Stochastic Processes 1

STAT 333

Fall 2021 (1219)

Cameron Roopnarine[1] Steve Drekic[2] Mirabelle Huynh[3]

4th December 2021

[1] LaTeXer
[2] Main Instructor
[3] Assisting Instructor

Contents

Chapter 1

Review of Elementary Probability

WEEK 1
8th to 15th September

Fundamental Definition of a Probability Function

Probability Model: A probability model consists of 3 essential components: a *sample space*, a collection of *events*, and a *probability function (measure)*.

- **Sample Space**: For a random experiment in which all possible outcomes are known, the set of all possible outcomes is called the sample space (denoted by Ω).
- **Event**: Every subset A of a sample space Ω is an event.
- **Probability Function**: For each event A of Ω, $\mathbb{P}(A)$ is defined as the *probability of an event A*, satisfying 3 conditions:
  (i) $0 \le \mathbb{P}(A) \le 1$,
  (ii) $\mathbb{P}(\Omega) = 1$, or equivalently, $\mathbb{P}(\emptyset) = 0$, where $\emptyset$ is the *null event*,
  (iii) For $n \in \mathbb{Z}^+$ (in fact, $n = \infty$ as well), $\mathbb{P}(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n \mathbb{P}(A_i)$ if the sequence of events $\{A_i\}_{i=1}^n$ is *mutually exclusive* (i.e., $A_i \cap A_j = \emptyset \; \forall i \neq j$).

As a result of conditions (ii) and (iii), and noting that A^c is the complement of A, it follows that

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c) \implies \mathbb{P}(A^c) = 1 - \mathbb{P}(A).$$

Conditional Probability

Conditional Probability: The *conditional probability of event A given event B occurs* is defined as

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)},$$

provided that $\mathbb{P}(B) > 0$.

Remarks:

(1) When $B = \Omega$, $\mathbb{P}(A \mid \Omega) = \mathbb{P}(A \cap \Omega)/\mathbb{P}(\Omega) = \mathbb{P}(A)/1 = \mathbb{P}(A)$, as one would expect.

(2) Rewriting the above formula, $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\,\mathbb{P}(B)$, which is often referred to as the basic "multiplication rule." For a sequence of events $\{A_i\}_{i=1}^n$, the generalized multiplication rule is given by

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1) \cdots \mathbb{P}(A_n \mid A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

Example 1.1. Suppose that we roll a fair six-sided die once (i.e., $\Omega = \{1, 2, 3, 4, 5, 6\}$). Let A denote the event of rolling a number less than 4 (i.e., $A = \{1, 2, 3\}$), and let B denote the event of rolling an odd number (i.e., $B = \{1, 3, 5\}$). Given that the roll is odd, what is the probability that number rolled is less than 4?

Solution: Since the die is fair, it immediately follows that $\mathbb{P}(A) = 3/6 = 1/2$ and $\mathbb{P}(B) = 3/6 = 1/2$. Moreover,

$$\begin{aligned}
\mathbb{P}(A \cap B) &= \mathbb{P}\big(\{1, 2, 3\} \cap \{1, 3, 5\}\big) && (1.1)\\
&= \mathbb{P}\big(\{1, 3\}\big) && (1.2)\\
&= \frac{2}{6} && (1.3)\\
&= \frac{1}{3}. && (1.4)
\end{aligned}$$

Therefore,

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1/3}{1/2} = \frac{2}{3}.$$

Independence of Events

Independence of Events: Two events A and B are *independent* if and only if (iff)

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$$

In general, if an experiment consists of a sequence of independent trials, and $A_1, A_2, \ldots, A_n$ are events such that A_i depends only on the i^{th} trial, then $A_1, A_2, \ldots, A_n$ are independent events and

$$\mathbb{P}(\cap_{i=1}^n A_i) = \prod_{i=1}^n \mathbb{P}(A_i).$$

Law of Total Probability

Law of Total Probability: For $n \in \mathbb{Z}^+$ (and even $n = \infty$), suppose that $\Omega = \cup_{i=1}^n B_i$, where the sequence of events $\{B_i\}_{i=1}^n$ is mutually exclusive. Then,

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A \cap \Omega)\\ &= \mathbb{P}\big(A \cap \{\cup_{i=1}^n B_i\}\big)\\ &= \mathbb{P}\big(\cup_{i=1}^n \{A \cap B_i\}\big)\\ &= \sum_{i=1}^n \mathbb{P}(A \cap B_i)\\ &= \sum_{i=1}^n \mathbb{P}(A \mid B_i)\,\mathbb{P}(B_i),\end{aligned}$$

where the second last equality follows from the fact that the sequence of events $\{A \cap B_i\}_{i=1}^n$ is also mutually exclusive.

Bayes' Formula

Bayes' Formula: Under the same assumptions as in the previous slide,

$$\mathbb{P}(B_j \mid A) = \frac{\mathbb{P}(A \cap B_j)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \mid B_j)\,\mathbb{P}(B_j)}{\sum_{i=1}^n \mathbb{P}(A \mid B_i)\,\mathbb{P}(B_i)}.$$

Definition of a Random Variable

Definition: A *random variable* (rv) X is a real-valued function which maps a sample space Ω onto a state space $\mathcal{S} \subseteq \mathbb{R}$ (i.e., $X : \Omega \to \mathcal{S}$).

Discrete type: $\mathcal{S}$ consists of a finite or countable number of possible values. Important functions include:

$$\begin{aligned}p(a) &= \mathbb{P}(X = a) && \text{(pmf)},\\ F(a) &= \mathbb{P}(X \le a) = \sum_{x \le a} p(x) && \text{(cdf)},\\ \bar{F}(a) &= \mathbb{P}(X > a) = 1 - F(a) && \text{(tpf)},\end{aligned}$$

where pmf stands for *probability mass function*, cdf stands for *cumulative distribution function*, and tpf stands for *tail probability function*.

Remark: If X takes on values in the set $\mathcal{S} = \{a_1, a_2, a_3, \ldots\}$ where $a_1 < a_2 < a_3 < \cdots$ such that $p(a_i) > 0$ $\forall i$, then we can recover the pmf from knowledge of the cdf via

$$\begin{aligned}p(a_1) &= F(a_1),\\ p(a_i) &= F(a_i) - F(a_{i-1}),\ i = 2, 3, 4, \ldots\end{aligned}$$

Discrete Distributions

Special Discrete Distributions:

1. **Bernoulli**: If we consider a *Bernoulli trial*, which is a random trial with probability p of being a "success" (denoted by 1) and a probability $1-p$ of being a "failure" (denoted by 0), then X is *Bernoulli* (i.e., $X \sim \text{BERN}(p)$) with pmf

$$p(x) = p^x(1-p)^{1-x},\ x = 0, 1.$$

2. **Binomial**: If X denotes the number of successes in $n \in \mathbb{Z}^+$ independent Bernoulli trials, each with probability p of being a success, then X is Binomial (i.e., $X \sim \text{BIN}(n,p)$) with pmf

$$p(x) = \binom{n}{x} p^x(1-p)^{n-x},\ x = 0, 1, \ldots, n,$$

where

$$\binom{n}{x} = \frac{n!}{(n-x)!x!} = \frac{(n)_x}{x!} = \frac{n(n-1)\cdots(n-x+1)}{x!}$$

is the number of distinct groups of x objects chosen from a set of n objects.

Remarks:

(1) A $\text{BIN}(1,p)$ distribution simplifies to become the $\text{BERN}(p)$ distribution.

(2) The binomial pmf is even defined for $n = 0$, in which case $p(x) = 1$ for $x = 0$. Such a distribution is said to be degenerate at 0.

(3) Note that $\binom{n}{x} = 0$ if $n, x \in \mathbb{N}$ with $n < x$.

3. **Negative Binomial**: If X denotes the number of Bernoulli trials (each with success probability p) required to observe $k \in \mathbb{Z}^+$ successes, then X is *Negative Binomial* (i.e., $X \sim \text{NB}_t(k,p)$) with pmf

$$p(x) = \binom{x-1}{k-1} p^k(1-p)^{x-k},\ x = k, k+1, k+2, \ldots.$$

Remarks:

(1) In the above pmf, $\binom{x-1}{k-1}$ appears rather than $\binom{x}{k}$ since the final trial must always be a success.

(2) Sometimes, a negative binomial distribution is alternatively defined as the number of failures observed to achieve k successes. If Y denotes such a rv and $X \sim \text{NB}_t(k,p)$, then we clearly have the relationship $X = Y + k$, which immediately leads to the following pmf for Y:

$$p_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}(X = y + k) = \binom{y+k-1}{k-1} p^k(1-p)^y,\ y = 0, 1, 2, \ldots.$$

To refer to this negative binomial distribution, we will write $Y \sim \text{NB}_f(k,p)$.

4. **Geometric**: If $X \sim \text{NB}_t(1,p)$, then X is *Geometric* (i.e., $X \sim \text{GEO}_t(p)$) with pmf

$$p(x) = p(1-p)^{x-1},\ x = 1, 2, 3 \ldots.$$

In other words, the geometric distribution models the number of Bernoulli trials required to observe the first success.

Remark: Similarly, if $X \sim \text{NB}_f(1, p)$ then we obtain an alternative geometric distribution (denoted by $X \sim \text{GEO}_f(p)$) which models the number of failures observed prior to the first success.

5. **Discrete Uniform**: If X is equally likely to take on values in the (finite) set $\{a, a+1, \ldots, b\}$ where $a, b \in \mathbb{Z}$ with $a \leq b$, then X is *Discrete Uniform* (i.e., $X \sim \text{DU}(a, b)$) with pmf

$$p(x) = \frac{1}{b-a+1}, \ x = a, a+1, \ldots, b.$$

6. **Hypergeometric**: If X denotes the number of success objects in n draws without replacement from a finite population of size N containing exactly r success objects, then X is *Hypergeometric* (i.e., $X \sim \text{HG}(N, r, n)$) with pmf

$$p(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}, \ x = \max\{0, n-N+r\}, \ldots, \min\{n, r\}.$$

7. **Poisson**: A rv X is *Poisson* (i.e., $X \sim \text{POI}(\lambda)$) with parameter $\lambda > 0$ if its pmf is one of the form

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \ x = 0, 1, 2, \ldots.$$

Remark: The pmf is even defined for $\lambda = 0$ (if we use the standard convention that $0^0 = 1$), in which case $p(x) = 1$ for $x = 0$ (i.e., X is degenerate at 0).

Example 1.2. Show that when n is large and p is small, the $\text{BIN}(n, p)$ distribution may be approximated by a $\text{POI}(\lambda)$ distribution where $\lambda = np$.

Solution: Recall $e^z = \lim\limits_{n\to\infty}(1 + z/n)^n$, $z \in \mathbb{R}$. Letting $X \sim \text{BIN}(n, p)$, we have

$$\begin{aligned}
\mathbb{P}(X = x) &= \binom{n}{x}p^x(1-p)^{n-x} \\
&= \frac{n(n-1)\cdots(n-x+1)}{x!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \\
&= \frac{n}{n}\frac{n-1}{n}\cdots\frac{n-x+1}{n}\frac{\lambda^x}{x!}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^x} \\
&\simeq (1)(1)\cdots(1)\frac{\lambda^x}{x!}\frac{e^{-\lambda}}{1} && \text{when } n \text{ is large} \\
&= \frac{e^{-\lambda}\lambda^x}{x!}
\end{aligned}$$

Continuous Random Variables

Continuous type: A rv X takes on a continuum of possible values (which is uncountable) with cdf

$$F(x) = \mathbb{P}(X \le x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y,$$

where $f(x)$ denotes the *probability density function* (pdf) of X, which is a non-negative real-valued function that satisfies

$$\mathbb{P}(X \in B) = \int_{x \in B} f(x)\,\mathrm{d}x,$$

where B is the set of real numbers (e.g., an interval).

Remarks:

(1) If $F(x)$ (or the tpf $\bar{F}(x) = 1 - F(x)$) is known, we can recover the pdf using the relation

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x} F(x) = F'(x) = -\bar{F}'(x),$$

which holds by the *Fundamental Theorem of Calculus*.

(2) When working with pdfs in general, it is usually not necessary to be precise about specifying whether a range of numbers includes the endpoints. This is quite different from the situation we encounter with discrete rvs. Throughout this course, however, we will adopt the convention of **not including** the endpoints when specifying the range of values for pdfs.

