Introduction: Topics Covered.

- What is Probability
- Set Theory Basics
- Probability Models
- Conditional Probability
- Total Probability and Bayes Rule
- Independence
- Counting

What is Probability?

- Measured relative frequency of occurrence of an event.
  Example: toss a coin 100 times, measure frequency of heads or compute probability of raining on a particular day and month (using past years' data)

- Or subjective belief about how "likely" an event is (when do not have data to estimate frequency).
  Example: any one-time event in history or "how likely is it that a new experimental drug will work?"
  This may either be a subjective belief or derived from the physics, for e.g. if I flip a symmetric coin (equal weight on both sides), I will get a head with probability $1/2$.

- For probabilistic reasoning, **two** types of problems need to be solved

1. Specify the probability “model” or learn it (covered in a statistics class).
2. Use the “model” to compute probability of different events (covered here).

- We will assume the model is given and will focus on problem 2.

Set Theory Basics

- Set: any collection of objects (elements of a set).
- Discrete sets
  - Finite number of elements, e.g. numbers of a die
  - Or infinite but countable number of elements, e.g. set of integers
- Continuous sets
  - Cannot count the number of elements, e.g. all real numbers between 0 and 1.
- "Universe" (denoted Ω): consists of all possible elements that could be of interest. In case of random experiments, it is the set of all possible outcomes. Example: for coin tosses, $\Omega = \{H, T\}$.
- Empty set (denoted ϕ): a set with no elements

Set Theory Basics

$$A \cup B = \{\xi \mid \xi \in A \quad \text{or} \quad \xi \in B\}$$

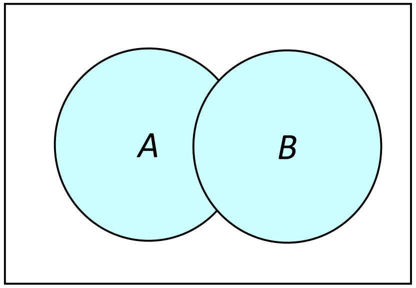
$$A \cap B = \{\xi \mid \xi \in A \quad \text{and} \quad \xi \in B\}$$

$$\overline{A} \quad = \{\xi \mid \xi \notin A\}$$

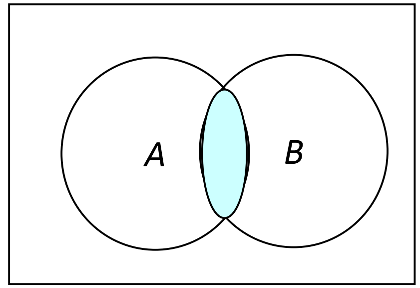
- If $A \cap B = \phi,$ the empty set, then A and B are said to be mutually exclusive (M.E).
- A partition of Ω is a collection of mutually exclusive subsets of Ω such that their union is Ω.

$$A_i \cap A_j = \phi, \quad \text{and} \quad \bigcup_{i=1} A_i = \Omega.$$

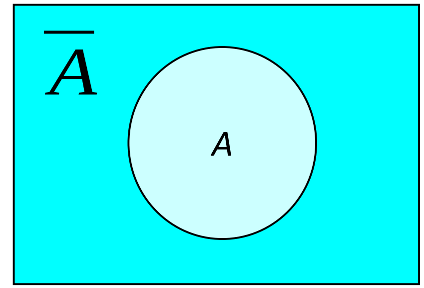
Using Venn Diagram



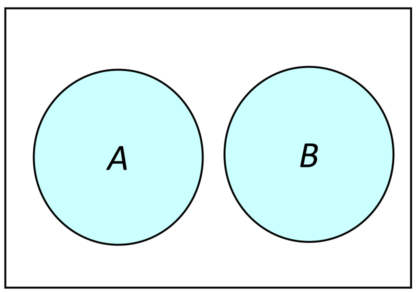
$A \cup B$



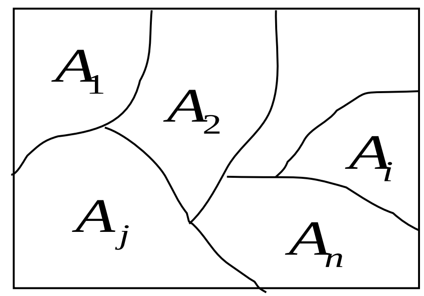
$A \cap B$



$\overline{A}$



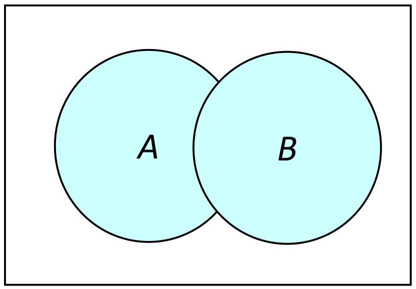
$A \cap B = \phi$



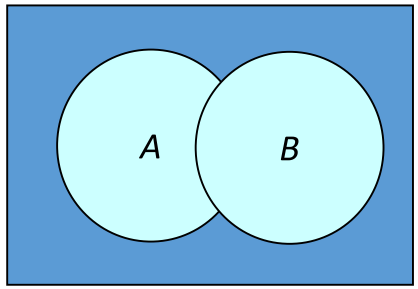
$A_i \cap A_j = \phi, \quad \text{and} \quad \bigcup_{i=1} A_i = \Omega.$

De-Morgan's Laws:

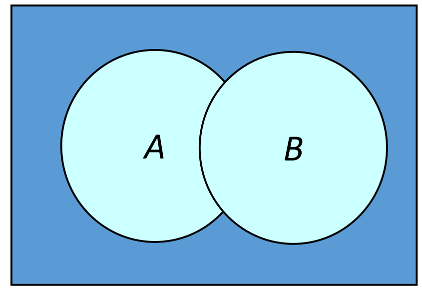
$$\overline{A\cup B}=\overline{A}\cap\overline{B}; \qquad \overline{A\cap B}=\overline{A}\cup\overline{B}$$



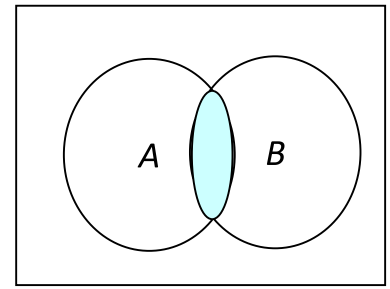
$A\cup B$



$\overline{A\cup B}$



$\overline{A}\cap\overline{B}$



$A\cap B$

Example: Consider the experiment where two coins are simultaneously tossed. Elementary events are ---

$$\xi_1=(H,H),\ \ \xi_2=(H,T),\ \ \xi_3=(T,H),\ \ \xi_4=(T,T)$$

$$W=\{x_1,x_2,x_3,x_4\} \qquad A=\{\xi_1,\xi_2,\xi_3\}$$

- Subset: $A \subseteq B$: if every element of A also belongs to B.
- Strict subset: $A \subset B$: if every element of A also belongs to B and B has more elements than A.
- Belongs: $\in$, Does not belong: $\notin$
- Complement: A' or A^c, Union: $A \cup B$, Intersection: $A \cap B$
  - $A' \triangleq \{x \in \Omega | x \notin A\}$
  - $A \cup B \triangleq \{x | x \in A, or\ x \in B\}$, $x \in \Omega$ is assumed.
  - $A \cap B \triangleq \{x | x \in A, and\ x \in B\}$
- **Disjoint sets: A and B are disjoint if $A \cap B = \phi$ (empty), i.e. they have no common elements.**

- DeMorgan's Laws

$$(A \cup B)' = A' \cap B' \quad (1)$$

$$(A \cap B)' = A' \cup B' \quad (2)$$

  - Proofs: Need to show that every element of LHS (left hand side) is also an element of RHS (right hand side), i.e. LHS $\subseteq$ RHS and show vice versa, i.e. RHS $\subseteq$ LHS.
  - We show the proof of the first property
    * If $x \in (A \cup B)'$, it means that x does not belong to A or B. In other words x does not belong to A and x does not B either. This means x belongs to the complement of A and to the complement of B, i.e. $x \in A' \cap B'$.
    * Just showing this much does not complete the proof, need to show the other side also.
    * If $x \in A' \cap B'$, it means that x does not belong to A and it does not

belong to B, i.e. it belongs to neither A nor B, i.e. $x \in (A \cup B)'$

* This completes the argument

Probabilistic models

- There is an underlying process called **experiment** that produces exactly ONE **outcome**.
- A probabilistic model: consists of a sample space and a probability law
  - Sample space (denoted Ω): set of all possible outcomes of an experiment
  - Event: any subset of the sample space
  - Probability Law: assigns a probability to every set A of possible outcomes (event)
  - Choice of sample space (or universe): every element should be distinct and mutually exclusive (disjoint); and the space should be "collectively exhaustive" (every possible outcome of an experiment should be included).

