \ldots, Y_n \leq y_n)$, $x = (x_1, \ldots, x_m)^T$, $y = (y_1, \ldots, y_n)^T$. The marginal CDFs are $F_X(x)$ and $F_Y(y)$.
- If $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ are jointly continuous, the joint pdf of X, Y is denoted by $f_{XY}(x, y)$, the marginals by $f_X(x)$ and $f_Y(y)$, and for $f_Y(y) > 0$, the conditional pdf of X given Y by $f_{X|Y}(x|y) = f_{XY}(x, y)/f_Y(y)$.

Transformation of random vectors

- Let $X \in \mathbb{R}^n$ be continuous with pdf $f_X(x)$. Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be a one-to-one mapping. Let $y = g(x) = (g_1(x), \ldots, g_n(x))^T$ where for $i = 1, \ldots, n$, $g_i : \mathbb{R}^n \to \mathbb{R}$.
- If the $n \times n$ Jacobian matrix of partial derivatives of g:

$$\frac{\partial y}{\partial x}(x) = \frac{\partial g}{\partial x}(x) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(x) & \cdots & \frac{\partial g_1}{\partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_1}(x) & \cdots & \frac{\partial g_n}{\partial x_n}(x) \end{pmatrix}$$

  exists, is continuous at x, and nonsingular for all x, then the RV $Y := g(X)$ is continuous and for all y in the range of g,

$$f_Y(y) = \frac{f_X(x)}{\left|\frac{\partial y}{\partial x}(x)\right|} = f_X(x) \left|\frac{\partial x}{\partial y}(y)\right|,$$

  where $x = g^{-1}(y)$, $|\cdot| := |\det(\cdot)|$, and $\frac{\partial x}{\partial y}(y) = \left(\frac{\partial y}{\partial x}(x)\right)^{-1}$ is the inverse of the Jacobian of matrix of g.

Auto and cross correlation/covariance matrices

$X \in \mathbb{R}^m, Y \in \mathbb{R}^n$ two random vectors on same probability space.

- **Cross correlation matrix:** $R_{XY} := E[XY^T]$ is an $m \times n$ matrix of correlations whose ij-th entry is $E[X_i Y_j] = R_{X_i Y_j}$.
- **Cross covariance matrix:** Denoted by $\mathrm{Cov}(X, Y)$, C_{XY}, K_{XY}, and Σ_{XY}. $\mathrm{Cov}(X, Y) := E[(X - E[X])(Y - E[Y])^T]$ is an $m \times n$ matrix of covariances whose ij-th entry is $\mathrm{Cov}(X_i, Y_j)$.
- $\mathrm{Cov}(X, Y) = R_{XY} - E[X](E[Y])^T$. Thus if $E[X]$ or $E[Y]$ is zero, $\mathrm{Cov}(X, Y) = R_{XY}$.
- **(Auto) correlation matrix:** $R_{XX} = E[XX^T]$ the correlation matrix of X with itself. Often the prefix 'auto' and suffix 'matrix' are omitted and R_{XX} is shortened to R_X.
- **(Auto) covariance matrix:** $\mathrm{Cov}(X, X)$ often shortened to $\mathrm{Cov}(X)$ with prefix 'auto' and suffix 'matrix' omitted.
- **Note:** (auto) correlation and covariance matrices are square but cross correlation and cross covariance matrices need not be.

Conditional mean, correlation, covariance

Let $X \in \mathbb{R}^m, Y \in \mathbb{R}^n, Z \in \mathbb{R}^k$ be three jointly continuous random vectors on the same probability space.