Continuous Distributions

Special Continuous Distributions:

1. **Uniform**: A rv X is *Uniform* on the real interval (a, b) (i.e., $X \sim \mathrm{U}(a, b)$) if it has pdf

$$f(x) = \frac{1}{b-a},\ a < x < b,$$

where $a, b \in \mathbb{R}$ with $a < b$.

Remark: The choice of name is because X takes on values in (a, b) with all subintervals of a fixed length being equally likely.

2. **Beta**: A rv X is *Beta* with parameters $m \in \mathbb{Z}^+$ and $n \in \mathbb{Z}^+$ (i.e., $X \sim \mathrm{Beta}(m, n)$) if it has pdf

$$f(x) = \frac{(m+n-1)!}{(m-1)!(n-1)!} x^{m-1}(1-x)^{n-1},\ 0 < x < 1.$$

Remark: A $\mathrm{Beta}(1, 1)$ distribution simplifies to become the $\mathrm{U}(0, 1)$ distribution.

3. **Erlang**: A rv X is *Erlang* with parameters $n \in \mathbb{Z}^+$ and $\lambda > 0$ (i.e., $X \sim \text{Erlang}(n, \lambda)$) if it has pdf

$$f(x) = \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}, \ x > 0.$$

Remark: The Erlang(n, λ) distribution is actually a special case of the more general Gamma distribution in which n is extended to be any positive real number.

4. **Exponential**: A rv X is *Exponential* with parameter $\lambda > 0$ (i.e., $X \sim \text{EXP}(\lambda)$) if it has pdf

$$f(x) = \lambda e^{-\lambda x}, \ x > 0.$$

Remark: An Erlang$(1, \lambda)$ distribution actually simplifies to become the EXP(λ) distribution.

Expectation

Expectation: If $g(\,\cdot\,)$ is an arbitrary real-valued function, then

$$\mathbb{E}\big[g(X)\big] = \begin{cases} \sum_x g(x)p(x) & \text{, if } X \text{ is a discrete rv,} \\ \int_{-\infty}^{\infty} g(x) f(x) \, \mathrm{d}x & \text{, if } X \text{ is a continuous rv.} \end{cases}$$

Special choices of $g(\,\cdot\,)$:

1. $g(X) = X^n$, $n \in \mathbb{N} \implies \mathbb{E}\big[g(X)\big] = \mathbb{E}[X^n]$ is the n^{th} moment of X. In general, moments serve to describe the shape of a distribution. If $n = 0$, then $\mathbb{E}[X^0] = 1$. If $n = 1$, then $\mathbb{E}[X] = \mu_X$ is the *mean* of X.

2. $g(X) = \big(X - \mathbb{E}[X]\big)^2 \implies \mathbb{E}\big[g(X)\big] = \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big]$ is the *variance* of X. Note that

$$\text{Var}(X) = \sigma_X^2 = \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big] = \mathbb{E}[X^2] - \mathbb{E}[X]^2,$$

   or equivalently

$$\sigma_X^2 = \mathbb{E}\big[X(X-1)\big] + \mathbb{E}[X] - \mathbb{E}[X]^2.$$

   Related to this quantity, the *standard deviation* of X is $\sqrt{\text{Var}(X)} = \sigma_X$.

3. $g(X) = aX + b$, $a, b \in \mathbb{R}$ (i.e., $g(X)$ is a linear function of X). Note that

$$\begin{aligned} \mu_{aX+b} &= \mathbb{E}[aX + b] = a\mu_X + b, \\ \sigma_{aX+b}^2 &= \text{Var}(aX + b) = a^2 \sigma_X^2, \\ \sigma_{aX+b} &= \sqrt{\text{Var}(aX + b)} = |a| \sigma_X. \end{aligned}$$

Moment Generating Function

4. $g(X) = e^{tX}$, $t \in \mathbb{R} \implies \mathbb{E}\big[g(X)\big] = \mathbb{E}[e^{tX}]$ is the *moment generating function* (mgf) of X. This quantity is a function of t and is denoted by

$$\phi_X(t) = \mathbb{E}[e^{tX}].$$

First, $\phi_X(0) = \mathbb{E}[e^{0X}] = \mathbb{E}[1] = 1$. Moreover, making use of the linearity property of the expected value operator, note that

$$\begin{aligned}
\phi_X(t) &= \mathbb{E}[e^{tX}] \\
&= \mathbb{E}\left[\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right] \\
&= \mathbb{E}\left[\frac{t^0X^0}{0!} + \frac{t^1X^1}{1!} + \frac{t^2X^2}{2!} + \cdots + \frac{t^nX^n}{n!} + \cdots\right] \\
&= \mathbb{E}[X^0]\frac{t^0}{0!} + \mathbb{E}[X]\frac{t^1}{1!} + \mathbb{E}[X^2]\frac{t^2}{2!} + \cdots + \mathbb{E}[X^n]\frac{t^n}{n!} + \cdots,
\end{aligned}$$

implying that the n^{th} moment of X is simply the coefficient of $t^n/n!$ in the above series expansion.

We have: $\phi_X(t) = \mathbb{E}[t^X] = \mathbb{E}[X^0]\frac{t^0}{0!} + \mathbb{E}[X]\frac{t^1}{1!} + \mathbb{E}[X^2]\frac{t^2}{2!} + \cdots + \mathbb{E}[X^n]\frac{t^n}{n!} + \cdots$.

Remarks:

(1) Given the mgf of X, we can extract its n^{th} moment via

$$\mathbb{E}[X^n] = \phi_X^{(n)}(0) = \frac{\mathrm{d}^n}{\mathrm{d}t^n}\phi_X(t)\bigg|_{t=0} = \lim_{t\to 0}\frac{\mathrm{d}^n}{\mathrm{d}t^n}\phi_X(t),\ n \in \mathbb{N}.$$

Note that the 0^{th} derivative of a function is simply the function itself.

(2) A mgf uniquely characterizes the probability distribution of a rv (i.e., there exists a one-to-one correspondence between the mgf and the pmf/pdf of a rv). In other words, if two rvs X and Y have the same mgf, then they must have the same probability distribution (which we denote by $X \sim Y$). Thus, by finding the mgf of a rv, one has indeed determined its probability distribution.

Example 1.3. Suppose that $X \sim \text{BIN}(n, p)$. Find the mgf of X and use it to find $\mathbb{E}[X]$ and $\text{Var}(X)$.

Solution: Recall the binomial series formula

$$(a+b)^m = \sum_{x=0}^{m}\binom{m}{x}a^x b^{m-x},\ a, b \in \mathbb{R},\ m \in \mathbb{N}.$$

Using this formula, we obtain

$$\begin{aligned}\phi_X(t) &= \mathbb{E}[e^{tX}] \\ &= \sum_{x=0}^{n} e^{tx}\binom{n}{x}p^x(1-p)^{n-x} \\ &= \sum_{x=0}^{n}\binom{n}{x}(pe^t)^x(1-p)^{n-x} \\ &= (pe^t+1-p)^n,\ t\in\mathbb{R}.\end{aligned}$$

Then,

$$\phi'_X(t) = n(pe^t+1-p)^{n-1}pe^t \quad \text{and} \quad Q''_X(t) = n(pe^t+1-p)^{n-1}pe^t + npe^t(n-1)(pe^t+1-p)^{n-2}pe^t.$$

Thus,

$$\mathbb{E}[X] = \phi'_X(0) = n(pe^0+1-p)^{n-1}pe^0 = np,$$
$$\mathrm{Var}(X) = \mathbb{E}[X^2]-\mathbb{E}[X]^2 = \phi''_X(0) - n^2p^2 = np + np(n-1)p - n^2p^2 = np.$$

Joint Distributions

Joint Distributions: The following results are presented for the bivariate case mostly, but these ideas extend naturally to an arbitrary number of rvs.

Definition: The *joint cdf* of X and Y is

$$\begin{aligned}F(a,b) &= \mathbb{P}(X\le a, Y\le b) \\ &= \mathbb{P}\big(\{X\le a\}\cap\{Y\le b\}\big),\ a,b\in\mathbb{R}.\end{aligned}$$

Remark: If the joint cdf is known, then we can recover their marginal counterparts as follows:

$$F_X(a) = \mathbb{P}(X\le a) = F(a,\infty) = \lim_{b\to\infty} F(a,b),$$
$$F_Y(a) = \mathbb{P}(Y\le b) = F(\infty,b) = \lim_{a\to\infty} F(a,b).$$

Jointly Discrete Case:
Joint pmf:

$$p(x,y) = \mathbb{P}(X=x, Y=y)$$

Marginals:

$$p_X(x) = \mathbb{P}(X=x) = \sum_y p(x,y)$$
$$p_Y(y) = \mathbb{P}(Y=y) = \sum_x p(x,y)$$

Multinomial Distribution: Consider an experiment which is repeated $n\in\mathbb{Z}^+$ times, with one of $k\ge 2$ distinct outcomes possible each time. Let $p_1,p_2,\ldots,p_k$ denote the probabilities of the k types of outcomes (with $\sum_{i=1}^k p_i = 1$). If X_i, $i=1,2,\ldots,k$, counts the number of type-i outcomes to occur, then

$(X_1, X_2, \ldots, X_k)$ is *Multinomial* (i.e., $(X_1, X_2, \ldots, X_k) \sim \text{MN}(n, p_1, p_2, \ldots, p_k)$) with joint pmf

$$p(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}, \; x_i = 0, 1, \ldots, n \; \forall i \text{ and } \sum_{i=1}^{k} x_i = n$$

Remark: A $\text{MN}(n, p_1, 1-p_1)$ distribution simplifies to become the $\text{BIN}(n, p_1)$ distribution.

Jointly Continuous Case:
Joint pdf: The joint pdf $f(x,y)$ is a non-negative real-valued function which enables one to calculate probabilities of the form

$$\mathbb{P}(X \in A, Y \in B) = \int_B \int_A f(x,y) \, \mathrm{d}x \, \mathrm{d}y = \int_A \int_B f(x,y) \, \mathrm{d}x \, \mathrm{d}y$$

where A and B are sets of real numbers (e.g., intervals). As a result,

$$F(a,b) = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x,y) \, \mathrm{d}x \, \mathrm{d}y = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x,y) \, \mathrm{d}y \, \mathrm{d}x$$

Marginals:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y) \, \mathrm{d}y$$
$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y) \, \mathrm{d}x$$

Jointly Continuous Case:
Important Relationship:

$$f(x,y) = \frac{\partial^2}{\partial x \, \partial y} F(x,y)$$

Transformations: Let (X,Y) be jointly continuous with joint pdf $f(x,y0)$ and region of support $\mathcal{S}(X,Y)$. Suppose that the rvs V and W are given by $V = b_1(X,Y)$ and $W = b_2(X,Y)$, where the functions $v = b_1(x,y)$ and $w = b_2(x,y)$ defined a one-to-one transformation that maps the set $\mathcal{S}(X,Y)$ onto the set $\mathcal{S}(V,W)$. If x and y are expressed in terms of v and w (i.e., $x = h_1(v,w)$ and $y = h_2(v,w)$), then the joint pdf of V and W is given by

$$g(v,w) = \begin{cases} f\big(h_1(v,w), h_2(v,w)\big)|J|, & \text{if } (v,w) \in \mathcal{S}(V,W), \\ 0, & \text{elsewhere,} \end{cases}$$

where J is the *Jacobian* of the transformation given by

$$J = \frac{\partial x}{\partial v}\frac{\partial y}{\partial w} - \frac{\partial x}{\partial w}\frac{\partial y}{\partial v}.$$

Expectation

Expectation: If $g(\,\cdot\,,\,\cdot\,)$ denotes an arbitrary real-valued function, then

$$\mathbb{E}\big[g(X,Y)\big] = \begin{cases} \sum_x \sum_y g(x,y)p(x,y) & \text{, if } X \text{ and } Y \text{ are jointly discrete,} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)\,\mathrm{d}y\,\mathrm{d}x & \text{, if } X \text{ and } Y \text{ are jointly continuous.} \end{cases}$$

Remark: The order of summation/integration is irrelevant and can be interchanged.