- **Probability Axioms:**
  1. **Nonnegativity.** $P(A) \geq 0$ for every event A.
  2. **Additivity.** If A and B are two **disjoint** events, then
     $P(A \cup B) = P(A) + P(B)$
     (also extends to any countable number of disjoint events).
  3. **Normalization.** Probability of the entire sample space, $P(\Omega) = 1$.
- Probability of the empty set, $P(\phi) = 0$ (follows from Axioms 2 & 3).
- Discrete probability law: sample space consists of a finite number of possible outcomes, law specified by probability of single element events.
  - Example: for a fair coin toss, $\Omega = \{H, T\}$, $P(H) = P(T) = 1/2$
  - Discrete uniform law for any event A:

$$P(A) = \frac{\text{number of elements in A}}{n}$$

- Continuous probability law: e.g. $\Omega = [0, 1]$: probability of any single element event is zero, need to talk of probability of a subinterval, $[a, b]$ of $[0, 1]$.

  Properties of probability laws

  1. If $A \subseteq B$, then $P(A) \leq P(B)$
  2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  3. $P(A \cup B) \leq P(A) + P(B)$
  4. $P(A \cup B \cup C) = P(A) + P(A' \cap B) + P(A' \cap B' \cap C)$
  5. Note: Some books use A^c for A' (complement of set A).

Conditional Probability

- Given that we know that an event B has occurred, what is the probability that event A occurred? Denoted by $P(A|B)$. Example: Roll of a 6-sided die. Given that the outcome is even, what is the probability of a 6? Answer: 1/3
- When number of outcomes is finite and all are equally likely,

$$P(A|B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B} \tag{3}$$

- In general,

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)} \tag{4}$$

- $P(A|B)$ is a probability law (satisfies axioms) on the universe B. Exercise: show this.

Total Probability and Bayes Rule

- Total Probability Theorem: Let $A_1, \ldots A_n$ be disjoint events which form a partition of the sample space ($\cup_{i=1}^{n} A_i = \Omega$). Then for any event B,

$$\begin{aligned} P(B) &= P(A_1 \cap B) + \ldots P(A_n \cap B) \\ &= P(A_1)P(B|A_1) + \ldots P(A_n)P(B|A_n) \end{aligned} \tag{5}$$

- Bayes rule: Let $A_1, \ldots A_n$ be disjoint events which form a partition of the sample space. Then for any event B, s.t. $P(B) > 0$, we have

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(B)} = \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1) + \ldots P(A_n)P(B|A_n)} \tag{6}$$

- Inference using Bayes rule
  - There are multiple "causes" $A_1, A_2, ..A_n$ that result in a certain "effect" B. Given that we observe the effect B, what is the probability that the cause was A_i? Answer: use Bayes rule.
  - Radar detection: what is the probability of the aircraft being present given that the radar registers it?

Independence

- $P(A|B) = P(A)$ and so $P(A \cap B) = P(B)P(A)$: the fact that B has occurred gives no information about the probability of occurrence of A. Example: A= head in first coin toss, B = head in second coin toss.
- **"Independence": DIFFERENT from "mutually exclusive" (disjoint)**
  - Events A and B are disjoint if $P(A \cap B) = 0$: cannot be independent if $P(A) > 0$ and $P(B) > 0$.

    Example: A = head in a coin toss, B = tail in a coin toss
  - Independence: a concept for events in a sequence. Independent events with $P(A) > 0, P(B) > 0$ cannot be disjoint
- Independence of a collection of events

- $P(\cap_{i \in S} A_i) = \Pi_{i \in S} P(A_i)$ for every subset S of $\{1, 2, ..n\}$

- Reliability analysis of complex systems: independence assumption often simplifies calculations
  - What is $P(\text{system fails})$ of the system $A \to B$?$*$ Let p_i = probability of success of component i.
    - $*$ m components in series: $P(\text{system fails}) = 1 - p_1 p_2 \ldots p_m$ (succeeds if all components succeed).
    - $*$ m components in parallel: $P(\text{system fails}) = (1 - p_1) \ldots (1 - p_m)$ (fails if all the components fail).

- Independent Bernoulli trials and Binomial probabilities
  - A Bernoulli trial: a coin toss (or any experiment with two possible outcomes, e.g. it rains or does not rain, bit values)
  - Independent Bernoulli trials: sequence of independent coin tosses

- Binomial: Given n independent coin tosses,what is the probability of k heads (denoted $p(k)$)?
  - probability of any one sequence with k heads is $p^k(1-p)^{n-k}$
  - number of such sequences (from counting arguments): $\left(\begin{array}{c} n \\ k \end{array}\right)$
  - $p(k) = \left(\begin{array}{c} n \\ k \end{array}\right) p^k(1-p)^{n-k}$, where $\left(\begin{array}{c} n \\ k \end{array}\right) \triangleq \frac{n!}{(n-k)!k!}$
- Application: what is the probability that more than c customers need an internet connection at a given time? We know that at a given time, the probability that any one customer needs connection is p.

  Answer: $\displaystyle\sum_{k=c+1}^{n} p(k)$

1 What is a random variable (r.v.)?

- A real valued function of the outcome of an experiment
- Example: Coin tosses. r.v. $X = 1$ if heads and $X = 0$ if tails (Bernoulli r.v.).
- A function of a r.v. defines another r.v.
- Discrete r.v.: X takes values from the set of integers

2 Discrete Random Variables & Probability Mass Function (PMF)

- **Probability Mass Function (PMF):** Probability that the r.v. X takes a value x is PMF of X computed at $X = x$. Denoted by $p_X(x)$. Thus

$$p_X(x) = P(\{X = x\}) = P(\text{all possible outcomes that result in the event } \{X = x\}) \tag{1}$$

- Everything that we learnt in Chap 1 for events applies. Let Ω is the sample space (space of all possible values of X in an experiment). Applying the axioms,
  - $p_X(x) \geq 0$
  - $P(\{X \in S\}) = \sum_{x \in S} p_X(x)$ (follows from Additivity since different events $\{X = x\}$ are disjoint)
  - $\sum_{x \in \Omega} p_X(x) = 1$ (follows from Additivity and Normalization).
  - Example: X = number of heads in 2 fair coin tosses ($p = 1/2$). $P(X > 0) = \sum_{x=1}^{2} p_X(x) = 0.75$.
- Can also define a binary r.v. for any event A as: $X = 1$ if A occurs and $X = 0$ otherwise. Then X is a Bernoulli r.v. with $p = P(A)$.
- Bernoulli ($X = 1$ (heads) or $X = 0$ (tails)) r.v. with probability of heads p

$$Bernoulli(p): \ p_X(x) = p^x (1-p)^{1-x}, \ x = 0, \ or \ x = 1 \tag{2}$$

- Binomial ($X = x$ heads out of n independent tosses, probability of heads p)

$$Binomial(n,p): \ p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \ x = 0, 1, \ldots n \tag{3}$$

- Geometric r.v., X, with probability of heads p (X= number of coin tosses needed for a head to come up for the first time or number of independent trials needed to achieve the first "success").