- **Conditional mean:** $\mu_{X|z} = E[X|Z = z] := \int x f_{X|Z}(x|z) dx$. $\mu_{Y|z}$ is similarly defined.
- **Conditional cross correlation:** $R_{XY|z} := E[XY^T|Z = z] = \int xy^T f_{XY|Z}(x, y|z) dx dy$.
- **Conditional cross covariance:** $\text{Cov}(X, Y|z) := \int (x - \mu_{X|z})(y - \mu_{Y|z})^T f_{XY|Z}(x, y|z) dx dy$. Also denoted by $C_{XY|z}$, $K_{XY|z}$, and $\Sigma_{XY|z}$.
- **Conditional (auto) correlation:** $R_{X|z} := E[XX^T|Z = z] = \int xx^T f_{X|Z}(x|z) dx$. Similarly for $R_{Y|z}$.
- **Conditional (auto) covariance:** $\text{Cov}(X, X|z) := \int (x - \mu_{X|z})(x - \mu_{X|z})^T f_{X|Z}(x|z) dx$. Also denoted by $\text{Cov}(X|z)$, $C_{X|z}$, $K_{X|z}$, and $\Sigma_{X|z}$.
- **Note:** Can define above quantities even when X, Y, Z are not jointly continuous.

Orthogonal, uncorrelated, independent random vectors

- **Orthogonal:** X and Y are called orthogonal if their **cross correlation matrix** $R_{XY} = E[XY^T] = 0$ and we write $X \perp Y$.
  Note: there are no conditions on R_X and R_Y.

- **Uncorrelated:** X and Y are called uncorrelated if their **cross covariance matrix** $\text{Cov}(X, Y) = 0$.
  Note-1: there are no conditions on $\text{Cov}(X)$ and $\text{Cov}(Y)$.
  Note-2: The **components** of a random vector X are uncorrelated or decorrelated if $\text{Cov}(X)$ is a **diagonal matrix.**

- **Independent:** X and Y are independent if $F_{XY}(x, y) = F_X(x)F_Y(y)$ (or corresponding conditions for pdfs/pmfs).
  Note: the components of X (respectively Y) need not be independent.
- $X \perp\!\!\!\perp Y \Rightarrow \text{Cov}(X, Y) = 0$ but in general $X \perp\!\!\!\perp Y \not\Leftarrow \text{Cov}(X, Y) = 0$.

Properties of auto/cross correlation/covariance matrices

For A, C nonrandom matrices and b, d nonrandom vectors,

1. $E[AX + b] = AE[X] + b$
2. $$\begin{aligned} \mathsf{Cov}(X, Y) &= E[X(Y - E[Y])^T] \\ &= E[(X - E[X])Y^T] \\ &= E[XY^T] - E[X](E[Y])^T \end{aligned}$$
3. $E[(AX)(CY)^T] = AE[XY^T]C^T$
4. $\mathsf{Cov}(AX + b, CY + d) = A\mathsf{Cov}(X, Y)C^T$
5. $\mathsf{Cov}(AX + b) = A\mathsf{Cov}(X)A^T$
6. $$\begin{aligned} \mathsf{Cov}(W + X, Y + Z) &= \mathsf{Cov}(W, Y) + \mathsf{Cov}(W, Z) \\ &\quad + \mathsf{Cov}(X, Y) + \mathsf{Cov}(X, Z). \end{aligned}$$

Properties of auto/cross correlation/covariance matrices

- $\mathsf{Cov}(X, Y) = (\mathsf{Cov}(Y, X))^T$: ij-th element of $\mathsf{Cov}(X, Y)$ $= \mathsf{Cov}(X_i, Y_j) = \mathsf{Cov}(Y_j, X_i) = ji$-th element of $\mathsf{Cov}(Y, X)$.

- Auto correlation/covariance matrices are symmetric: for all i, j, $\mathsf{Cov}(X_i, X_j) = \mathsf{Cov}(X_j, X_i)$. Therefore $\mathsf{Cov}(X) = (\mathsf{Cov}(X))^T$.

- The diagonal elements of auto correlation/covariance matrices are nonnegative since for all i, $\mathsf{Cov}(X_i, X_i) = \mathsf{var}(X_i) \geq 0$.

- For all i, j, $|\mathsf{Cov}(X_i, X_j)| \leq \sqrt{\mathsf{Cov}(X_i, X_i)\mathsf{Cov}(X_j, X_j)}$ (Cauchy-Schwarz inequality).

Linear Algebra Facts

Let A be an $n \times n$ real square matrix.