Special choices of $g(\,\cdot\,)$:

1. $g(X,Y) = \big(X - \mathbb{E}[X]\big)\big(Y - \mathbb{E}[Y]\big) \implies \mathbb{E}\big[g(X,Y)\big] = \mathbb{E}\big[\big(X - \mathbb{E}[X]\big)\big(Y - \mathbb{E}[Y]\big)\big]$ is the *covariance* of X and Y. Note that

$$\operatorname{Cov}(X,Y) = \mathbb{E}\big[\big(X - \mathbb{E}[X]\big)\big(Y - \mathbb{E}[Y]\big)\big] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

and $\operatorname{Cov}(X,X) = \operatorname{Var}(X)$.

2. $g(X,Y) = aX + bY$, $a, b \in \mathbb{R}$ (i.e., $g(X,Y)$ is a linear combination of X and Y). Note that:

$$\begin{aligned} \mathbb{E}[aX + bY] &= a\,\mathbb{E}[X] + b\,\mathbb{E}[Y], \\ \operatorname{Var}(aX + bY) &= a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) + 2ab\operatorname{Cov}(X,Y). \end{aligned}$$

3. $g(X,Y) = e^{sX+tY}$, $s, t \in \mathbb{R} \implies \mathbb{E}\big[g(X,Y)\big] = \mathbb{E}[e^{sX+tY}]$ is the joint mgf of X and Y. A joint mgf (denoted by $\phi(s,t)$) also uniquely characterizes a joint probability distribution and can be used to calculate joint moments of X and Y via the formula

$$\mathbb{E}[X^m Y^n] = \phi^{(m,n)}(0,0) = \left(\frac{\partial^{m+n}}{\partial s^m\,\partial t^n}\phi(s,t)\right)_{s=0,t=0} = \lim_{s\to 0, t\to 0} \frac{\partial^{m+n}}{\partial s^m\,\partial t^n}\phi(s,t),\ m,n \in \mathbb{N}$$

Independence of Random Variables

Formal Definition: If X and Y are *independent* rvs if

$$\begin{aligned} F(a,b) &= \mathbb{P}(X \le a, Y \le b) \\ &= \mathbb{P}(X \le a)\,\mathbb{P}(Y \le b) \\ &= F_X(a)F_Y(b)\ \forall a, b \in \mathbb{R}. \end{aligned}$$

Equivalently, independence exists iff $p(x,y) = p_X(x)p_Y(y)$ (in the jointly discrete case) or $f(x,y) = f_X(x)f_Y(y)$ (in the jointly continuous case) $\forall x, y \in \mathbb{R}$.

Important Property: For arbitrary real-valued functions $g(\,\cdot\,)$ and $h(\,\cdot\,)$, if X and Y are independent, then

$$\mathbb{E}\big[g(X)h(Y)\big] = \mathbb{E}\big[g(X)\big]\,\mathbb{E}\big[h(Y)\big].$$

Remark: As a consequence of this property, $\operatorname{Cov}(X,Y) = 0$ if X and Y are independent, implying that $\operatorname{Var}(aX + bY) = a^2\operatorname{Var}(X) + b^2\operatorname{Var}(Y)$. However, if $\operatorname{Cov}(X,Y) = 0$, we cannot conclude that X and Y are independent (we can only say that X and Y are *uncorrelated*).

Example 1.4. Suppose that X and Y have joint pmf (and corresponding marginals) of the form

	$p(x,y)$	y = 0	1	$p_X(x)$
x	0	0.2	0	0.2
	1	0	0.6	0.6
	2	0.2	0	0.2
	$p_Y(y)$	0.4	0.6	1

Show that $\text{Cov}(X,Y) = 0$ holds, but X and Y are not independent.

Solution: Recall that $\text{Cov}(X,X) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$. Note that

$$\begin{aligned}\mathbb{E}[XY] &= \sum_x \sum_y xyp(x,y) \\ &= (0)(0)(0.2) + (0)(1)(0) + (1)(0)(0) + (1)(1)(0.6) + (2)(0)(0.2) + (2)(1)(0) \\ &= 0.6,\end{aligned}$$

$$\begin{aligned}\mathbb{E}[X] &= \sum_x xp_X(x) = (0)(0.2) + (1)(0.6) + (2)(0.2) = 1, \\ \mathbb{E}[Y] &= \sum_y yp_Y(y) = (0)(0.4) + (1)(0.6) = 0.6.\end{aligned}$$

Thus,

$$\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0.6 - (1)(0.6) = 0.$$

However, from the given table, it is clear that $p(2,0) = 0.2 \neq 0.08 = (0.2)(0.4) = p_X(2)p_Y(0)$. Thus, we conclude that while $\text{Cov}(X,Y) = 0$, X and Y are not independent.

Theorem 1.1. If $X_1, X_2, \ldots, X_n$ are independent rvs where $\phi_{X_i}(t)$ is the mgf of X_i, $i = 1, 2, \ldots, n$, then $T = \sum_{i=1}^n X_i$ has mgf $\phi_T(t) = \prod_{i=1}^n \phi_{X_i}(t)$.

Proof: Note that the mgf of T given by

$$\begin{aligned}\phi_T(t) &= \mathbb{E}[e^{tT}] \\ &= \mathbb{E}[e^{t(X_1+X_2+\cdots+X_n)}] \\ &= \mathbb{E}[e^{tX_1}e^{tX_2}\cdots e^{tX_n}] \\ &= \mathbb{E}[e^{tX_1}]\,\mathbb{E}[e^{tX_2}]\cdots\mathbb{E}[e^{tX_n}] && \text{by independence of } \{X_i\}_{i=1}^n \\ &= \phi_{X_1}(t)\phi_{X_2}(t)\cdots\phi_{X_n}(t) \\ &= \prod_{i=1}^n \phi_{X_i}(t).\end{aligned}$$

<u>Remarks</u>:

(1) Simply put, Theorem 1.1 states that the mgf of a sum of independent rvs is just the product of their individual mgfs.

(2) As a special case of the above result, note that $\phi_T(t) = \phi_{X_1}(t)^n$ if $X_1, X_2, \ldots, X_n$ is an independent and identically distributed (iid) sequence of rvs.

Example 1.5. Let $X_1, X_2, \ldots, X_m$ be an independent sequence of rvs where $X_i \sim \text{BIN}(n_i, p)$, $i = 1, 2, \ldots, m$. Find the distribution of $T = \sum_{i=1}^m X_i$.

Solution: Looking at the mgf of T, note that

$$\begin{aligned}
\phi_T(t) &= \prod_{i=1}^m \phi_{X_i}(t) && \text{by Theorem 1.1} \\
&= \prod_{i=1}^m (pe^t + 1 - p)^{n_i} && \text{using the result of Example of 1.3} \\
&= (pe^t + 1 - p)^{\sum_{i=1}^m n_i}, \ t \in \mathbb{R}.
\end{aligned}$$

By the mgf uniqueness property we recognize that $T = \sum_{i=1}^m X_i \sim \text{BIN}\big(\sum_{i=1}^m n_i, p\big)$.
Remark: As a special case of the above example, if $X_1, X_2, \ldots, X_m$ are iid BERN(p) rvs, then $T = \sum_{i=1}^m X_i \sim \text{BIN}(m, p)$.

Convergence of Random Variables

Modes of Convergence: If X_n, $n \in \mathbb{Z}^+$, and X are rvs, then

1. $X_n \to X$ *in distribution* iff
$$\lim_{n\to\infty} \mathbb{P}(X_n \le x) = \mathbb{P}(X \le x), \ \forall x \in \mathbb{R} \text{ at which } \mathbb{P}(X \le x) \text{ is continuous,}$$
2. $X_n \to X$ *in probability*, iff $\forall \varepsilon > 0$,
$$\lim_{n\to\infty} \mathbb{P}\big(|X_n - X| > \varepsilon\big) = 0,$$
3. $X_n \to X$ *almost surely (a.s.)* iff
$$\mathbb{P}\big(\lim_{n\to\infty} X_n = X\big) = 1.$$

Remarks:

(1) In probability theory, an event is said to happen a.s. if it happens with probability 1.

(2) The following implications hold true in general:
$$X_n \to X \text{ a.s.} \implies X_n \to X \text{ in probability} \implies X_n \to X \text{ in distribution.}$$

Strong Law of Large Numbers

Strong Law of Large Numbers (SLLN): If $X_1, X_2, \ldots, X_n$ is an iid sequence of rvs with common mean μ and $\mathbb{E}\big[|X_1|\big] < \infty$, then

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu \text{ a.s. as } n \to \infty.$$

Remark: The SLLN is one of the most important results in probability and statistics, indicating that the sample mean will, with probability 1, converge to the true mean of the underlying distribution as the sample size approaches infinity. In other words, if the same experiment or study is repeated independently many times, the average of the results of the trials must be close to the mean. The result gets closer to the mean as the number of trials is increased.

Chapter 2

Conditional Distributions and Conditional Expectation

Week 2
15th to 22nd September

2.1 Definitions and Construction

Jointly Discrete Case

Formulation: If X_1 and X_2 are both discrete rvs with joint pmf $p(x_1, x_2)$ and marginal pmfs $p_1(x_1)$ and $p_2(x_2)$, respectively, then the conditional distribution of X_1 given $X_2 = x_2$, denoted by $X_1 \mid (X_2 = x_2)$, is defined via its *conditional pmf*

$$p_{1|2}(x_1 \mid x_2) = \mathbb{P}(X_1 = x_1 \mid X_2 = x_2) = \frac{\mathbb{P}(X_1 = x_1, X_2 = x_2)}{\mathbb{P}(X_2 = x_2)} = \frac{p(x_1, x_2)}{p_2(x_2)},$$

provided that $p_2(x_2) > 0$. Similarly, the conditional distribution of $X_2 \mid (X_1 = x_1)$ is defined via its conditional pmf

$$p_{2|1}(x_2 \mid x_1) = \mathbb{P}(X_2 = x_2 \mid X_1 = x_1) = \frac{p(x_1, x_2)}{p_1(x_1)}, \text{ provided that } p_1(x_1) > 0.$$

Remarks:

(1) If X_1 and X_2 are independent, then $p(x_1, x_2) = p_1(x_1)p_2(x_2)\ \forall x_1, x_2 \in \mathbb{R}$, and so $p_{1|2}(x_1 \mid x_2) = p_1(x_1)$ and $p_{2|1}(x_2 \mid x_1) = p_2(x_2)$.

(2) These ideas extend beyond the simple bivariate case naturally. For example, suppose that X_1, X_2, and X_3 are discrete rvs. We can define the conditional distribution of (X_1, X_2) given $X_3 = x_3$ via its conditional pmf as follows:

$$p_{12|3}(x_1, x_2 \mid x_3) = \frac{\mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3)}{\mathbb{P}(X_3 = x_3)} = \frac{p(x_1, x_2, x_3)}{p_3(x_3)},$$

provided that $p_3(x_3) > 0$. Alternatively, we can define the conditional distribution of X_2 given $(X_1 = x_1, X_3 = x_3)$ via its conditional pmf given by

$$p_{2|13}(x_2 \mid x_1, x_3) = \frac{p(x_1, x_2, x_3)}{p_{13}(x_1, x_3)}, \text{ provided that } p_{13}(x_1, x_3) > 0,$$

where $p_{13}(x_1, x_3)$ is the joint pmf of X_1 and X_3.