- Example: I keep taking a test until I pass it. Probability of passing the test in the x^{th} try is $p_X(x)$.
- Easy to see that

$$Geometric(p): \; p_X(x) = (1-p)^{x-1}p, \quad x = 0, 1, 2, \ldots \infty \tag{4}$$

- Poisson r.v. X with expected number of arrivals Λ (e.g. if X = number of arrivals in time τ with arrival rate λ, then $\Lambda = \lambda\tau$)

$$Poisson(\Lambda): \; p_X(x) = \frac{e^{-\Lambda}(\Lambda)^x}{x!}, \; x = 0, 1, \ldots \infty \tag{5}$$

- Uniform(a,b):

$$p_X(x) = \begin{cases} 1/(b-a+1), & if \; x = a, a+1, \ldots b \\ 0, & otherwise \end{cases} \tag{6}$$

- pmf of $Y = g(X)$
  - $p_Y(y) = P(\{Y = y\}) = \sum_{x|g(x)=y} p_X(x)$

    Example $Y = |X|$. Then $p_Y(y) = p_X(y) + p_X(-y), \; if \; y > 0$ and $p_Y(0) = p_X(0)$. Exercise: $X \sim Uniform(-4, 4)$ and $Y = |X|$, find $p_Y(y)$.

- Expectation, mean, variance
  - Motivating example: Read pg 81
  - Expected value of X (or mean of X): $E[X] \triangleq \sum_{x \in \Omega} x p_X(x)$
  - Interpret mean as center of gravity of a bar with weights $p_X(x)$ placed at location x
  - Expected value of $Y = g(X)$: $E[Y] = E[g(X)] = \sum_{x \in \Omega} g(x) p_X(x)$. Exercise: show this.
  - n^{th} moment of X: $E[X^n]$. n^{th} central moment: $E[(X - E[X])^n]$.
  - Variance of X: $var[X] \triangleq E[(X - E[X])^2]$ (2nd central moment)
  - $Y = aX + b$ (linear fn): $E[Y] = aE[X] + b$, $var[Y] = a^2 var[X]$
  - Poisson: $E[X] = \Lambda$, $var[X] = \Lambda$ (show this)
  - Bernoulli: $E[X] = p$, $var[X] = p(1-p)$ (show this)
  - Uniform(a,b): $E[X] = (a+b)/2$, $var[X] = \frac{(b-a+1)^2-1}{12}$ (show this)

3 Multiple Discrete Random Variables: Topics

- Joint PMF, Marginal PMF of 2 and or more than 2 r.v.'s
- PMF of a function of 2 r.v.'s
- Expected value of functions of 2 r.v's
- Expectation is a linear operator. Expectation of sums of n r.v.'s
- Conditioning on an event and on another r.v.
- Bayes rule
- Independence

4 Joint & Marginal PMF, PMF of function of r.v.s, Expectation

- For everything in this handout, you can think in terms of events $\{X = x\}$ and $\{Y = y\}$ and apply what you have learnt in Chapter 1.
- The **joint PMF** of two random variables X and Y is defined as

$$p_{X,Y}(x,y) \triangleq P(X = x, Y = y)$$

  where $P(X = x, Y = y)$ is the same as $P(\{X = x\} \cap \{Y = y\})$.

  - Let A be the set of all values of x, y that satisfy a certain property, then $P((X,Y) \in A) = \sum_{(x,y)\in A} p_{X,Y}(x,y)$
  - e.g. X = outcome of first die toss, Y is outcome of second die toss, A = sum of outcomes of the two tosses is even.

- **Marginal PMF** is another term for the PMF of a single r.v. obtained by **"marginalizing"** the joint PMF over the other r.v., i.e. the marginal PMF of X, $p_X(x)$ can be computed as follows:
  Apply Total Probability Theorem to $p_{X,Y}(x,y)$, i.e. sum over $\{Y = y\}$ for different values y (these are a set of disjoint events whose union is the sample space):

$$p_X(x) = \sum_y p_{X,Y}(x,y)$$

  Similarly the marginal PMF of Y, $p_Y(y)$ can be computed by "marginalizing" over X

$$p_Y(y) = \sum_x p_{X,Y}(x,y)$$

- **PMF of a function of r.v.'s:** If $Z = g(X,Y)$,

$$p_Z(z) = \sum_{(x,y):g(x,y)=z} p_{X,Y}(x,y)$$

  - Read the above as $p_Z(z) = P(Z = z) = P(\text{all values of } (X,Y) \text{ for which } g(X,Y) = z)$

- **Expected value of functions of multiple r.v.'s**
  If $Z = g(X,Y)$,
$$E[Z] = \sum_{(x,y)} g(x,y) p_{X,Y}(x,y)$$

- **More than 2 r.v.s.**
  - Joint PMF of n r.v.'s: $p_{X_1,X_2,\ldots X_n}(x_1, x_2, \ldots x_n)$
  - We can **marginalize** over one or more than one r.v.,

    e.g. $p_{X_1,X_2,\ldots X_{n-1}}(x_1, x_2, \ldots x_{n-1}) = \sum_{x_n} p_{X_1,X_2,\ldots X_n}(x_1, x_2, \ldots x_n)$
    e.g. $p_{X_1,X_2}(x_1, x_2) = \sum_{x_3,x_4,\ldots x_n} p_{X_1,X_2,\ldots X_n}(x_1, x_2, \ldots x_n)$
    e.g. $p_{X_1}(x_1) = \sum_{x_2,x_3,\ldots x_n} p_{X_1,X_2,\ldots X_n}(x_1, x_2, \ldots x_n)$

- **Expectation is a linear operator.** *Exercise: show this*

$$E[a_1X_1 + a_2X_2 + \ldots a_nX_n] = a_1E[X_1] + a_2E[X_2] + \ldots a_nE[X_n]$$

5 Conditioning and Bayes rule

- **PMF of r.v. X conditioned on an event A with $P(A) > 0$**
$$p_{X|A}(x) \triangleq P(\{X = x\}|A) = \frac{P(\{X = x\} \cap A)}{P(A)}$$
  - $p_{X|A}(x)$ is a legitimate PMF, i.e. $\sum_x p_{X|A}(x) = 1$.

- **PMF of r.v. X conditioned on r.v. Y.** Replace A by $\{Y = y\}$
$$p_{X|Y}(x|y) \triangleq P(\{X = x\}|\{Y = y\}) = \frac{P(\{X = x\} \cap \{Y = y\})}{P(\{Y = y\})} = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$
  The above holds for all y for which $p_y(y) > 0$. The above is equivalent to
$$p_{X,Y}(x,y) = p_{X|Y}(x|y)p_Y(y)$$
$$p_{X,Y}(x,y) = p_{Y|X}(y|x)p_X(x)$$
  - $p_{X|Y}(x|y)$ (with $p_Y(y) > 0$) is a legitimate PMF, i.e. $\sum_x p_{X|Y}(x|y) = 1$.
  - Similarly, $p_{Y|X}(y|x)$ is also a legitimate PMF, i.e. $\sum_y p_{Y|X}(y|x) = 1$. *Show this.*

- **Bayes rule.** How to compute $p_{X|Y}(x|y)$ using $p_X(x)$ and $p_{Y|X}(y|x)$,

$$\begin{aligned} p_{X|Y}(x|y) &= \frac{p_{X,Y}(x,y)}{p_Y(y)} \\ &= \frac{p_{Y|X}(y|x)p_X(x)}{\sum_{x'} p_{Y|X}(y|x')p_X(x')} \end{aligned}$$

- **Conditional Expectation given event A**

$$E[X|A] = \sum_x x p_{X|A}(x)$$

$$E[g(X)|A] = \sum_x g(x) p_{X|A}(x)$$

- **Conditional Expectation given r.v. $Y = y$.** Replace A by $\{Y = y\}$

$$E[X|Y = y] = \sum_x x p_{X|Y}(x|y)$$

Note this is a function of $Y = y$.