- $u \neq 0$ is an **eigenvector** of A with **eigenvalue** λ if $Au = \lambda u$.
- The eigenvalues of an $n \times n$ matrix A are the roots of its degree-n **characteristic polynomial**: $p(\lambda) := \det(\lambda I - A) = 0$.
- All the eigenvalues of a real symmetric matrix are real-valued.
- For any real symmetric matrix there is an orthonormal basis made up of its eigenvectors (which are real-valued).
- **Real Spectral Theorem (eigendecomposition):** Every $n \times n$ real symmetric matrix K can be decomposed as

$$K = U \Lambda U^T = \begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} \begin{pmatrix} u_1^T \\ \vdots \\ u_n^T \end{pmatrix} = \sum_{i=1}^{n} \lambda_i u_i u_i^T$$

  where U is an $n \times n$ real **orthonormal** matrix, i.e., $UU^T = U^T U = I_n$, whose columns are orthonormal eigenvectors of K, i.e., $\forall i$, $Ku_i = \lambda_i u_i$, and Λ is an $n \times n$ real diagonal matrix of eigenvalues.

Linear Algebra Facts (cont.)

- The **matrix square root** of an $n \times n$ real symmetric matrix K with eigendecomposition $U \Lambda U^T$ is given by $\sqrt{K} = U\sqrt{\Lambda}U^T$ where $\sqrt{\Lambda} := \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$.

- An $n \times n$ real matrix K is called positive semidefinite (or nonnegative definite) if $\forall a \in \mathbb{R}^n$, $a^T K a \geq 0$. It is called **positive definite** if for all **nonzero** $a \in \mathbb{R}^n$, $a^T K a > 0$.

- A real symmetric matrix is positive semidefinite (resp. positive definite) iff all its eigenvalues are nonnegative (resp. strictly positive).

- **Sylvester's test:** A real symmetric matrix is positive semidefinite (resp. positive definite) iff all its leading principal minors are nonnegative (resp. strictly positive). The j-th leading principal minor of a matrix K is the determinant of its upper-left $j \times j$ sub-matrix.

Characterization of auto correlation/covariance matrices

- **Result:** Any $m \times n$ matrix A is a valid cross covariance matrix: *Proof:* Let $Y = (Y_1, \ldots, Y_n)^T$ where $Y_1, \ldots, Y_n$ are **Independent and Identically Distributed** (IID) with zero mean and unit variance. Then $\mathrm{Cov}(Y) = I_n$ the $n \times n$ identity matrix. If $X := AY$ then $\mathrm{Cov}(X, Y) = A\mathrm{Cov}(Y) = AI_n = A$.
- **Result:** Auto correlation/covariance matrices are **real**, **symmetric**, and **positive semidefinite**. Conversely, if K is a real, symmetric, positive semidefinite matrix, then K is the correlation/covariance matrix of some zero-mean random vector X.

  Proof: Auto correlation/covariance matrices are clearly real and symmetric by definition. They are also positive semidefinite because $\forall a \in \mathbb{R}^n$, $a^T E[XX^T]a = E[(a^T X)^2] \geq 0$. Conversely, if K is real, symmetric, and positive semidefinite, it has an eigendecomposition $K = U\Lambda U^T$. If Y is any zero-mean random vector with $R_Y = I_n$ and we set $X := U\sqrt{\Lambda}Y$ then $R_X = \mathrm{Cov}(X) = U\sqrt{\Lambda}R_Y\sqrt{\Lambda}U^T = K$.

Decorrelating linear transformation

- Let X be a random vector with $E[X] = \mu_X$ and $\text{Cov}(X) = K$.
- Let $K = U \Lambda U^T$ be the eigendecomposition of K.
- Define a new random vector Y via the following "change of coordinates": $Y = U^T(X - \mu_X)$.
  Note-1: Subtracting the mean shifts the origin to the mean.
  Note-2: Multiplying by U^T is like rotating the coordinate system: If $b_1 = U^T a_1$ and $b_2 = U^T a_2$ then since $UU^T = I$ (U is an orthonormal matrix), $b_1^T b_2 = a_1^T a_2$, i.e., U^T preserves angles and lengths.
- Then $E[Y] = 0$ and $R_Y = \text{Cov}(Y) = U^T K U = \Lambda$ a diagonal matrix.
- The components of Y are uncorrelated!
- U^T called the **(Kosambi) Karhunen-Loeve transform (KLT)**.
- Application: **Principal Component Analysis (PCA)** in statistical signal processing and machine learning.