Conditional Expectation: The *conditional mean* of $X_1 \mid (X_2 = x_2)$ is

$$\mathbb{E}[X_1 \mid X_2 = x_2] = \sum_{x_1} x_1 p_{1|2}(x_1 \mid x_2).$$

More generally, if $w(\,\cdot\,,\,\cdot\,)$, $h(\,\cdot\,)$, and $g(\,\cdot\,)$ are arbitrary real-valued functions, then

$$\mathbb{E}\big[w(X_1, X_2) \bigm| X_2 = x_2\big] = \mathbb{E}\big[w(X_1, x_2) \bigm| X_2 = x_2\big] = \sum_{x_1} w(x_1, x_2) p_{1|2}(x_1 \mid x_2)$$

and

$$\mathbb{E}\big[g(X_1)h(X_2) \bigm| X_2 = x_2\big] = \mathbb{E}\big[g(X_1)h(x_2) \bigm| X_2 = x_2\big] = h(x_2)\,\mathbb{E}\big[g(X_1) \bigm| X_2 = x_2\big].$$

As an immediate consequence, if $a, b \in \mathbb{R}$, then we obtain

$$\mathbb{E}\big[ag(X_1) + bh(X_1) \bigm| X_2 = x_2\big] = a\,\mathbb{E}\big[g(X_1) \bigm| X_2 = x_2\big] + b\,\mathbb{E}\big[h(X_1) \bigm| X_2 = x_2\big].$$

Furthermore, if we recall that $\mathbb{E}[X_1 + X_2] = \sum_{x_1} \sum_{x_2} (x_1 + x_2) p(x_1, x_2)$, then it correspondingly follows that

$$\begin{aligned}
\mathbb{E}[X_1 + X_2 \mid X_3 = x_3] &= \sum_{x_1}\sum_{x_2}(x_1 + x_2)p_{12|3}(x_1, x_2 \mid x_3) \\
&= \sum_{x_1}\sum_{x_2}(x_1 + x_2)\frac{p(x_1, x_2, x_3)}{p_3(x_3)} \\
&= \sum_{x_1}\sum_{x_2} x_1 \cdot \frac{p(x_1, x_2, x_3)}{p_3(x_3)} + \sum_{x_1}\sum_{x_2} x_2 \cdot \frac{p(x_1, x_2, x_3)}{p_3(x_3)} \\
&= \sum_{x_1}\frac{x_1}{p_3(x_3)}\sum_{x_2} p(x_1, x_2, x_3) + \sum_{x_2}\frac{x_2}{p_3(x_3)}\sum_{x_1} p(x_1, x_2, x_3) \\
&= \sum_{x_1}\frac{x_1}{p_3(x_3)}p_{13}(x_1, x_3) + \sum_{x_2}\frac{x_2}{p_3(x_3)}p_{23}(x_2, x_3) \\
&= \sum_{x_1} x_1 p_{1|3}(x_1 \mid x_3) + \sum_{x_2} x_2 p_{2|3}(x_2 \mid x_3) \\
&= \mathbb{E}[X_1 \mid X_3 = x_3] + \mathbb{E}[X_2 \mid X_3 = x_3].
\end{aligned}$$

We have: $\mathbb{E}[X_1 + X_2 \mid X_3 = x_3] = \mathbb{E}[X_1 \mid X_3 = x_3] + \mathbb{E}[X_2 \mid X_3 = x_3]$. In other words, the conditional expected value is also a **linear** operator. In fact, more generally, if $a_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$, then the same essential approach can be used to show that

$$\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i \,\middle|\, Y = y\right] = \sum_{i=1}^{n} a_i \,\mathbb{E}[X_i \mid Y = y].$$

Conditional Variance: If we take $g(X_1) = \big(X_1 - \mathbb{E}[X_1 \mid X_2 = x_2]\big)^2$, then

$$\mathbb{E}\big[g(X_1) \mid X_2 = x_2\big] = \mathbb{E}\Big[\big(X_1 - \mathbb{E}[X_1 \mid X_2 = x_2]\big)^2 \,\Big|\, X_2 = x_2\Big] = \mathrm{Var}(X_1 \mid X_2 = x_2)$$

is the *conditional variance* of $X_1 \mid (X_2 = x_2)$.

As with the calculation of variance, the following result provides an alternative (and often times preferred) way to calculate $\mathrm{Var}(X_1 \mid X_2 = x_2)$.

Theorem 2.1. $\mathrm{Var}(X_1 \mid X_2 = x_2) = \mathbb{E}[X_1^2 \mid X_2 = x_2] - \mathbb{E}[X_1 \mid X_2 = x_2]^2$.

Proof:

$$\begin{aligned}
\mathrm{Var}(X_1 \mid X_2 = x_2) &= \mathbb{E}\Big[\big(X_1 - \mathbb{E}[X_1 \mid X_2 = x_2]\big)^2 \,\Big|\, X_2 = x_2\Big] \\
&= \mathbb{E}\big[X_1^2 - 2X_1\,\mathbb{E}[X_1 \mid X_2 = x_2] + \mathbb{E}[X_1 \mid X_2 = x_2]^2 \,\big|\, X_2 = x_2\big] \\
&= \mathbb{E}[X_1^2] - 2\,\mathbb{E}[X_1 \mid X_2 = x_2]^2 + \mathbb{E}[X_1 \mid X_2 = x_2]^2 \\
&= \mathbb{E}[X_1^2 \mid X_2 = x_2] - \mathbb{E}[X_1 \mid X_2 = x_2]^2
\end{aligned}$$

Example 2.1. Suppose that X_1 and X_2 are discrete rvs having joint pmf of the form

$$p(x_1, x_2) = \begin{cases} 1/5 & \text{, if } x_1 = 1 \text{ and } x_2 = 0, \\ 2/15 & \text{, if } x_1 = 0 \text{ and } x_2 = 1, \\ 1/15 & \text{, if } x_1 = 1 \text{ and } x_2 = 2, \\ 1/5 & \text{, if } x_1 = 2 \text{ and } x_2 = 0, \\ 2/5 & \text{, if } x_1 = 1 \text{ and } x_2 = 1, \\ 0 & \text{, otherwise.} \end{cases}$$

Find the conditional distribution of $X_1 \mid (X_2 = 1)$. Also, calculate $\mathbb{E}[X_1 \mid X_2 = 1]$ and $\text{Var}(X_1 \mid X_2 = 1)$.

Solution: Note that for problems of this nature, it often helps to create a table summarizing the information:

		x_2			
	$p(x_1, x_2)$	0	1	2	$p_1(x_1)$
	0	0	2/15	0	2/15
x_1	1	1/5	2/5	1/15	2/3
	2	1/5	0	0	1/5
	$p_2(x_2)$	2/5	8/15	1/15	1

Then,

- $p_{1|2}(0 \mid 1) = \mathbb{P}(X_1 = 0 \mid X_2 = 1) = (2/15)/(8/15) = 1/4$, and
- $p_{1|2}(1 \mid 1) = \mathbb{P}(X_1 = 1 \mid X_2 = 1) = (2/5)/(8/15) = 3/4$.

Thus, the conditional pmf of $X_1 \mid (X_2 = 1)$ can be represented as follows:

x_1	0	1
$p_{1|2}(x_1 \mid 1)$	1/4	3/4

Note that $X_1 \mid (X_2 = 1) \sim$ BERN$(3/4)$. Thus, $\mathbb{E}[X_1 \mid X_2 = 1] = 3/4$ and $\text{Var}(X_1 \mid X_2 = 1) = 3/4(1 - 3/4) = 3/16$.

Example 2.2. For $i = 1, 2$, suppose that $X_i \sim$ BIN(n_i, p) where X_1 and X_2 are independent. Find the conditional distribution of X_1 given $X_1 + X_2 = m$.

Solution: We want to find the conditional pmf of $X_1 \mid (Y = m)$, where $Y = X_1 + X_2$. Let this conditional pmf be denoted by $p_{X_1|Y}(x_1 \mid m) = \mathbb{P}(X_1 = x_1 \mid Y = m)$. Recall from Example 1.5 that

$X_1 + X_2 \sim$ BIN$(n_1 + n_2, p)$.

$$\begin{aligned}
p_{X_1|Y}(x_1 \mid m) &= \frac{\mathbb{P}(X_1 = x_1, Y = m)}{\mathbb{P}(Y = m)} \\
&= \frac{\mathbb{P}(X_1 = x_1, X_1 + X_2 = m)}{\mathbb{P}(X_1 + X_2 = m)} \\
&= \frac{\mathbb{P}(X_1 = x_1, X_2 = m - x_1)}{\binom{n_1+n_2}{m} p^m (1-p)^{n_1+n_2-m}} \\
&= \frac{p_1(x_1) p_2(m - x_1)}{\binom{n_1+n_2}{m} p^m (1-p)^{n_1+n_2-m}} \\
&= \frac{\binom{n_1}{x_1} p^{x_1} (1-p)^{n_1 - x_1} \binom{n_2}{m-x_1} p^{m-x_1} (1-p)^{n_2-(m-x_1)}}{\binom{n_1+n_2}{m} p^m (1-p)^{n_1+n_2-m}}
\end{aligned}$$

provided that $0 \le x_1 \le n_1$, and $0 \le m - x_1 \le n_2$ (i.e., $m - n_2 \le x \le m$). Simplifying,

$$p_{X_1|Y}(x_1 \mid m) = \frac{\binom{n_1}{x_1}\binom{n_2}{m-x_1}}{\binom{n_1+n_2}{m}},$$

for $x_1 = \max\{0, m - n_2\}, \ldots, \min\{n_1, m\}$.

Remark: Looking at the conditional pmf we just obtained, we recognize that $X_1 \mid (X_1 + X_2 = m) \sim$ HG$(n_1 + n_2, n_1, m)$. The result that $X_1 \mid (X_1 + X_2 = m)$ has a hypergeometric distribution should not be all that surprising. Consider the sequence of $n_1 + n_2$ Bernoulli trials represented visually as follows:

$$\underbrace{\boxed{1}\ \ \boxed{2}\ \ \cdots\ \ \boxed{n_1}\ \ \boxed{1}\ \ \boxed{2}\ \ \cdots\ \ \boxed{n_2}}_{n_1 + n_2 \text{ trials}}$$

Of these $n_1 + n_2$ trials in which m of them were known to be successes, we want x_1 successes to have occurred among the first n_1 trials (thereby implying that $m - x_1$ successes are obtained during the final n_2 trials). Since any of these trials were equally likely to be a success (i.e., the same success probability p is assumed), the desired result ends up being the obtained hypergeometric probability.

Example 2.3. Let $X_1, X_2, \ldots, X_m$ be independent rvs where $X_i \sim$ POI(λ_i), $i = 1, 2, \ldots, m$. Define $Y = \sum_{i=1}^m X_i$. Find the conditional distribution of $X_j \mid (Y = n)$.

Solution: We are interested in the conditional pmf of $X_j \mid (Y = n)$, to be denoted by

$$\begin{aligned}
p_{X_j|Y}(x_j \mid n) &= \mathbb{P}(X_j = x_j \mid Y = n) \\
&= \frac{\mathbb{P}(X_j = x_j, Y = n)}{\mathbb{P}(Y = n)} \\
&= \frac{\mathbb{P}\Big(X_j = x_j, \sum_{i=1}^m X_i = n\Big)}{\mathbb{P}(Y = n)}
\end{aligned}$$

First, we investigate the numerator:

$$\begin{aligned}\mathbb{P}\Big(X_j = x_j, \sum_{i=1}^{m} X_i = n\Big) &= \mathbb{P}\Big(X_j = x_j, X_j + \sum_{i=1, i\neq j}^{m} X_i = n\Big)\\ &= \mathbb{P}\Big(X_j = x_j, \sum_{i=1, i\neq j}^{m} X_i = n - x_j\Big)\\ &= \mathbb{P}(X_j = x_j)\,\mathbb{P}\Big(\sum_{i=1, i\neq j}^{m} X_i = n - x_j\Big)\end{aligned}$$

where the last equality follows due to the independence of $\{X_i\}_{i=1}^m$. We are given that $X_j \sim \text{POI}(\lambda_j)$. Due to the result of Exercise 1.1, it follows that

$$\sum_{i=1, i\neq j}^{m} X_i \sim \text{POI}\Big(\sum_{i=1, i\neq j}^{m} \lambda_i\Big).$$

By the same result, we also have that

$$Y = \sum_{i=1}^{m} X_i \sim \text{POI}\Big(\sum_{i=1}^{m} \lambda_i\Big).$$

Therefore,

$$p_{X_j|Y}(x_j \mid n) = \frac{\frac{e^{-\lambda_j}\lambda_j^{x_j}}{x_j!}\,\frac{e^{-\sum_{i=1,i\neq j}^{m}\lambda_i}(\sum_{i=1,i\neq j}^{m}\lambda_i)^{n-x_j}}{(n-x_j)}}{\frac{e^{-\sum_{i=1}^{m}\lambda_i}(\sum_{i=1}^{m}\lambda_i)^{n}}{n!}}$$

provided that $x_j \geq 0$ and $n - x_j \geq 0$ which implies $0 \leq x_j \leq n$. Thus,

$$\begin{aligned}p_{X_j|Y}(x_j \mid n) &= \binom{n}{x_j}\frac{\lambda_j^{x_j}(\lambda_Y - \lambda_j)^{n-x_j}}{\lambda_Y^n}\\ &= \binom{n}{x_j}\Big(\frac{\lambda_j}{\lambda_Y}\Big)^{x_j}\Big(1 - \frac{\lambda_j}{\lambda_Y}\Big)^{n-x_j},\ x_j = 0, 1, \ldots, n\end{aligned}$$

where $\lambda_Y = \sum_{i=1}^m \lambda_i$ and note that $\lambda_Y^{x_j}\lambda_Y^{n-x_j} = \lambda_Y$. We see that

$$X_j \mid (Y = n) \sim \text{BIN}\Big(n, \frac{\lambda_j}{\sum_{i=1}^m \lambda_i}\Big).$$

Example 2.4. Suppose that $X \sim \text{POI}(\lambda)$ and $Y \mid (X = x) \sim \text{BIN}(x, p)$. Find the conditional distribution of $X \mid (Y = y)$.