- **Total Expectation Theorem**

$$E[X] = \sum_y p_Y(y) E[X|Y = y]$$

- **Total Expectation Theorem for disjoint events $A_1, A_2, \ldots A_n$ which form a partition of sample space.**

$$E[X] = \sum_{i=1}^{n} P(A_i) E[X|A_i]$$

Note A_i's are disjoint and $\cup_{i=1}^{n} A_i = \Omega$

6 Independence

- **Independence of a r.v. & an event A.** r.v. X is independent of A with $P(A) > 0$, iff

$$p_{X|A}(x) = p_X(x), \quad \text{for all } x$$

.

  - This also implies: $P(\{X = x\} \cap A) = p_X(x) P(A)$.
  -

- **Independence of 2 r.v.'s.** R.v.'s X and Y are independent iff

$$p_{X|Y}(x|y) = p_X(x), \text{ for all } x \text{ and for all } y \text{ for which } p_Y(y) > 0$$

  This is equivalent to the following two things(*show this*)

$$p_{X,Y}(x,y) = p_X(x)p_Y(y)$$

$$p_{Y|X}(y|x) = p_Y(y), \text{ for all } y \text{ and for all } x \text{ for which } p_X(x) > 0$$

- **Conditional Independence of r.v.s X and Y given event A with** $P(A) > 0$ **
  $p_{X|Y,A}(x|y) = p_{X|A}(x)$ for all x and for all y for which $p_{Y|A}(y) > 0$ or that
  $p_{X,Y|A}(x,y) = p_{X|A}(x)p_{Y|A}(y)$

- **Expectation of product of independent r.v.s.**
  - If X and Y are independent, $E[XY] = E[X]E[Y]$.

$$\begin{aligned} E[XY] &= \sum_y \sum_x xy p_{X,Y}(x,y) \\ &= \sum_y \sum_x xy p_X(x) p_Y(y) \\ &= \sum_y y p_Y(y) \sum_x x p_X(x) \\ &= E[X]E[Y] \end{aligned}$$

  - If X and Y are independent, $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$. (Show).

- If $X_1, X_2, \ldots X_n$ are independent,

$$p_{X_1,X_2,\ldots X_n}(x_1, x_2, \ldots x_n) = p_{X_1}(x_1)p_{X_2}(x_2)\ldots p_{X_n}(x_n)$$

- **Variance of sum of 2 independent r.v.'s.**
  Let X, Y are independent, then $Var[X+Y] = Var[X] + Var[Y]$.

- **Variance of sum of n independent r.v.'s.**
  If $X_1, X_2, \ldots X_n$ are independent,

$$Var[X_1 + X_2 + \ldots X_n] = Var[X_1] + Var[X_2] + \ldots Var[X_n]$$

  .

  - **Application: Variance of a Binomial**, See Example 2.20
    Binomial r.v. is a sum of n independent Bernoulli r.v.'s. So its variance is $np(1-p)$
  - **Application: Mean and Variance of Sample Mean**, Example 2.21
    Let $X_1, X_2, \ldots X_n$ be independent and *identically distributed*, i.e. $p_{X_i}(x) = p_{X_1}(x)$ for all i. Thus all have the same mean (denote by a) and same variance (denote by v).
    Sample mean is defined as $S_n = \frac{X_1+X_2+\ldots X_n}{n}$.
    Since $E[.]$ is a linear operator, $E[S_n] = \sum_{i=1}^n \frac{1}{n} E[X_i] = \frac{na}{n} = a$.
    Since the X_i's are independent, $Var[S_n] = \sum_{i=1}^n \frac{1}{n^2} Var[X_i] = \frac{nv}{n^2} = \frac{v}{n}$

1 Continuous R.V. & Probability Density Function (PDF)

- A r.v. X is called **continuous** if there is a function $f_X(x)$ with $f_X(x) \geq 0$, called **probability density function (PDF)**, s.t. $P(X \in B) = \int_B f_X(x)dx$ for all subsets B of the real line.
- Specifically, for $B = [a, b]$,

$$P(a \leq X \leq b) = \int_{x=a}^{b} f_X(x)dx \tag{1}$$

and can be interpreted as the area under the graph of the PDF $f_X(x)$.

- For any single value a, $P(\{X = a\}) = \int_{x=a}^{a} f_X(x)dx = 0$.
- Thus $P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$
- Sample space $\Omega = (-\infty, \infty)$
- Normalization: $P(\Omega) = P(-\infty < X < \infty) = 1$. Thus $\int_{x=-\infty}^{\infty} f_X(x)dx = 1$
- Interpreting the PDF: For an interval $[x, x + \delta]$ with very small δ,

$$P([x, x + \delta]) = \int_{t=x}^{x+\delta} f_X(t)dt \approx f_X(x)\delta \tag{2}$$

Thus $f_X(x)$= probability mass per unit length near x.

- Expected value: $E[X] = \int_{x=-\infty}^{\infty} x f_X(x)dx$. Similarly define $E[g(X)]$ and $var[X]$
- Mean and variance of uniform, Example 3.4
- Exponential r.v.

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & if\ x \geq 0 \\ 0, & otherwise \end{cases} \tag{3}$$

  - Show it is a legitimate PDF.
  - $E[X] = 1/\lambda$, $var[X] = 1/\lambda^2$ (show).
  - Example: X= amount of time until an equipment breaks down or a bulb burns out.
  - Example 3.5 (Note: you need to use the correct time unit in the problem, here days).

2 Cumulative Distribution Function (CDF)

- Cumulative Distribution Function (CDF), $F_X(x) \triangleq P(X \leq x)$ (probability of event $\{X \leq x\}$).
- Defined for discrete and continuous r.v.'s

$$\text{Discrete:} \quad F_X(x) = \sum_{k \leq x} p_X(k) \tag{4}$$

$$\text{Continuous:} \quad F_X(x) = \int_{-\infty}^{x} f_X(t)dt \tag{5}$$

- Note the PDF $f_X(x)$ is NOT a probability of any event, it can be > 1.
- But $F_X(x)$ is the probability of the event $\{X \leq x\}$ for both continuous and discrete r.v.'s.
- Properties
  - $F_X(x)$ is monotonically nondecreasing in x.
  - $F_X(x) \to 0$ as $x \to -\infty$ and $F_X(x) \to 1$ as $x \to \infty$
  - $F_X(x)$ is continuous for continuous r.v.'s and it is piecewise constant for discrete r.v.'s
- Relation to PMF, PDF

$$\text{Discrete:} \quad p_X(k) = F_X(k) - F_X(k-1) \tag{6}$$

$$\text{Continuous:} \quad f_X(x) = \frac{dF_X}{dx}(x) \tag{7}$$

- Using CDF to compute PMF.
  - Example 3.6: Compute PMF of maximum of 3 r.v.'s: What is the PMF of the maximum score of 3 test scores, when each test score is independent of others and each score takes any value between 1 and 10 with probability 1/10?
    Answer: Compute $F_X(k) = P(X \leq k) = P(\{X_1 \leq k\}, and \{X_2 \leq k\}, and \{X_3 \leq k\}) = P(\{X_1 \leq k\})P(\{X_2 \leq k\})P\{X_3 \leq k\})$ (follows from independence of the 3 events) and then compute the PMF using (6).
  - For continuous r.v.'s, in almost all cases, the correct way to compute the CDF of a function of a continuous r.v. (or of a set of continuous r.v.'s) is to compute the CDF first and then take its derivative to get the PDF. We will learn this later.
  -
  - The PDF of a discrete r.v. X, $f_X(x) \triangleq \sum_{j=-\infty}^{\infty} p_X(j)\delta(x-j)$.
  - If I integrate this, I get $F_X(x) = \int_{t \leq x} f_X(t)dt = \sum_{j \leq x} p_X(j)$ which is the same as the CDF definition given in (4)