Covariance singularity and deterministic linear dependency

- A covariance matrix is nonsingular/invertible iff it is positive definite.
- $X = (X_1, \ldots, X_n)^T$ has a **deterministic linear dependency** if $a_0 1 + \sum_{i=1}^n a_i X_i = 0$ with probability one for some constants $a_0, a_1, \ldots, a_n$, not all zero. Compactly, $a_0 1 + a^T X = 0$ where $a = (a_1, \ldots, a_n)^T$.
- **Result:** X has a deterministic linear dependency iff Cov(X) is singular.
  Proof: Exercise.
- **Example 6** The covariance matrix of a 2-dimensional random vector $W = (X, Y)^T$ is of the form:

$$\Sigma_W = \begin{pmatrix} \sigma_X^2 & \sigma_X \sigma_Y \rho_{XY} \\ \sigma_X \sigma_Y \rho_{XY} & \sigma_Y^2 \end{pmatrix}$$

  where ρ_{XY} is the correlation coefficient of X and Y. If $\sigma_X, \sigma_Y > 0$, Σ_W is singular iff $\det(\Sigma_W) = (1 - \rho_{XY}^2)\sigma_X^2 \sigma_Y^2 = 0 \Leftrightarrow \rho_{XY} = \pm 1 \Leftrightarrow X = aY + b$, $a \neq 0, b$ const.

Covariance singularity and determistic linear dependency

- **Caution:** If the covariance matrix of a random vector is nonsingular, it is still possible that there is a deterministic **nonlinear** dependency among the component random variables as the following example shows:

Example 7 Let $X_1 \sim \mathcal{N}(0,1)$ and $X_2 = X_1^2$. Then since $E[X_1] = E[X_1^3] = 0$, $E[X_2] = E[X_1^2] = 1$, and $E[X_2^2] = E[X_1^4] = 3$,

$$\text{Cov}((X_1, X_2)^T) = \begin{pmatrix} 1 & 0 \\ 0 & 3-1 \end{pmatrix}$$

which is nonsingular even though there is a deterministic **nonlinear** dependency between X_1 and X_2.

Jointly Gaussian random variables/vectors

- A collection of random variables $(X_i : i \in \mathcal{I})$ is **jointly** Gaussian $\Leftrightarrow$ every finite linear combination is a scalar Gaussian random variable:
  For all n, all $i_1, i_2, \ldots, i_n \in \mathcal{I}$, and all $a = (a_1, \ldots, a_n)^T \in \mathbb{R}^n$, $\sum_{j=1}^n a_j X_{i_j} \sim \mathcal{N}(\mu, \sigma^2)$ where $\mu = \sum_{j=1}^n a_j E[X_{i_j}]$ and $\sigma^2 = a^T \mathrm{Cov}((X_{i_1}, \ldots, X_{i_n})^T) a$.
- A collection of random vectors is jointly Gaussian $\Leftrightarrow$ the collection of all components of all vectors is jointly Gaussian.
- If $(X_i : i \in \mathcal{I})$ is jointly Gaussian then so is:
  1. $(X_j : j \in \mathcal{J})$ for all $\mathcal{J} \subseteq \mathcal{I}$. In particular, each X_i is Gaussian.
  2. The collection of all finite linear combinations of X_i's.
  3. The collection of all limits of sequences of X_i's.
- If each X_i is Gaussian and $(X_i : i \in \mathcal{I})$ are independent then they are jointly Gaussian. Without the independence condition the result may not be true.

Jointly Gaussian random variables/vectors

- $X = (X_1, \dots, X_n)^T$ Gaussian $\Leftrightarrow$

$$\Phi_X(v) = E[e^{jv^T X}] = e^{jv^T E[X] - \frac{1}{2} v^T \mathsf{Cov}(X) v}.$$

  Thus the distribution of a collection of jointly Gaussian random variables is completely specified by their means and covariances.

- We say that X is a $\mathcal{N}(\mu, K)$ random vector if X is a Gaussian random vector with mean vector μ and covariance matrix K.