Solution: We want to calculate the conditional pmf of $X \mid (Y = y)$, to be denoted by

$$p_{X|Y}(x \mid y) = \mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}.$$

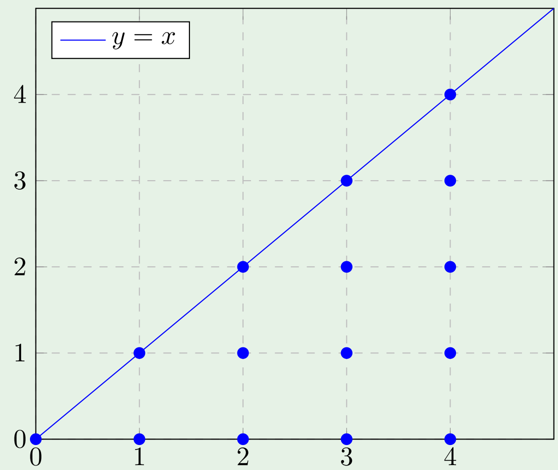
First, note that

$$\mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)},$$

which implies that

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(Y = y \mid X = x)\,\mathbb{P}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}\binom{x}{y}p^y(1-p)^{x-y},$$

for $x = 0, 1, 2, \ldots$ and $y = 0, 1, \ldots, x$. Note that the range of y depends on the values of x. A graphical display of the region is given below:



We may rewrite this region with the range of x depending on the values of y. Specifically, note that $x = 0, 1, 2, \ldots$ and $y = 0, 1, \ldots, x$ is equivalent to $y = 0, 1, 2, \ldots$ and $x = y, y+1, y+2, \ldots$. We use this alternative region to find the marginal pmf of Y.

$$\begin{aligned}
\mathbb{P}(Y = y) &= \sum_x \mathbb{P}(X = x, Y = y) \\
&= \sum_{x=y}^{\infty} e^{-\lambda}\frac{\lambda^x}{x!}\binom{x}{y}p^y(1-p)^{x-y} \\
&= \sum_{x=y}^{\infty} e^{-\lambda}\frac{\lambda^x}{x!}\frac{x!}{y!(x-y)!}p^y(1-p)^{x-y} \\
&= \frac{e^{-\lambda}}{y!}p^y\sum_{x=y}^{\infty}\frac{\lambda^x(1-p)^{x-y}}{(x-y)!}\lambda^{-y}\lambda^y \\
&= \frac{e^{-\lambda}(\lambda p)^y}{y!}\sum_{x=y}^{\infty}\frac{\big(\lambda(1-p)\big)^{x-y}}{(x-y)!} && \text{let } z = x - y \\
&= \frac{e^{-\lambda}(\lambda p)^y}{y!}e^{\lambda(1-p)} \\
&= \frac{e^{-\lambda p}(\lambda p)^y}{y!} && y = 0, 1, 2, \ldots
\end{aligned}$$

In fact, $Y \sim \text{POI}(\lambda p)$. Therefore,

$$\begin{aligned}
p_{X|Y}(x \mid y) &= \frac{\frac{e^{-\lambda}\lambda^x}{x!}\frac{x!}{y!(x-y)!}p^y(1-p)^{x-y}}{\frac{e^{-\lambda p}(\lambda p)^y}{y!}} \\
&= \frac{e^{-\lambda(1-p)}\big(\lambda(1-p)\big)^{x-y}}{(x-y)!},
\end{aligned}$$

for $x = y, y+1, \ldots$.

Remark: The above conditional pmf is recognized as that of a **shifted** Poisson distribution (y units to the right). Specifically, we have that

$$X \mid (Y = y) \sim W + y$$

where $W \sim \text{POI}\big(\lambda(1-p)\big)$.

Formulation: In the jointly discrete case, it was natural to define:

$$p_{X|Y}(x \mid y) = \mathbb{P}(X = x \,|\, Y = y) = \mathbb{P}(X = x, Y = y) / \, \mathbb{P}(Y = y).$$

Strictly speaking, this no longer makes sense in a continuous context since $f(x, y) \neq \mathbb{P}(X = x, Y = y)$ and $f_Y(y) \neq \mathbb{P}(Y = y)$. However, for small positive values of $\mathrm{d}y$ (as the figure below shows), $\mathbb{P}(y \leq Y \leq y + \mathrm{d}y) \approx f_Y(y)\,\mathrm{d}y$.

Formally,

$$f_Y(y) = \lim_{\mathrm{d}y \to 0} \frac{\mathbb{P}(y \leq Y \leq y + \mathrm{d}y)}{\mathrm{d}y}.$$

Similarly,

$$f(x, y) = \lim_{\mathrm{d}x \to 0, \mathrm{d}y \to 0} \frac{\mathbb{P}(x \leq X \leq x + \mathrm{d}x, y \leq Y \leq y + \mathrm{d}y)}{\mathrm{d}x\,\mathrm{d}y},$$

which implies that $\mathbb{P}(x \leq X \leq x + \mathrm{d}x, y \leq Y \leq y + \mathrm{d}y) \approx f(x, y)\,\mathrm{d}x\,\mathrm{d}y$. For small positive values of $\mathrm{d}x$ and $\mathrm{d}y$, consider now

$$\begin{aligned}
\mathbb{P}(x \leq X \leq x + \mathrm{d}x \,|\, y \leq Y \leq y + \mathrm{d}y) &= \frac{\mathbb{P}(x \leq X \leq x + \mathrm{d}x \,|\, y \leq Y \leq y + \mathrm{d}y)}{\mathbb{P}(y \leq Y \leq y + \mathrm{d}y)} \\
&\approx \frac{f(x, y)\,\mathrm{d}x\,\mathrm{d}y}{f_Y(y)\,\mathrm{d}y} \\
&= \frac{f(x, y)}{f_Y(y)}\,\mathrm{d}x.
\end{aligned}$$

As a result, we formally define the *conditional pdf* of X given $Y = y$ (again to be denoted by $X \mid (Y = y)$) as

$$f_{X|Y}(x \mid y) = \frac{f(x, y)}{f_Y(y)} = \lim_{\mathrm{d}x \to 0, \mathrm{d}y \to 0} \frac{\mathbb{P}(x \leq X \leq x + \mathrm{d}x, y \leq Y \leq y + \mathrm{d}y)}{\mathrm{d}x}.$$

Remark: In the jointly continuous case, the conditional probability of an event of the form $\{a \leq X \leq b\}$ given $Y = y$ would be calculated as

$$\mathbb{P}(a \leq X \leq b \,|\, Y = y) = \int_a^b f_{X|Y}(x \mid y)\,\mathrm{d}x = \frac{\int_a^b f(x, y)\,\mathrm{d}x}{f_Y(y)},$$

which we can also express as

$$\mathbb{P}(a \leq X \leq b \mid Y = y) = \frac{\int_a^b f(x, y)\,\mathrm{d}x}{\int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x}.$$

In other words, we could view this as a way of assigning probability to an event $\{a \leq X \leq b\}$ over a "slice," $Y = y$, of the (joint) region of support for the pair of rvs X and Y.

Example 2.5. Suppose that the joint pdf of X and Y is given by

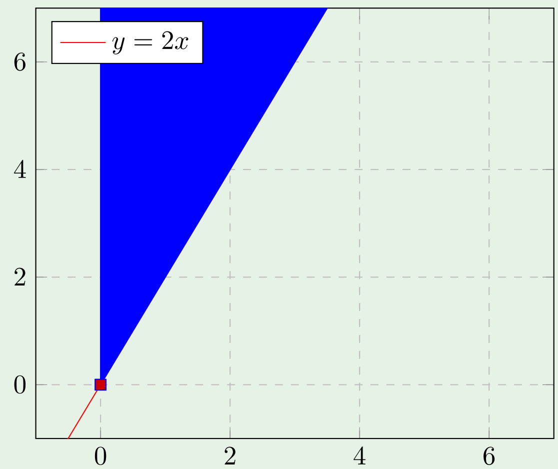
$$f(x, y) = \begin{cases} 5e^{-3x-y}, & \text{if } 0 \leq 2x \leq y < \infty, \\ 0, & \text{elsewhere.} \end{cases}$$

Determine the conditional distribution of $Y \mid (X = x)$ where $0 \leq x < \infty$.

Solution: We wish to find the conditional pdf of $Y \mid (X = x)$ given by

$$f_{Y|X}(y \mid x) = \frac{f(x,y)}{f_X(x)}$$

The region of support for this joint distribution looks like:



For $0 < x < \infty$:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y \\ &= \int_{2x}^{\infty} 5e^{-3x-y}\,\mathrm{d}y \\ &= \left[5e^{-3x}(-e^{-y})\right]_{y=2x}^{y=\infty} \\ &= 5e^{-3x}e^{-2x} \\ &= 5e^{-5x} \end{aligned}$$

Note that $X \sim \text{EXP}(5)$. Finally, we get:

$$f_{Y|X}(y \mid x) = \frac{5e^{-3x-y}}{5e^{-5x}} = e^{-y+2x},\ y > 2x.$$

Remark: The conditional pdf of $Y \mid (X = x)$ is recognized as that of a *shifted exponential distribution* ($2x$ units to the right). Specifically, we have that $Y \mid (X = x) \sim W + 2x$, where $W \sim \text{EXP}(1)$.

Conditional Expectation: If X and Y are jointly continuous rvs and $g(\,\cdot\,)$ is an arbitrary real-valued function, then the *conditional expectation* of $g(X)$ given $Y = y$ is

$$\mathbb{E}[g(X) \mid Y = y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x \mid y)\,\mathrm{d}x,$$

and so the conditional mean of $X \mid (Y = y)$ is given by

$$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\,\mathrm{d}x.$$

Example 2.6. Suppose that the joint pdf of X and Y is given by

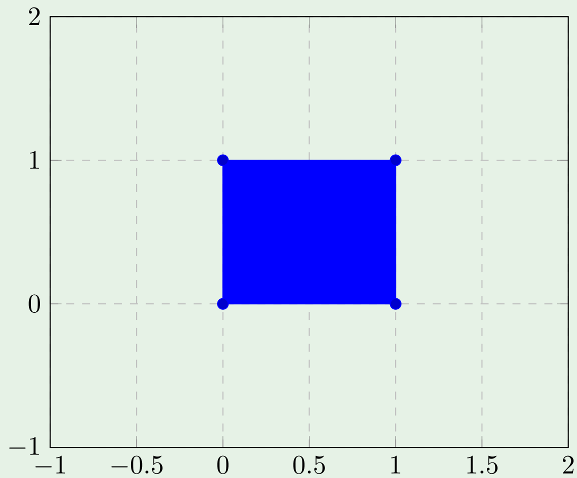
$$f(x,y) = \begin{cases} \dfrac{12}{5}x(2-x-y), & \text{if } 0<x<1 \text{ and } 0<y<1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the conditional distribution of X given $Y=y$ where $0<y<1$, and use it to calculate its conditional mean.