- Geometric and exponential CDF **
  - Let $X_{geo,p}$ be the number of trials required for the first success (geometric) with probability of success $= p$. Then we can show that the probability of $\{X_{geo,p} \leq k\}$ is equal to the probability of an exponential r.v. $\{X_{expo,\lambda} \leq k\delta\}$ with parameter λ, if δ satisfies $1 - p = e^{-\lambda\delta}$ or $\delta = -\ln(1-p)/\lambda$
    Proof: Equate $F_{X_{geo,p}}(k) = 1 - (1-p)^k$ to $F_{X_{expo,\lambda}}(k\delta) = 1 - e^{-\lambda k\delta}$
  - Implication: When δ (time interval between two Bernoulli trials (coin tosses)) is small, then $F_{X_{geo,p}}(k) \approx F_{X_{expo,\lambda}}(k\delta)$ with $p = \lambda\delta$ (follows because $e^{-\lambda\delta} \approx 1 - \lambda\delta$ for δ small).

- $Binomial(n, p)$ becomes $Poisson(np)$ for small time interval, δ, between coin tosses
  Proof idea:
  - Consider a sequence of n independent coin tosses with probability of heads p in any toss (number of heads $\sim Binomial(n, p)$).
  - Assume the time interval between two tosses is δ.
  - Then expected value of X in one toss (in time δ) is p.
  - When δ small, expected value of X per unit time is $\lambda = p/\delta$.
  - The total time duration is $\tau = n\delta$.
  - When $\delta \to 0$, but λ and τ are finite, $n \to \infty$ and $p \to 0$.
  - When δ small, can show that the PMF of a $Binomial(n, p)$ r.v. is approximately equal to the PMF of $Poisson(\lambda\tau)$ r.v. with $\lambda\tau = np$

- The Poisson process is a continuous time analog of a Bernoulli process (Details in Chap 5) **

3 Normal (Gaussian) Random Variable

- The most commonly used r.v. in Communications and Signal Processing

- X is normal or Gaussian if it has a PDF of the form

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

  where one can show that $\mu = E[X]$ and $\sigma^2 = var[X]$.

- Standard normal: Normal r.v. with $\mu = 0$, $\sigma^2 = 1$.

- Cdf of a standard normal Y, denoted $\Phi(y)$

$$\Phi(y) \triangleq P(Y \leq y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-t^2/2} dt$$

- Let X is a normal r.v. with mean μ, variance σ^2. Then can show that $Y = \frac{X-\mu}{\sigma}$ is a standard normal r.v.

- Computing CDF of any normal r.v. X using the table for Φ: $F_X(x) = \Phi(\frac{x-\mu}{\sigma})$. See Example 3.7.

4 Multiple Continuous Random Variables: Topics

- Conditioning on an event
- Joint and Marginal PDF
- Expectation, Independence, Joint CDF, Bayes rule
- Derived distributions
  - Function of a Single random variable: $Y = g(X)$ for any function g
  - Function of a Single random variable: $Y = g(X)$ for linear function g
  - Function of a Single random variable: $Y = g(X)$ for strictly monotonic g
  - Function of Two random variables: $Z = g(X, Y)$ for any function g

5 Conditioning on an event.

$$f_{X|A}(x) := \begin{cases} \frac{f_X(x)}{P(A)} & \text{if } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

Consider the special case when $A := \{X \in R\}$, e.g. the region R can be the interval $[a, b]$. In this case, we should be writing $f_{X|\{X \in R\}}$. But to keep things simple, we misuse notation to also write

$$\begin{aligned} f_{X|R}(x) &:= \begin{cases} \frac{f_X(x)}{P(X \in R)} & \text{if } x \in R \\ 0 & \text{otherwise} \end{cases} \\ &:= \begin{cases} \frac{f_X(x)}{\int_{t \in R} f_X(t) dt} & \text{if } x \in R \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

6 Joint and Marginal PDF

- Two r.v.s X and Y are **jointly continuous** iff there is a function $f_{X,Y}(x,y)$ with $f_{X,Y}(x,y) \geq 0$, called the **joint PDF**, s.t. $P((X,Y) \in B) = \int_B f_{X,Y}(x,y)dxdy$ for all subsets B of the 2D plane.

- Specifically, for $B = [a,b] \times [c,d] \triangleq \{(x,y) : a \leq x \leq b, c \leq y \leq d\}$,

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_{y=c}^{d} \int_{x=a}^{b} f_{X,Y}(x,y)dxdy$$

- **Interpreting the joint PDF:** For small positive numbers δ_1, δ_2,

$$P(a \leq X \leq a+\delta_1, c \leq Y \leq c+\delta_2) = \int_{y=c}^{c+\delta_2} \int_{x=a}^{a+\delta_1} f_{X,Y}(x,y)dxdy \approx f_{X,Y}(a,c)\delta_1\delta_2$$

  Thus $f_{X,Y}(a,c)$ is the probability mass per unit area near (a,c).

- **Marginal PDF:** The PDF obtained by integrating the joint PDF over the entire range of one r.v. (in general, integrating over a set of r.v.'s)

$$\begin{aligned} P(a \leq X \leq b) &= P(a \leq X \leq b, -\infty \leq Y \leq \infty) = \int_{x=a}^{b} \int_{y=-\infty}^{\infty} f_{X,Y}(x,y)dydx \\ \implies f_X(x) &= \int_{y=-\infty}^{\infty} f_{X,Y}(x,y)dy \end{aligned}$$

- Example 3.12, 3.13

7 Conditional PDF

- Conditional PDF of X given that $Y = y$ is defined as

$$f_{X|Y}(x|y) \triangleq \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

- For any y, $f_{X|Y}(x|y)$ is a legitimate PDF: integrates to 1.

- Example 3.15

- **Interpretation:** For small positive numbers δ_1, δ_2, consider the probability that X belongs to a small interval $[x, x+\delta_1]$ given that Y belongs to a small interval $[y, y+\delta_2]$

$$\begin{aligned} P(x \leq X \leq x+\delta_1 | y \leq Y \leq y+\delta_2) &= \frac{P(x \leq X \leq x+\delta_1, y \leq Y \leq y+\delta_2)}{P(y \leq Y \leq y+\delta_2)} \\ &\approx \frac{f_{X,Y}(x,y)\delta_1\delta_2}{f_Y(y)\delta_2} \\ &= f_{X|Y}(x|y)\delta_1 \end{aligned}$$

- **Since $f_{X|Y}(x|y)\delta_1$ does not depend on δ_2, we can think of the limiting case when $\delta_2 \to 0$ and so we get**

$$P(x \leq X \leq x+\delta_1 | Y = y) = \lim_{\delta_2 \to 0} P(x \leq X \leq x+\delta_1 | y \leq Y \leq y+\delta_2) \approx f_{X|Y}(x|y)\delta_1 \quad \delta_1 \text{ small}$$

- In general, for any region A, we have that

$$P(X \in A|Y = y) = \lim_{\delta \to 0} P(X \in A|y \leq Y \leq y + \delta) = \int_{x \in A} f_{X|Y}(x|y)dx$$

8 Expectation, Independence, Joint & Conditional CDF, Bayes

- **Expectation:** See page 172 for $E[g(X)|Y = y]$, $E[g(X,Y)|Y = y]$ and total expectation theorem for $E[g(X)]$ and for $E[g(X,Y)]$.
- **Independence:** X and Y are independent iff $f_{X|Y} = f_X$ (or iff $f_{X,Y} = f_X f_Y$, or iff $f_{Y|X} = f_Y$)
- If X and Y independent, any two events $\{X \in A\}$ and $\{Y \in B\}$ are independent.
- If X and Y independent, $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$ and $Var[X+Y] = Var[X]+Var[Y]$ Exercise: show this.
- **Joint CDF:**

$$F_{X,Y}(x,y) := P(X \leq x, Y \leq y) = \int_{t=-\infty}^{y} \int_{s=-\infty}^{x} f_{X,Y}(s,t)dsdt$$