- If $X = (X_1, \dots, X_n)^T$ is Gaussian then

$$\mathsf{Cov}(X) \text{ diagonal } \Leftrightarrow \perp\!\!\!\perp (X_1, \dots, X_n).$$

- If X, Y are jointly Gaussian random vectors then

$$X \perp\!\!\!\perp Y \Leftrightarrow \mathsf{Cov}(X, Y) = 0.$$

Jointly Gaussian random variables/vectors

- If $X = (X_1, \ldots, X_n)^T$ is Gaussian and $\mathsf{Cov}(X) > 0$ (non-singular), then X is continuous with pdf

$$f_X(x) = \frac{\exp\left\{-\frac{1}{2}\,(x - E[X])^T\,(\mathsf{Cov}(X))^{-1}\,(x - E[X])\right\}}{\sqrt{(2\pi)^n \det(\mathsf{Cov}(X))}}.$$

- The contours of constant pdf value $\{x \in \mathbb{R}^n : f_X(x) = \text{constant}\}$ are given by the equation:

$$(x - \mu_X)^T \Sigma_X^{-1} (x - \mu_X) = \text{ constant}$$

  which are concentric ellipsoids centered at μ_X with principal axes given by the eigenvectors of $\mathsf{Cov}(X)$.

- Σ_X is a diagonal matrix iff the principal axes are aligened with the coordinate axes and then the components are all independent.
- $\mathsf{Cov}(X) = \sigma^2 I_n$ iff the contours are concentric spheres. Then X is called a **spherical/white Gaussian** random vector.

Jointly Gaussian random variables/vectors

If $X \in \mathbb{R}^m$ and $Y \in \mathbb{R}^n$ are jointly Gaussian random vectors then $X|Y = y$ is also a Gaussian random vector with

- (Conditional) mean vector $\mu_{X|y}$:

$$\begin{aligned}\mu_{X|y} &= E[X|Y = y] = E[X] + \mathsf{Cov}(X, Y)(\mathsf{Cov}(Y))^{-1}(y - E[Y]) \\ &= \mu_X + \Sigma_{XY}\Sigma_Y^{-1}(y - \mu_Y).\end{aligned}$$

- (Conditional) covariance matrix $\Sigma_{X|y}$:

$$\begin{aligned}\Sigma_{X|y} &= \mathsf{Cov}(X - \mu_{X|y}|Y = y) \\ &= \mathsf{Cov}(X) - \mathsf{Cov}(X, Y)(\mathsf{Cov}(Y))^{-1}\mathsf{Cov}(Y, X) \\ &= \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}.\end{aligned}$$

Note: $\mu_{X|y}$ depends on the value of y but $\Sigma_{X|y}$ does not.

Individually Gaussian $\nRightarrow$ jointly Gaussian

- If X is a Gaussian random vector and Cov$(X) > 0$ (positive definite) then its pdf, being of exponential form, cannot be zero anywhere.
- Let $X = (X_1, X_2)^T$ with joint pdf:

$$f_X(x) = f_{X_1X_2}(x_1, x_2) = \begin{cases} 0 & x_1 \cdot x_2 < 0 \\ \frac{2}{2\pi} e^{-\frac{1}{2}(x_1^2 + x_2^2)} & \text{otherwise.} \end{cases}$$

- $f_X(x) = 0$ in quadrants 2, 4 of the x_1–x_2 plane and $f_X(x) = 2\mathcal{N}(0, I_2)$ in quadrants 1, 3. It is as if the probability mass of $\mathcal{N}(0, I_2)$ from the second quadrant has been "folded" into the first and all the mass from the fourth into the third.
- From this and a little thought it follows that the marginal pdfs of both X_1 and X_2 are $\mathcal{N}(0, 1)$.
- Since the joint pdf is symmetric about the origin: $f_X(x) = f_X(-x)$, Cov$(X_1, X_2) = 0$. Thus Cov$(X) = I_2$.
- Since Cov$(X) > 0$ and the pdf is zero in quadrants 2, 4, X_1, X_2 cannot be jointly Gaussian, yet each of them are individually!