Solution: Using our earlier theory, we wish to find the conditional pdf of $X \mid (Y=y)$ given by

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

The region of support for this joint distribution of X and Y look like:



For $0<y<1$,

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x \\ &= \int_0^1 \frac{12}{5}x(2-x-y)\,\mathrm{d}x \\ &= \frac{12}{5}\int_0^1 (2x-x^2-xy)\,\mathrm{d}x \\ &= \frac{12}{5}\left[x^2-\frac{x^3}{3}-\frac{x^2y}{3}\right]_{x=0}^{x=1} \\ &= \frac{12}{5}\left(1-\frac{1}{3}-\frac{y}{2}\right) \\ &= \frac{2(4-3y)}{5} \end{aligned}$$

You can verify this by integrating $f_Y(y)$ over the support of Y (to get 1). Thus,

$$f_{X|Y}(x \mid y) = \frac{12/5x(2-x-y)}{2/5(4-3y)} = \frac{6x(2-x-y)}{4-3y}, \ 0<x<1$$

The conditional mean of X given $Y = y$ is:

$$\begin{aligned}
\mathbb{E}[X \mid Y = y] &= \int_0^1 x \frac{6x(2-x-y)}{4-3y}\,\mathrm{d}x \\
&= \frac{6}{4-3y}\int_0^1 (2x^2 - x^3 - x^2 y)\,\mathrm{d}x \\
&= \frac{6}{4-3y}\left[\frac{2x^3}{3} - \frac{x^4}{4} - \frac{x^3 y}{3}\right]_{x=0}^{x=1} \\
&= \frac{6}{4-3y}\left(\frac{2}{3} - \frac{1}{4} - \frac{y}{3}\right) \\
&= \frac{5-4y}{2(4-3y)}
\end{aligned}$$

Conditional Variance: Likewise, as in the jointly discrete case, we can also consider the notion of conditional variance, which retains the same definition as before:

$$\mathrm{Var}(X \mid Y = y) = \mathbb{E}\Big[\big(X - \mathbb{E}[X \mid Y = y]\big)^2 \,\Big|\, Y = y\Big] = \mathbb{E}[X^2 \mid Y = y] - \mathbb{E}[X \mid Y = y]^2.$$

A fact that is becoming more and more evident is that conditional expectation inherits many of the properties from regular expectation. Moreover, the same properties concerning conditional expectation that held in the jointly discrete case continue to hold true in the jointly continuous case (as we are effectively replacing summation with integration).

Example 2.6. (*continued*) Calculate $\mathrm{Var}(X \mid Y = y)$ where $0 < y < 1$ and the joint pdf of X and Y is given by

$$f(x,y) = \begin{cases} \frac{12}{5}x(2-x-5), & \text{if } 0 < x < 1 \text{ and } 0 < y < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Solution: Our earlier results tell us that

$$\begin{aligned}
\mathbb{E}[X^2 \mid Y = y] &= \int_0^1 x^2 \frac{6x(2-x-y)}{4-3y}\,\mathrm{d}x \\
&= \frac{6}{4-3y}\int_0^1 (2x^3 - x^4 - x^3 y)\,\mathrm{d}x \\
&= \frac{6}{4-3y}\left[\frac{x^4}{2} - \frac{x^5}{5} - \frac{x^4 y}{4}\right]_{x=0}^{x=1} \\
&= \frac{6}{4-3y}\left(\frac{1}{2} - \frac{1}{5} - \frac{y}{4}\right) \\
&= \frac{3(6-5y)}{10(4-3y)}.
\end{aligned}$$

Therefore, this leads to

$$\begin{aligned}
\mathrm{Var}(X \mid Y = y) &= \mathbb{E}[X^2 \mid Y = y] - \mathbb{E}[X \mid Y = y]^2 \\
&= \frac{3(6-5y)}{10(4-3y)} - \frac{(5-4y)^2}{4(4-3y)^2} \\
&= \frac{19 + 2y(5y-14)}{20(4-3y)^2}.
\end{aligned}$$

Mixed Case

We can also consider conditional distributions where the rvs are neither jointly continuous nor jointly discrete. To consider such a situation, suppose X is a continuous rv having pdf $f_X(x)$ and Y is a discrete rv having pmf $p_Y(y)$.
If we focus on the conditional distribution of X given $Y = y$, then let us look at the following quantity:

$$\begin{aligned}\frac{\mathbb{P}(x \le X \le x + \mathrm{d}x \mid Y = y)}{\mathrm{d}x} &= \frac{\mathbb{P}(x \le X \le x + \mathrm{d}x, Y = y)}{\mathrm{d}x\,\mathbb{P}(Y = y)} \\ &= \frac{\mathbb{P}(x \le X \le x + \mathrm{d}x)\,\mathbb{P}(Y = y \mid x \le X \le x + \mathrm{d}x)}{\mathrm{d}x\,\mathbb{P}(Y = y)} \\ &= \frac{\mathbb{P}(Y = y \mid x \le X \le x + \mathrm{d}x)}{\mathbb{P}(Y = y)}\,\frac{\mathbb{P}(x \le X \le x + \mathrm{d}x)}{\mathrm{d}x},\end{aligned}$$

where $\mathrm{d}x$ is again, a small positive value.
By letting $\mathrm{d}x \to 0$, we can formally define the conditional pdf of $X \mid (Y = y)$ as follows:

$$\begin{aligned}f(x \mid y) &= \lim_{\mathrm{d}x \to 0} \frac{\mathbb{P}(x \le X \le x + \mathrm{d}x \mid Y = y)}{\mathrm{d}x} \\ &= \lim_{\mathrm{d}x \to 0} \frac{\mathbb{P}(Y = y \mid x \le X \le x + \mathrm{d}x)}{\mathbb{P}(Y = y)}\,\frac{\mathbb{P}(x \le X \le x + \mathrm{d}x)}{\mathrm{d}x} \\ &= \frac{\mathbb{P}(Y = y \mid X = x)}{\mathbb{P}(Y = y)} f_X(x) \\ &= \frac{p(y \mid x) f_X(x)}{p_Y(y)},\end{aligned}$$

where $p(y \mid x) = \mathbb{P}(Y = y \mid X = x)$ is defined as the conditional pmf of $Y \mid (X = x)$. Note that since $f(x \mid y)$ is a pdf, it follows that

$$\int_{-\infty}^{\infty} f(x \mid y)\,\mathrm{d}x = 1 \implies p_Y(y) = \int_{-\infty}^{\infty} p(y \mid x) f_X(x)\,\mathrm{d}x.$$

Similarly, we can also write

$$p(y \mid x) = \frac{f(x \mid y) p_Y(y)}{f_X(x)}.$$

Since $p(y \mid x)$ is a pmf, we have that

$$\sum_y p(y \mid x) = 1 \implies f_X(x) = \sum_y f(x \mid y) p_Y(y).$$

Example 2.7. Suppose that $X \sim \mathrm{U}(0, 1)$ and $Y \mid (X = x) \sim \mathrm{BERN}(x)$. Find the conditional distribution of $X \mid (Y = y)$.

Solution: We wish to find the conditional pdf of $X \mid (Y = y)$ given by

$$f(x \mid y) = \frac{p(y \mid x) f_X(x)}{p_Y(y)}$$

Based on the given information, we have

$$\begin{aligned}f_X(x) &= 1,\ 0 < x < 1, \\ p(y \mid x) &= x^y (1 - x)^{1-y},\ y = 0, 1.\end{aligned}$$

For $y = 0, 1$, note that

$$\begin{aligned} p_Y(y) &= \int_{-\infty}^{\infty} p(y \mid x) f_X(x) \, \mathrm{d}x \\ &= \int_0^1 x^y (1-x)^{1-y} (1) \, \mathrm{d}x \end{aligned}$$

- For $y = 0 \implies p_Y(0) = \int_0^1 (1-x) \, \mathrm{d}x = \big[x - x^2/2\big]_{x=0}^{x=1} = 1/2$.
- For $y = 1 \implies p_Y(1) = \int_0^1 x \, \mathrm{d}x = \big[x^2/2\big]_{x=0}^{x=1} = 1/2$.

In other words, we have that

$$p_Y(y) = \frac{1}{2}, \; y = 0, 1 \implies Y \sim \text{BERN}\left(\frac{1}{2}\right)$$

Thus, for $y = 0, 1$, we ultimately obtain

$$f(x \mid y) = \frac{x^y (1-x)^{1-y} (1)}{1/2} = 2x^y (1-x)^{1-y}, \; 0 < x < 1.$$

2.2 Computing Expectation by Conditioning

An Important Observation

As before, let $g(\,\cdot\,)$ be an arbitrary real-valued function. In general, we recognize that $\mathbb{E}\big[g(X) \bigm| Y = y\big] = v(y)$, where $v(y)$ is some function of y. With this in mind, let us make the following definition:

$$\mathbb{E}\big[g(X) \bigm| Y\big] = \mathbb{E}\big[g(X) \bigm| Y = y\big]\big|_{y=Y} = v(Y).$$

Functions of rvs are, once again, rvs themselves. Therefore, it makes sense to consider the expected value of $v(Y)$. In this regard, we would obtain:

$$\begin{aligned} \mathbb{E}\big[\mathbb{E}\big[g(X) \bigm| Y\big]\big] &= \mathbb{E}\big[v(Y)\big] \\ &= \begin{cases} \sum_y v(y) p_Y(y) & \text{, if } Y \text{ is discrete,} \\ \int_{-\infty}^{\infty} v(y) f_Y(y) \, \mathrm{d}y & \text{, if } Y \text{ is continuous,} \end{cases} \\ &= \begin{cases} \sum_y \mathbb{E}\big[g(X) \bigm| Y = y\big] p_Y(y) & \text{, if } Y \text{ is discrete,} \\ \int_{-\infty}^{\infty} \mathbb{E}\big[g(X) \bigm| Y = y\big] f_Y(y) \, \mathrm{d}y & \text{, if } Y \text{ is continuous.} \end{cases} \end{aligned}$$

Law of Total Expectation

The following important result is regarded as the *law of total expectation*.

Theorem 2.2. For rvs X and Y, $\mathbb{E}\big[g(X)\big] = \mathbb{E}\big[\mathbb{E}\big[g(X) \bigm| Y\big]\big]$.

Proof: Without loss of generality, assume that X and Y are jointly continuous rvs. From above, we have

$$\begin{aligned}\mathbb{E}\big[\mathbb{E}\big[g(X) \,\big|\, Y\big]\big] &= \int_{-\infty}^{\infty} \mathbb{E}\big[g(X) \,\big|\, Y = y\big] f_Y(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x) f_{X|Y}(x \mid y)\,\mathrm{d}x f_Y(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)\frac{f(x,y)}{f_Y(y)} f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x) f(x,y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} g(x)\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x \\ &= \mathbb{E}\big[g(X)\big]\end{aligned}$$

Remark: Using a similar method of proof, the result of Theorem 2.2 can naturally be extended as follows:

$$\mathbb{E}\big[g(X,Y)\big] = \mathbb{E}\big[\mathbb{E}\big[g(X,Y) \,\big|\, Y\big]\big].$$

The usefulness of the law of total expectation is well-demonstrated in the following example.

Example 2.8. Suppose that $X \sim \text{GEO}_t(p)$ with pmf $p_X(x) = (1-p)^{x-1}p$, $x = 1, 2, 3, \ldots$. Calculate $\mathbb{E}[X]$ and $\mathrm{Var}(X)$ using the law of total expectation.

Solution: With $X \sim \text{GEO}_t(p)$, recall that X actually models the number of (independent) trials necessary to obtain the first success. Define:

$$Y = \begin{cases} 0 & \text{, if the 1}^{\text{st}}\text{ trial is a failure,} \\ 1 & \text{, if the 1}^{\text{st}}\text{ trial is a success.}\end{cases}$$

We observe that $Y \sim \text{BERN}(p)$, so that $p_Y(0) = 1 - p$ and $p_Y(1) = p$.
Note:

- $X \mid (Y = 1)$ is degenerate at 1 (i.e., X given $Y = 1$ is equal to 1 with probability 1).
- $X \mid (Y = 0)$ is equivalent in distribution $1 + X$ (i.e., $X \mid (Y = 0) \sim 1 + X$).