- Obtain joint PDF from joint CDF:

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x,y)$$

- **Conditional CDF:**

$$F_{X|Y}(x|y) := P(X \leq x|Y = y) = \lim_{\delta \to 0} P(X \leq x|y \leq Y \leq y + \delta) = \int_{t=-\infty}^{x} f_{X|Y}(t|y)dt$$

- **Bayes rule when unobserved phenomenon is continuous.** Pg 175 and Example 3.18. Recall that $f_{X|Y}(x|y)$ is, by definition, such that, for δ small,

$$P(X \in [x, x+\delta]|Y = y) = f_{X|Y}(x|y)\delta$$

Also, for δ, δ_2 small,

$$P(X \in [x, x+\delta], Y \in [y, y+\delta_2]) = f_{X,Y}(x,y)\delta\delta_2$$

Using Bayes rule for events,

$$P(X \in [x, x+\delta]|Y \in [y, y+\delta_2]) = \frac{P(X \in [x, x+\delta], Y \in [y, y+\delta_2])}{P(Y \in [y, y+\delta_2])} = \frac{f_{X,Y}(x,y)\delta\delta_2}{f_Y(y)\delta_2} = \frac{f_{X,Y}(x,y)\delta}{f_Y(y)}$$

Notice that the right hand side does not depend on δ_2. Taking the limit $\delta_2 \to 0$, we get

$$P(X \in [x, x+\delta]|Y = y) = \lim_{\delta_2 \to 0} P(X \in [x, x+\delta]|Y \in [y, y+\delta_2]) = \frac{f_{X,Y}(x,y)\delta}{f_Y(y)}$$

Thus,

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

- **Bayes rule when unobserved phenomenon is discrete.** Pg 176 and Example 3.19. For e.g., say discrete r.v. N is the unobserved phenomenon. Then for δ small,

$$\begin{aligned} P(N=i|X\in[x,x+\delta]) &= P(N=i|X\in[x,x+\delta]) \\ &= \frac{P(N=i)P(X\in[x,x+\delta]|N=i)}{P(X\in[x,x+\delta])} \\ &= \frac{p_N(i)f_{X|N}(x|i)\delta}{\sum_j p_N(j)f_{X|N}(x|j)\delta} \\ &= \frac{p_N(i)f_{X|N}(x|i)}{\sum_j p_N(j)f_{X|N}(x|j)} \end{aligned}$$

Notice that the right hand side is independent of δ. Thus we can take $\lim_{\delta\to 0}$ on both sides and the right side will not change. Thus we get

$$p_{N|X}(i|x) = P(N=i|X=x) = \lim_{\delta\to 0} P(N=i|X\in[x,x+\delta]) = \frac{p_N(i)f_{X|N=i}(x)}{\sum_j p_N(j)f_{X|N=j}(x)}$$

- **Bayes rule with conditioning on events.** The derivation is analogous to the above conditioning on discrete r.v.'s case.
Suppose that events $A_1, A_2, \ldots A_n$ form *a partition*, i.e. they are disjoint and their union is the entire sample space. The simplest example is $n=2$, $A_1 = A$, $A_2 = A^c$.
Then

$$P(A_i|X=x) = \frac{P(A_i)f_{X|A_i}(x)}{\sum_j P(A_j)f_{X|A_j}(x)}$$

- More than 2 random variables (Pg 178, 179) **

9 Derived distributions: PDF of $g(X)$ and of $g(X,Y)$

- **Obtaining PDF of $Y = g(X)$.** ALWAYS use the following 2 step procedure:
  - Compute CDF first. $F_Y(y) = P(g(X)\le y) = \int_{x|g(x)\le y} f_X(x)dx$
  - Obtain PDF by differentiating F_Y, i.e. $f_Y(y) = \frac{\partial F_Y}{\partial y}(y)$
- **Special Case 1: Linear Case: $Y = aX + b$.** Can show that

$$f_Y(y) = \frac{1}{|a|} f_X(\frac{y-b}{a})$$

- **Special Case 2: Strictly Monotonic Case.**
  - Consider $Y = g(X)$ with g being a **strictly monotonic** function of X.
  - Thus g is a one to one function.

- Thus there exists a function h s.t. $y = g(x)$ iff $x = h(y)$ (i.e. h is the inverse function of g, often denotes as $h \triangleq g^{-1}$).
- Then can show that

$$f_Y(y) = f_X(h(y))|\frac{dh}{dy}(y)|$$

- Proof for strictly monotonically increasing g:
  $F_Y(y) = P(g(X) \leq Y) = P(X \leq h(Y)) = F_X(h(y))$.
  Differentiate both sides w.r.t y (apply chain rule on the right side) to get:

$$f_Y(y) = \frac{dF_Y}{dy}(y) = \frac{dF_X(h(y))}{dy} = f_X(h(y))\frac{dh}{dy}(y)$$

  For strictly monotonically decreasing g, using a similar procedure, we get $f_Y(y) = -f_X(h(y))\frac{dh}{dy}(y)$

- **Functions of two random variables.** Two possible ways to solve this depending on which is easier. Try the first method first: if easy to find the region to integrate over then just do that. Else use the second method.

  1. Do the following
     (a) Compute CDF of $Z = g(X, Y)$, i.e compute $F_Z(z)$. In general, this computed as:

     $F_Z(z) = P(g(X, Y) \leq z) = \int_{x,y:g(x,y)\leq z} f_{X,Y}(x, y)dydx$.
     (b) Differentiate w.r.t. z to get the PDF, i.e. compute $f_Z(z) = \frac{\partial F_Z(z)}{\partial z}$.
  2. Use a three step procedure
     (a) Compute conditional CDF, $F_{Z|X}(z|x) := P(Z \leq z|X = x)$
     (b) Differentiate w.r.t. z to get conditional PDF, $f_{Z|X}(z|x) = \frac{\partial F_{Z|X}(z|x)}{\partial z}$
     (c) Compute $f_Z(z) = \int f_{Z,X}(z, x)dx = \int f_{Z|X}(z|x)f_X(x)dx$

- Special case: PDF of $Z = X + Y$ when X, Y are independent: convolution of PDFs of X and Y.

Outline

Chapter 1 - Some Topics On Probability

- Chain rule: $P(A_1, A_2, A_3, ..., A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \dots P(A_n|A_1, A_2, \dots, A_{n-1})$
- Total Probability: if $B_1, B_2, \dots, B_n$ from a partition of the sample space, then $P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$
  Partition: The events are mutually disjoint and their union is equal to the sample space.
- Union bound: suppose $P(A_i) \geq 1 - p_i$ for small probabilities p_i, then
  $P(\cap_i A_i) = 1 - P(\cup_i A_i^c) \geq 1 - \sum_i P(A_i^c) \geq 1 - \sum_i p_i$

Chapter 1 - Some Topics On Probability

- Independence: Events A,B are independent if $P(A, B) = P(A)P(B)$
- events $A_1, A_2, A_3, ...A_n$ are mutually independent if for any subset $S \subseteq \{1, 2, 3, ...n\}$, $P(\cap_{i \in s} A_i) = \Pi_{i \in s} P(A_i)$
- analogous definition for random variables: for mutually independent r.v.'s the joint pdf of any subset of r.v.'s is equal to the product of the marginal pdf's.