Laws of large numbers

Weak Law of Large Numbers (WLLN):

- Let $X_1, X_2, \ldots$ be a sequence of IID random variables with finite mean $\mu = E[X_i] < \infty$.
- Let $\widehat{\mu}_n = \frac{1}{n}\Big(X_1 + \ldots + X_n\Big)$ denote the sample mean.
- For any $\epsilon > 0$, the WLLN implies that

$$\lim_{n\to\infty} P\Big(\big|\widehat{\mu}_n - \mu\big| \geq \epsilon\Big) = 0 \ .$$

- That is, the sample mean converges (in probability) to the true mean.

Laws of large numbers

Central Limit Theorem (CLT):

- Let $X_1, X_2, \ldots$ be a sequence of IID random variables with finite mean μ and finite variance σ^2.
- Let

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{(X_i - \mu)}{\sigma}$$

  denote the normalized sum which has zero mean and unit variance for each n.
- The CLT implies that for all $z \in (-\infty, \infty)$,

$$\lim_{n \to \infty} P\Big(Z_n \le z\Big) = \lim_{n \to \infty} F_{Z_n}(z) = 1 - Q(z)$$

  where $1 - Q(z)$ is the CDF of a $\mathcal{N}(0, 1)$ RV.
- That is, the normalized sum converges (in distribution) to a standard Gaussian (normal) RV.

Confidence intervals

- Let $X_1, X_2, \ldots$ be a sequence of IID random variables with finite mean μ and finite variance σ^2.
- Let $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \widehat{\sigma^2}_n = \frac{1}{n}\sum_{i=1}^{n}\Big(X_i - \widehat{\mu}_n\Big)^2$ denote the empirical estimates of the mean and variance respectively.
- Then, $E\Big[\widehat{\mu}_n\Big] = \mu, \quad \mathrm{var}\Big(\widehat{\mu}_n\Big) = \frac{1}{n}\sigma^2, \quad E\Big[\widehat{\sigma^2}_n\Big] = \frac{n-1}{n}\sigma^2$
- The CLT implies that for all sufficiently large n,

$$P\Big(\big|\widehat{\mu}_n - \mu\big| \geq \tau\Big) \approx 2Q\Big(\frac{\tau}{\sigma}\sqrt{n}\Big)$$

$$\Rightarrow P\Big(\widehat{\mu}_n \in (\mu - \tau, \mu + \tau)\Big) > \begin{cases} 0.68 & \text{if } \tau = \sigma/\sqrt{n} \\ 0.95 & \text{if } \tau = 2\sigma/\sqrt{n} \\ 0.99 & \text{if } \tau = 3\sigma/\sqrt{n}. \end{cases}$$

In practice, σ is replaced by $\widehat{\sigma}_n$ or $\widehat{\sigma}_n\sqrt{n/(n-1)}$ (if $n > 1$).

Distribution-free bounds

- **Hoeffding's inequality:** Let $X_1, \ldots, X_n$ be independent RVs with $X_i \in [a_i, b_i]$ with certainty for each i. If $S_n = \sum_{i=1}^n X_i$ then for all $\epsilon > 0$,

$$P\Big(S_n - E[S_n] \geq \epsilon\Big) \leq e^{-2\epsilon^2 / \sum_{i=1}^n (b_i - a_i)^2}$$
$$P\Big(E[S_n] - S_n \leq -\epsilon\Big) \leq e^{-2\epsilon^2 / \sum_{i=1}^n (b_i - a_i)^2}$$

- **McDiarmid's inequality:** Let $X_1, \ldots, X_n$ be independent RVs and g a function of n variables whose value does not change by more than c_i if only the i-th variable is changed keeping others fixed. Then for all $\epsilon > 0$,

$$P\Big(g(X_1, \ldots, X_n) - E[g(X_1, \ldots, X_n)] \geq \epsilon\Big) \leq e^{-2\epsilon^2 / \sum_{i=1}^n c_i^2}$$
$$P\Big(E[g(X_1, \ldots, X_n)] - g(X_1, \ldots, X_n) \leq -\epsilon\Big) \leq e^{-2\epsilon^2 / \sum_{i=1}^n c_i^2}$$