By the law of total expectation, we obtain:

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}\big[\mathbb{E}[X \mid Y]\big] \\ &= \sum_{y=0}^{1} \mathbb{E}[X \mid Y = y] p_Y(y) \\ &= (1-p)\,\mathbb{E}[X \mid Y = y] p_Y(y) \\ &= (1-p)\,\mathbb{E}[X \mid Y = 0] + p\,\mathbb{E}[X \mid Y = 1] \\ &= (1-p)\,\mathbb{E}[1 + X] + p \\ &= (1-p) + (1-p)\,\mathbb{E}[X] + p \\ &= 1 + (1-p)\,\mathbb{E}[X],\end{aligned}$$

which implies that $(1 - (1-p))\,\mathbb{E}[X] = 1$, or simply $\mathbb{E}[X] = 1/p$. Similarly, we use the law of total expectation to get

$$\begin{aligned}
\mathbb{E}[X^2] &= \mathbb{E}\big[\mathbb{E}[X^2 \mid Y]\big] \\
&= \sum_{y=0}^{1} \mathbb{E}[X^2 \mid Y = y] p_Y(y) \\
&= (1-p)\,\mathbb{E}[X^2 \mid Y = 0] + p\,\mathbb{E}[X^2 \mid Y = 1] \\
&= (1-p)\,\mathbb{E}\big[(1+X)^2\big] + p \\
&= (1-p)\big(\mathbb{E}[X^2] + 2\,\mathbb{E}[X] + 1\big) + p \\
&= 1 + (1-p)\,\mathbb{E}[X^2] + \frac{2(1-p)}{p},
\end{aligned}$$

which implies that

$$\big(1-(1-p)\big)\,\mathbb{E}[X] = \frac{p + 2(1-p)}{p}$$

or simply

$$\mathbb{E}[X^2] = \frac{p+2-2p}{p^2} = \frac{2-p}{p^2}$$

Finally,

$$\mathrm{Var}(X) = \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{1-p}{p^2}$$

Remarks:

(1) Note that the obtained mean and variance agree with known results. Moreover, the above procedure relied only on basic manipulations and did not involve any complicated sums or the differentiation of a mgf.

(2) As part of the above solution, we claimed that $X \mid (Y = 0) \sim Z$ where $Z = 1 + X$, and this implied that $\mathbb{E}[X^2 \mid Y = 0] = \mathbb{E}\big[(1 + X)^2\big]$. To see why this holds true formally, consider first

$$p_{X|Y}(x \mid 0) = \mathbb{P}(X = x \mid Y = 0) = \frac{\mathbb{P}(X = x, Y = 0)}{\mathbb{P}(Y = 0)} = \frac{\mathbb{P}(X = x, Y = 0)}{1-p}.$$

Note that

$$\begin{aligned}
\mathbb{P}(X = x, Y = 0) &= \mathbb{P}(1^{\text{st}} \text{ trial is a failure and } x \text{ total trials needed to get } 1^{\text{st}} \text{ success}) \\
&= \mathbb{P}(1^{\text{st}} \text{ trial is a failure, next } x-2 \text{ trials are failures, and } x^{\text{th}} \text{ trial is a success}) \\
&= (1-p)(1-p)^{x-2}p \text{ due to independence of trials.}
\end{aligned}$$

Thus,

$$p_{X|Y}(x \mid 0) = \frac{(1-p)(1-p)^{x-2}p}{1-p} = (1-p)^{x-2}p,\ x = 2, 3, 4, \dots.$$

On the other hand, note that

$$\begin{aligned}
p_Z(z) &= \mathbb{P}(Z = z) \\
&= \mathbb{P}(1 + X = z) \\
&= \mathbb{P}(X = z-1) \\
&= (1-p)^{(z-1)-1}p \\
&= (1-p)^{z-2}p,\ z = 2, 3, 4, \dots.
\end{aligned}$$

Since these two pmfs are identical, it follows that $X \mid (Y = 0) \sim Z$. As a further consequence, for an arbitrary real-valued function $g(\,\cdot\,)$, we must have that

$$\mathbb{E}\big[g(X) \,\big|\, Y = 0\big] = \mathbb{E}\big[g(Z)\big] = \mathbb{E}\big[g(1+X)\big].$$

Computing Variances by Conditioning

In recognizing that $\mathbb{E}\big[g(X) \bigm| Y = y\big]$ is a function of y, it similarly follows that $\mathrm{Var}(X \mid Y = y)$ is also a function of y. Therefore, we can make the following definition:

$$\mathrm{Var}(X \mid Y) = \mathrm{Var}(X \mid Y = y)\big|_{y=Y}.$$

Since $\mathrm{Var}(X \mid Y)$ is a function of Y, it is a rv as well, meaning that we could take its expected value. The following result, usually referred to as the *conditional variance formula*, provides a convenient way to calculate variance through the use of conditioning.

Theorem 2.3. For rvs X and Y, $\mathrm{Var}(X) = \mathbb{E}\big[\mathrm{Var}(X \mid Y)\big] + \mathrm{Var}\big(\mathbb{E}[X \mid Y]\big)$.

Proof: First, consider the term $\mathbb{E}\big[\mathrm{Var}(X \mid Y)\big]$. Since

$$\mathrm{Var}(X \mid Y = y) = \mathbb{E}[X^2 \mid Y = y] - \mathbb{E}[X \mid Y = y]^2,$$

it follows that

$$\mathrm{Var}(X \mid Y) = \mathbb{E}[X^2 \mid Y] - \mathbb{E}[X \mid Y]^2,$$

which yields (by Theorem 2.2)

$$\begin{aligned}
\mathbb{E}\big[\mathrm{Var}(X \mid Y)\big] &= \mathbb{E}\big[\mathbb{E}[X^2 \mid Y] - \mathbb{E}[X \mid Y]^2\big] \\
&= \mathbb{E}\big[\mathbb{E}[X^2 \mid Y]\big] - \mathbb{E}\big[\mathbb{E}[X \mid Y]^2\big] \\
&= \mathbb{E}[X^2] - \mathbb{E}\big[\mathbb{E}[X \mid Y]^2\big].
\end{aligned}$$

Next, recall

$$\mathrm{Var}(v(Y)) = \mathbb{E}\big[v(Y)^2\big] - \mathbb{E}\big[v(Y)\big]^2.$$

Applying Theorem 2.2 once more,

$$\begin{aligned}
\mathrm{Var}\big(\mathbb{E}[X \mid Y]\big) &= \mathbb{E}\big[\mathbb{E}[X \mid Y]^2\big] - \mathbb{E}\big[\mathbb{E}[X \mid Y]\big]^2 \\
&= \mathbb{E}\big[\mathbb{E}[X \mid Y]^2\big] - \mathbb{E}[X]^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}\big[\mathrm{Var}(X \mid Y)\big] + \mathrm{Var}\big(\mathbb{E}[X \mid Y]\big) &= \mathbb{E}[X^2] - \mathbb{E}\big[\mathbb{E}[X \mid Y]^2\big] + \mathbb{E}\big[\mathbb{E}[X \mid Y]^2\big] - \mathbb{E}[X]^2 \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
&= \mathrm{Var}(X).
\end{aligned}$$

Example 2.9. Suppose that $\{X_i\}_{i=1}^{\infty}$ is an iid sequence of rvs with common mean μ and common variance σ^2. Let N be a discrete, non-negative integer-valued rv that is independent of each X_i. Find the mean and variance of $T = \sum_{i=1}^{N} X_i$ (referred to as a *random sum*).

Solution: By the law of total expectation,

$$\mathbb{E}[T] = \mathbb{E}\big[\mathbb{E}[T \mid N]\big].$$

Note that

$$\begin{aligned}
\mathbb{E}[T \mid N = n] &= \mathbb{E}\left[\sum_{i=1}^{N} X_i \,\middle|\, N = n\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} X_i \,\middle|\, N = n\right] \\
&= \sum_{i=1}^{n} \mathbb{E}[X_i \mid N = n] \\
&= \sum_{i=1}^{n} \mathbb{E}[X_i] && \text{since } N \text{ is independent of } \{X_i\}_{i=1}^{\infty} \\
&= n\mu.
\end{aligned}$$

Thus,

$$\mathbb{E}[T \mid N] = \mathbb{E}[T \mid N = n]\big|_{n=N} = N\mu,$$

and so $\mathbb{E}[T] = \mathbb{E}[N\mu] = \mu\,\mathbb{E}[N]$. To calculate $\operatorname{Var}(T)$, we employ Theorem 2.3 to obtain

$$\begin{aligned}
\operatorname{Var}(T) &= \mathbb{E}\big[\operatorname{Var}(T \mid N)\big] + \operatorname{Var}\big(\mathbb{E}[T \mid N]\big) \\
&= \mathbb{E}\big[\operatorname{Var}(T \mid N)\big] + \operatorname{Var}(N\mu) \\
&= \mathbb{E}\big[\operatorname{Var}(T \mid N)\big] + \mu^2 \operatorname{Var}(N).
\end{aligned}$$

Now,

$$\begin{aligned}
\operatorname{Var}(T \mid N = n) &= \operatorname{Var}\left(\sum_{i=1}^{N} X_i \,\middle|\, N = n\right) \\
&= \operatorname{Var}\left(\sum_{i=1}^{n} X_i \,\middle|\, N = n\right) \\
&= \operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) && \text{since } N \text{ is independent of } \{X_i\}_{i=1}^{\infty} \\
&= \sum_{i=1}^{n} \operatorname{Var}(X_i) \\
&= n\sigma^2.
\end{aligned}$$

Thus, $\operatorname{Var}(T \mid N) = \operatorname{Var}(T \mid N = n)\big|_{N=n} = N\sigma^2$. Finally,

$$\begin{aligned}
\operatorname{Var}(T) &= \mathbb{E}[N\sigma^2] + \mu^2 \operatorname{Var}(N) \\
&= \sigma^2\,\mathbb{E}[N] + \mu^2 \operatorname{Var}(N).
\end{aligned}$$

WEEK 3
22nd to 29th September

2.3 Computing Probabilities by Conditioning

For any two rvs, recall that

$$\mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X \mid Y]\big] = \begin{cases} \sum_y \mathbb{E}[X \mid Y = y] p_Y(y) & \text{, if } Y \text{ is discrete,} \\ \int_{-\infty}^{\infty} \mathbb{E}[X \mid Y = y] f_Y(y) \, \mathrm{d}y & \text{, if } Y \text{ is continuous.} \end{cases} \tag{2.1}$$

Now suppose that A represents some event of interest, and we wish to determine $\mathbb{P}(A)$. Define an indicator rv X such that

$$X = \begin{cases} 0 & \text{, if event } A^c \text{ occurs,} \\ 1 & \text{, if event } A \text{ occurs.} \end{cases}$$

Clearly, $\mathbb{P}(X = 1) = \mathbb{P}(A)$ and $\mathbb{P}(X = 0) = 1 - \mathbb{P}(A)$, so that $X \sim \text{BERN}\big(\mathbb{P}(A)\big)$. Thus,

$$\begin{aligned} \mathbb{E}[X \mid Y = y] &= \sum_x x \, \mathbb{P}(X = x \mid Y = y) \\ &= 0 \, \mathbb{P}(X = 0 \mid Y = y) + 1 \, \mathbb{P}(X = 1 \mid Y = y) \\ &= \mathbb{P}(X = 1 \mid Y = y) \\ &= \mathbb{P}(A \mid Y = y). \end{aligned}$$

Therefore, (2.1) becomes

$$\mathbb{P}(A) = \begin{cases} \sum_y \mathbb{P}(A \mid Y = y) p_Y(y) & \text{, if } Y \text{ is discrete,} \\ \int_{-\infty}^{\infty} \mathbb{P}(A \mid Y = y) f_Y(y) \, \mathrm{d}y & \text{, if } Y \text{ is continuous,} \end{cases} \tag{2.2}$$

which are analogues of the law of total probability. In other words, the expectation formula (2.1) can also be used to calculate probabilities of interest as indicated by (2.2).

Example 2.10. Suppose that X and Y are independent continuous rvs. Find an expression for $\mathbb{P}(X < Y)$.