Chapter 1 - Some Topics On Probability

- Conditional Independence: Events A,B are conditionally independent given an event C if $P\big(A, B|C\big) = P(A|C)P(B|C)$ extend to a set of events as above extend to r.v.'s as above
  Given X is independent of $\{Y, Z\}$. Then,
  - X is independent of Y; X is independent of Z;
  - X is conditionally independent of Y given Z
  - $E[XY/Z] = E[X|Z]E[Y/Z]$
  - $E[XY/Z] = E[X]E[Y/Z]$

Chapter 1 - Some Topics On Probability

- Law of Iterated Expectations: $E_{X,Y}[g(X,Y)] = E_Y[E_{X|Y}[g(X,Y)|Y]]$
- Conditional Variance Identity: $Var_{X,Y}[g(X,Y)] = E_Y[Var_{X|Y}[g(X,Y)|Y]] + Var_Y[E_{X|Y}[g(X,Y)|Y]]$
- Cauchy-Schwartz Inequality: For vectors v_1, v_2, $| < v_1, v_2 > |^2 \leq \|v_1\|_2^2 \|v_2\|_2^2$

Chapter 1 - Some Topics On Probability

- For scalar r.v.'s X, Y: $|E[XY]|^2 \leq E[X^2]E[Y^2]$
- For random vectors X, Y, $|E[X'Y]|^2 \leq E[\|X\|_2^2]E[\|Y\|_2^2]$
- Proof follows by using the fact that $E[(X - \alpha Y)^2] \geq 0$. Get a quadratic equation in α and use the condition to ensure that this is non-negative
- For random matrices $\mathcal{X}, \mathcal{Y}$, $\|E[\mathcal{X}\mathcal{Y}']\|_2^2 \leq \lambda_{\max}(E[\mathcal{X}\mathcal{X}'])\lambda_{\max}(E[\mathcal{Y}\mathcal{Y}']) = \|E[\mathcal{X}\mathcal{X}']\|_2 \|E[\mathcal{Y}\mathcal{Y}']\|_2$. Recall that for a positive semi-definite matrix M, $\|M\|_2 = \lambda_{\max}(M)$.

- Proof: use the following definition of $\|M\|_2$: $\|M\|_2 = \max_{x,y:\|x\|_2=1,\|y\|_2=1} |x'My|$, and then apply C-S for random vectors.
- Convergence in probability. A sequence of random variables, $X_1, X_2, \dots X_n$ converges to a constant a in probability means that for every $\epsilon > 0$, $\lim_{n\to\infty} Pr(|X_n - a| > \epsilon) = 0$
- Convergence in distribution. A sequence of random variables, $X_1, X_2, \dots X_n$ converges to random variable Z in distribution means that $\lim_{n\to\infty} F_{X_n}(x) = F_Z(x), \text{for almost all points}\, x.$

Chapter 1 - Some Topics On Probability

- Convergence in probability implies convergence in distribution
- Consistent Estimator. An estimator for θ based on n random variables, $\hat{\theta}_n(\underline{X})$, is consistent if it converges to θ in probability for large n.
- independent and identically distributed (iid) random variables: $X_1, X_2, \dots X_n$ are iid iff they are mutually independent and have the same marginal distribution.
- For all subsets $\subseteq \{1, 2, \dots n\}$ of size s, the following two things hold
  $F_{X_i, i \in S}(x_1, x_2, \dots x_s) = \prod_{i \in S} F_{X_i}(x_i)$ (independent) and
  $F_{X_i}(x_i) = F_{X_1}(x_1)$ (iid)
- Clearly the above two imply that the joint distribution for any subset of variables is also equal $F_{X_i, i \in S}(x_1, x_2, \dots x_s) = \prod_{i=1}^{s} F_{X_1}(x_i) = F_{X_1, X_2, \dots X_s}(x_1, x_2, \dots x_s)$.

Chapter 1 - Some Topics On Probability

- Moment Generating Function (MGF) $M_X(u)$
  $M_X(u) := E[e^{u^T X}]$
- It is the two-sided Laplace transform of the pdf of X for continuous r.v.'s X.
- For a scalar r.v. X, $M_X(t) := E[e^{tX}]$, differentiating this i times with respect to t and setting $t = 0$ gives the i-th moment about the origin.
- Characteristic Function $C_X(u) := M_X(iu) = E[e^{iu^T X}]$
- $C_X(-u)$ is the Fourier transform of the pdf or pmf of X : defined only when the Fourier transform exists.
- Can get back the pmf or pdf by inverse Fourier transform.

Chapter 1 - Some Topics On Probability

- Markov inequality and its implications
- Markov inequality: for a non-negative r.v. i.e. for X for which $P(X < 0) = 0$
  $P(X > a) \leq \frac{E[X]}{a}$.
- Chebyshev inequality: apply Markov to $(Y - \mu_Y)^2$
  $P((Y - \mu_Y)^2 > a) \leq \frac{\sigma_Y^2}{a}$
  if the variance is small, Y does not deviate too much from its mean.
- Chernoff bounds: apply Markov to e^{tY} for any $t > 0$.
  $P(X > a) \leq \min_{t>0} e^{-ta} E[e^{tX}]$
  $P(X < b) \leq \min_{t>0} e^{tb} E[e^{-tX}]$ or sometimes one gets a simpler expression by using a specific value of $t > 0$

- Using Chernoff bounding to bound $P(S_n \in [a,b])$, $S_n := \sum_{i=1}^n X_i$ when X_i's are iid $P(S_n \geq a) \leq \min_{t>0} e^{-ta} \prod_{i=1}^n E[e^{tX_i}] = \min_{t>0} e^{-ta}(E[e^{tX_1}])^n := p_1$. $P(S_n \leq b) \leq \min_{t>0} e^{tb} \prod_{i=1}^n E[e^{-tX_i}] = \min_{t>0} e^{tb}(E[e^{-tX_1}])^n := p_2$
- Thus, using the union bound with $A_1 = S_n < a$, $A_2 = S_n > b$ $P(b < S_n < a) \geq 1 - p_1 - p_2$ With $b = n(\mu - \epsilon)$ and $a = n(\mu + \epsilon)$, we can conclude that w.h.p. $\bar{X}_n := S_n/n$ lies between $\mu \pm \epsilon$.
- A similar thing can also be done when X_i's just independent and not iid. Sometimes have an upper bound for $E[e^{tX_1}]$ and that can be used.

Hoeffding inequality: Chernoff bound for sums of independent bounded random variables, followed by using Hoeffding's lemma.

- Given independent and *bounded* r.v.'s $X_1, \dots X_n$: $P(X_i \in [a_i, b_i]) = 1$, $P(|S_n - E[S_n]| \geq t) \leq 2\exp(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}) = P(E[S_n] - t \leq S_n \leq E[S_n] + t)$ or let $\bar{X}_n := S_n/n$ and $\mu_n := \sum_i E[X_i]/n$, then $P(|\bar{X}_n - \mu_n| \geq \epsilon) \leq 2\exp(\frac{-2\epsilon^2 n^2}{\sum_{i=1}^n (b_i - a_i)^2}) \leq 2\exp(\frac{-2\epsilon^2 n^2}{\max_i (b_i - a_i)^2})$
- Proof: use Chernoff bounding followed by Hoeffding's lemma

Chapter 1 - Some Topics On Probability

- Weak Law of Large Numbers (WLLN) for i.i.d. scalar random variables, $X_1, X_2, \ldots X_n$, with finite mean μ. Define $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ For any $\epsilon > 0$, $\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$
  Proof: use Chebyshev if σ^2 is finite. Else use characteristic function
- Central Limit Theorem for i.i.d. random variables. Given an iid sequence of random variables, $X_1, X_2, \ldots X_n$, with finite mean μ and finite variance σ^2 as the sample mean. Then $\sqrt{n}(\bar{X}_n - \mu)$ converges in distribution a Gaussian rv $Z \sim (0, \sigma^2)$
- Many of the above results also exist for certain types of non-iid rv's. Proofs much more difficult.
- Mean Value Theorem and Taylor Series Expansion

Chapter 1 - Some Topics On Probability

- Delta method: if $\sqrt{N}(X_N - \theta)$ converges in distribution to Z then $\sqrt{N}(g(X_N) - g(\theta))$ converges in distribution to $g'(\theta)Z$ as long as $g'(\theta)$ is well defined and non-zero. Thus if $Z \sim (0, \sigma^2)$, then $g'(\theta)Z \sim (0, g'(\theta)^2\sigma^2)$.
- If $g'(\theta) = 0$, then one can use what is called the second-order Delta method. This is derived by using a second order Taylor series expansion or second-order mean value theorem to expand out $g(X_N)$ around θ.
- Second order Delta method: Given that $\sqrt{N}(X_N - \theta)$ converges in distribution to Z. Then, if $g'(\theta) = 0$, $N(g(X_N) - g(\theta))$ converges in distribution to $\frac{g''(\theta)}{2}Z^2$. If $Z \sim (0, \sigma^2)$, then $Z^2 = \sigma^2\frac{g''(\theta)}{2}\chi_1^2$ where χ_1^2 is a r.v. that has a chi-square distribution with 1 degree of freedom.
- Slutsky's theorem

Chapter 2 - Jointly Gaussian Random Variables

- First note that a scalar Guassian r.v. X with mean μ and variance σ^2 has the following pdf $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ Its characteristic function can be computed by computing the Fourier transform at $-t$ to get $C_X(t) = e^{j\mu t} e^{-\frac{\sigma^2 t^2}{2}}$
  Next we prove a sequence of if and only if conditions for a random vector X to be jointly Gaussian. Any one of these could serve as the definition of j G. A random vector X is jointly Gaussian if and only if one of the following (and hence all of the following) holds
- The $n \times 1$ random vector X is jointly Gaussian if and only if the scalar $u^T X$ is Gaussian distributed for all $n \times 1$ vectors

Chapter 2 - Jointly Gaussian Random Variables

- The random vector X is jointly Gaussian if and only if its characteristic function, $C_X(u) := E[e^{iu^T X}]$ can be written as $C_X(u) = e^{iu^T\mu}e^{-u^T\Sigma u/2}$ where $\mu = E[X]$ and $\Sigma = cov(X)$.
- Proof: X is j G implies that $V = u^T X$ is G with mean $u^T\mu$ and variance $u^T\Sigma u$. Thus its characteristic function, $C_V(t) = e^{itu^T\mu}e^{-t^2u^T\Sigma u/2}$. But $C_V(t) = E[e^{itV}] = E[e^{itu^T X}]$. If we set $t = 1$, then this is $E[e^{iu^T X}]$ which is equal to $C_X(u)$. Thus, $C_X(u) = C_V(1) = e^{iu^T\mu}e^{-u^T\Sigma u/2}$.
- Proof (other side): we are given that the charac function of X, $C_X(u) = E[e^{iu^T X}] = e^{iu^T\mu}e^{-u^T\Sigma u/2}$. Consider $V = u^T X$. Thus, $C_V(t) = E[e^{itV}] = C_X(tu) = e^{iu^T\mu}e^{-t^2u^T\Sigma u/2}$. Also, $E[V] = u^T\mu$, $var(V) = u^T\Sigma u$. Thus V is G.

Chapter 2 - Jointly Gaussian Random Variables

- The random vector X is jointly Gaussian if and only if its joint pdf can be written as $f_X(x) = \frac{1}{(\sqrt{2\pi})^n det(\Sigma)} e^{-(X-\mu)^T \Sigma^{-1}(X-\mu)/2}$
- Proof: follows by computing the characteristic function from the pdf and vice versa.
- The random vector X is j G if and only if it can be written as an affine function of i.i.d. standard Gaussian r.v's.
- Proof uses 2.
- Proof: suppose $X = AZ + a$ where $Z \sim (0, I)$; compute its c.f. and show that it is a c.f. of a j G, joint pdf given by (18) and thus it is j G.
- Proof (other side): suppose X is j G; let $Z := \Sigma^{-1/2}(X - \mu)$ and write out its c.f.; can show that it is the c.f. of iid standard G.

Chapter 2 - Jointly Gaussian Random Variables

- The random vector X is j G if and only if it can be written as an affine function of jointly Gaussian r.v's.
- Proof: Suppose X is an affine function of a j G r.v. Y, i.e. $X = BY + b$. Since Y is j G, by 18, it can be written as $Y = AZ + a$ where $Z \sim (0, I)$ (i.i.d. standard Gaussian). Thus, $X = BAZ + (Ba + b)$, i.e. it is an affine function of Z, and thus, by 18, X is j G.
- Proof (other side): X is j G. So by 18, it can be written as $X = BZ + b$. But $Z \sim (0, I)$ i.e. Z is a j G r.v.

Properties:

- If X_1, X_2 are j G, then the conditional distribution of X_1 given X_2 is also j G.
- If the elements of a j G r.v. X are pairwise uncorrelated (i.e. non-diagonal elements of their covariance matrix are zero), then they are also mutually independent.
- Any subset of X is also j G.

Chapter-3 Optimization: basic fact

- Claim: $\min_{t_1,t_2} f(t_1,t_2) = \min_{t_1}(\min_{t_2} f(t_1,t_2))$
  Proof: show that LHS $\geq$ RHS and LHS $\leq$ RHS
  Let $[\hat{\hat{t}}_1, \hat{\hat{t}}_2] \in \arg\min_{t_1,t_2} f(t_1,t_2)$ (if the minimizer is not unique let $\hat{\hat{t}}_1, \hat{\hat{t}}_2$ be any one minimizer), i.e.
  $\min_{t_1,t_2} f(t_1,t_2) = f(\hat{\hat{t}}_1, \hat{\hat{t}}_2)$
  Let $\hat{t}_2(t_1) \in \arg\min_{t_2} f(t_1,t_2)$, i.e.
  $\min_{t_2} f(t_1,t_2) = f(t_1, \hat{t}_2(t_1))$
  Let $\hat{t}_1 \in \arg\min f(t_1, \hat{t}_2(t_1))$, i.e.
  $\min_{t_1} f(t_1, \hat{t}_2(t_1)) = f(\hat{t}_1, \hat{t}_2(\hat{t}_1))$
  Combining last two equations,
  $f(\hat{t}_1, \hat{t}_2(\hat{t}_1)) = \min_{t_1} f(t_1, \hat{t}_2(t_1)) = \min_{t_1}(\min_{t_2} f(t_1,t_2))$

- Notice that
  $f(t_1, t_2) \min_{t_2} f(t_1, t_2) f(t_1, \hat{t}_2(t_1)) \min_{t_1} f(t_1, \hat{t}_2(t_1)) f(\hat{t}_1, \hat{t}_2(\hat{t}_1))$
- The above holds for all t_1, t_2. In particular use $t_1 \equiv \hat{\hat{t}}_1$, $t_2 \equiv \hat{\hat{t}}_2$. Thus,
  $\min_{t_1,t_2} f(t_1, t_2) = f(\hat{\hat{t}}_1, \hat{\hat{t}}_2) \geq \min_{t_1} f(t_1, \hat{t}_2(t_1)) = \min_{t_1}(\min_{t_2} f(t_1, t_2))$
  Thus LHS $\geq$ RHS. Notice also that
  $\min_{t_1,t_2} f(t_1, t_2) f(t_1, t_2) = f(\hat{\hat{t}}_1, \hat{\hat{t}}_2)$ and this holds for all t_1, t_2. In particular, use $t_1 \equiv \hat{t}_1$, $t_2 \equiv \hat{t}_2(\hat{t}_1)$. Then,
  $\min_{t_1,t_2} f(t_1, t_2) f(\hat{t}_1, \hat{t}_2(\hat{t}_1)) = \min_{t_1}(\min_{t_2} f(t_1, t_2))$
  Thus, LHS $\leq$ RHS and this finishes the proof.