Solution: With the event defined as $A = \{X < Y\}$, we apply (2.2) to get

$$\begin{aligned} \mathbb{P}(X < Y) &= \mathbb{P}(A) \\ &= \int_{-\infty}^{\infty} \mathbb{P}(A \mid Y = y) f_Y(y) \, \mathrm{d}y \\ &= \int_{-\infty}^{\infty} \mathbb{P}(X < Y \mid Y = y) f_Y(y) \, \mathrm{d}y \\ &= \int_{-\infty}^{\infty} \mathbb{P}(X < y \mid Y = y) f_Y(y) \, \mathrm{d}y \\ &= \int_{-\infty}^{\infty} \mathbb{P}(X < y) f_Y(y) \, \mathrm{d}y && \text{since } X \text{ and } Y \text{ are independent rvs} \\ &= \int_{-\infty}^{\infty} \mathbb{P}(X \leq y) f_Y(y) \, \mathrm{d}y && \text{since } X \text{ is a continuous rv} \\ &= \int_{-\infty}^{\infty} F_X(y) f_Y(y) \, \mathrm{d}y \end{aligned} \tag{2.3}$$

Remark: If, in addition, X and Y are identically distributed, then the pdf $f_Y(y)$ is equal to $f_X(y)$ and the result of Example 2.10 simplifies to become

$$\begin{aligned}
\mathbb{P}(X < Y) &= \int_{-\infty}^{\infty} F_X(y) f_X(y)\,\mathrm{d}y \\
&= \int_0^1 u\,\mathrm{d}u \qquad\qquad \text{where } u = F_X(y) \implies \frac{\mathrm{d}u}{\mathrm{d}y} = f_X(y) \implies \mathrm{d}u = f_X(y)\,\mathrm{d}y \\
&= \left[\frac{u^2}{2}\right]_{u=0}^{u=1} \\
&= \frac{1}{2},
\end{aligned}$$

as one would expect.

Example 2.11. Suppose that $X \sim \text{EXP}(\lambda_1)$ and $Y \sim \text{EXP}(\lambda_2)$ are independent exponential rvs. Show that

$$\mathbb{P}(X < Y) = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

Solution: Since X and Y are both exponential rvs, it immediately follows that

$$\begin{aligned}
f_Y(y) &= \lambda_2 e^{-\lambda_2 y},\ y > 0, \\
F_X(y) &= \int_0^y \lambda_1 e^{-\lambda_1 x}\,\mathrm{d}x \\
&= \lambda_1 \left[-\frac{1}{\lambda_1} e^{-\lambda_1 x}\right]_{x=0}^{x=y} \\
&= 1 - e^{-\lambda_1 y},\ y \geq 0.
\end{aligned}$$

Therefore, (2.3) becomes

$$\begin{aligned}
\mathbb{P}(X < Y) &= \int_0^{\infty} (1 - e^{-\lambda_1 y})\lambda_2 e^{-\lambda_2 y}\,\mathrm{d}y \\
&= \int_0^{\infty} \lambda_2 e^{-\lambda_2 y}\,\mathrm{d}y - \lambda_2 \int_0^{\infty} e^{-(\lambda_1+\lambda_2)y}\,\mathrm{d}y \\
&= 1 - \frac{\lambda_2}{(\lambda_1 + \lambda_2)} \int_0^{\infty} (\lambda_1 + \lambda_2) e^{-(\lambda_1+\lambda_2)y}\,\mathrm{d}y \\
&= 1 - \frac{\lambda_2}{\lambda_1 + \lambda_2} \\
&= \frac{\lambda_1}{\lambda_1 + \lambda_2}.
\end{aligned}$$

Remark: As a matter of interest, this particular result will be featured quite prominently in Chapter 4.

Example 2.12. Suppose W, X, and Y are independent continuous rvs on $(0, \infty)$. If $Z = X \mid (X < Y)$, then show that $(W, X) \mid (W < X < Y)$ and $(W, Z) \mid (W < Z)$ are identically distributed.

Solution: Let us first consider the joint conditional cdf of $(W, X) \mid (W < X < Y)$:

$$
\begin{aligned}
G(w,x) &= \mathbb{P}(W \le w, X \le x \mid W < X < Y) \\
&= \frac{\mathbb{P}(W \le w, X \le x, W < X < Y)}{\mathbb{P}(W < X < Y)} \\
&= \frac{\mathbb{P}(W \le w, X \le x, W < X, X < Y)}{\mathbb{P}(W < X, X < Y)}, \ w, x \ge 0.
\end{aligned}
$$

Conditioning on the rv X and noting that W, X, and Y are independent rvs, it follows that

$$
\begin{aligned}
\mathbb{P}(W < X, X < Y) &= \int_0^\infty \mathbb{P}(W < X, X < Y \mid X = s) f_X(s)\,\mathrm{d}s \\
&= \int_0^\infty \mathbb{P}(W < s, Y > s \mid X = s) f_X(s)\,\mathrm{d}s \\
&= \int_0^\infty \mathbb{P}(W < s, Y > s) f_X(s)\,\mathrm{d}s \\
&= \int_0^\infty \mathbb{P}(W < s)\,\mathbb{P}(Y > s) f_X(s)\,\mathrm{d}s \qquad (2.4)
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{P}(W \le w, X \le x, W < X, X < Y) &= \int_0^\infty \mathbb{P}(W \le w, X \le x, W < X, X < Y \mid X = s) f_X(s)\,\mathrm{d}s \\
&= \int_0^\infty \mathbb{P}(W \le w, s \le x, W < s, Y > s \mid X = s) f_X(s)\,\mathrm{d}s \\
&= \int_0^\infty \mathbb{P}(W \le w, s \le x, W < s, Y > s) f_X(s)\,\mathrm{d}s \\
&= \int_0^x \mathbb{P}(W \le w, W < s, Y > s) f_X(s)\,\mathrm{d}s \\
&= \int_0^x \mathbb{P}\big(W \le \min\{w, s\}, Y > s\big) f_X(s)\,\mathrm{d}s \\
&= \int_0^x \mathbb{P}\big(W \le \min\{w, s\}\big)\,\mathbb{P}(Y > s) f_X(s)\,\mathrm{d}s \qquad (2.5)
\end{aligned}
$$

Next, consider the conditional rv $Z = X \mid (X < Y)$.

$$
\begin{aligned}
\mathbb{P}(Z \le z) &= \mathbb{P}(X \le z \mid X < Y) \\
&= \frac{\mathbb{P}(X \le z, X < Y)}{\mathbb{P}(X < Y)} \\
&= \frac{\int_0^\infty \mathbb{P}(X \le z, X < Y \mid X = s) f_X(s)\,\mathrm{d}s}{\mathbb{P}(X < Y)} \\
&= \frac{\int_0^\infty \mathbb{P}(s \le z, s < Y \mid X = s) f_X(s)\,\mathrm{d}s}{\mathbb{P}(X < Y)} \\
&= \frac{\int_0^\infty \mathbb{P}(s \le z, s < Y) f_X(s)\,\mathrm{d}s}{\mathbb{P}(X < Y)} \\
&= \frac{\int_0^z \mathbb{P}(Y > s) f_X(s)\,\mathrm{d}s}{\mathbb{P}(X < Y)}
\end{aligned}
$$

and so the pdf of Z is given by

$$\begin{aligned} h_Z(z) &= \frac{\mathrm{d}}{\mathrm{d}z}\,\mathbb{P}(Z \le z) \\ &= \frac{\frac{\mathrm{d}}{\mathrm{d}z}\int_0^z \mathbb{P}(Y > s) f_X(s)\,\mathrm{d}s}{\mathbb{P}(X < Y)} \\ &= \frac{\mathbb{P}(Y > z) f_X(z)}{\mathbb{P}(X < Y)},\ z > 0. \end{aligned}$$

Now, the joint conditional cdf of $(W, Z) \mid (W < Z)$ is given by

$$\mathbb{P}(W \le w, Z \le z \mid W < Z) = \frac{\mathbb{P}(W \le w, Z \le z, W < Z)}{\mathbb{P}(W < Z)},\ w, z \ge 0$$

Due to the independence of W with X and Y,

$$\begin{aligned} \mathbb{P}(W < Z) &= \int_0^\infty \mathbb{P}(W < Z \mid Z = s) h_Z(s)\,\mathrm{d}s \\ &= \int_0^\infty \mathbb{P}(W < z \mid Z = s) h_Z(s)\,\mathrm{d}s \\ &= \int_0^\infty \mathbb{P}(W < s) h_Z(s)\,\mathrm{d}s \\ &= \int_0^\infty \mathbb{P}(W < s) \frac{\mathbb{P}(Y > s) f_X(s)}{\mathbb{P}(X < Y)}\,\mathrm{d}s \\ &= \frac{\mathbb{P}(W < X, X < Y)}{\mathbb{P}(X < Y)} && \text{from (2.4)} \end{aligned}$$

Next,

$$\begin{aligned} \mathbb{P}(W \le w, Z \le z, W < Z) &= \int_0^\infty \mathbb{P}(W \le w, Z \le z, W < Z \mid Z = s) h_Z(s)\,\mathrm{d}s \\ &= \int_0^\infty \mathbb{P}(W \le w, s \le z, W < s) h_Z(s)\,\mathrm{d}s \\ &= \int_0^z \mathbb{P}(W \le w, W < s) h_Z(s)\,\mathrm{d}s \\ &= \int_0^z \mathbb{P}(W \le \min\{w, s\}) \frac{\mathbb{P}(Y > s) f_X(s)}{\mathbb{P}(X < Y)}\,\mathrm{d}s \\ &= \mathbb{P}(W \le w, X \le z, W < X, X < Y) && \text{from (2.5)} \end{aligned}$$

Therefore, we ultimately obtain:

$$\mathbb{P}(W \le w, Z \le z, W < Z) = \frac{\mathbb{P}(W \le w, X \le z, W < X, X < Y)}{\mathbb{P}(W < X, X < Y)} = G(w, z),\ w, z \ge 0.$$

This implies that

$$(W, X) \mid (W < X < Y) \sim (W, Z) \mid (W < Z).$$

Remark: It can likewise be shown that if $V = X \mid (W < X)$, then $(X, Y) \mid (W < X < Y)$ and $(V, Y) \mid (V < Y)$ are identically distributed (left as an upcoming exercise).

2.4 Some Further Extensions

If you consider our treatment of the conditional expectation $\mathbb{E}[X \mid Y = y]$, then one detail you should notice is that this kind of expectation behaves *exactly* the same as the regular (i.e., unconditional) expectation *except* that all pmfs/pdfs used now are conditional on the event $Y = y$. In this sense, conditional expectations essentially satisfy all the properties of regular expectation. Thus, for an arbitrary real-valued function $g(\,\cdot\,)$, a corresponding analogue of

$$\mathbb{E}\big[g(X)\big] = \begin{cases} \sum_w \mathbb{E}\big[g(X) \,\big|\, W = w\big] p_W(w) & \text{, if } W \text{ is discrete,} \\ \int_{-\infty}^{\infty} \mathbb{E}\big[g(X) \,\big|\, W = w\big] f_W(w)\,\mathrm{d}w & \text{, if } W \text{ is continuous,} \end{cases}$$

would be

$$\mathbb{E}\big[g(X) \,\big|\, Y = y\big] = \begin{cases} \sum_w \mathbb{E}\big[g(X) \,\big|\, W = w, Y = y\big] p_{W|Y}(w \mid y) & \text{, if } W \text{ is discrete,} \\ \int_{-\infty}^{\infty} \mathbb{E}\big[g(X) \,\big|\, W = w, Y = y\big] f_{W|Y}(w \mid y)\,\mathrm{d}w & \text{, if } W \text{ is continuous.} \end{cases}$$

We remark that the above relation makes sense, since if we assume (without loss of generality) that X and Y are discrete rvs, then we obtain (in the case when W is discrete too):

$$\begin{aligned} \sum_w \mathbb{E}\big[g(X) \,\big|\, W = w, Y = y\big] p_{W|Y}(w \mid y) &= \sum_w \sum_x g(x) p_{X|WY}(x \mid w, y) p_{W|Y}(w \mid y) \\ &= \sum_w \sum_x g(x) \frac{p_{XWY}(x, w, y)}{p_{WY}(w, y)} \frac{p_{WY}(w, y)}{p_Y(y)} \\ &= \sum_x \frac{g(x)}{p_Y(y)} \sum_w p_{XWY}(x, w, y) \\ &= \sum_x g(x) \frac{p_{XY}(x, y)}{p_Y(y)} \\ &= \sum_x g(x) p_{X|Y}(x, y) \\ &= \mathbb{E}\big[g(X) \,\big|\, Y = y\big]. \end{aligned}$$

Similarly, if one introduces an event of interest A and defines

$$g(X) = \begin{cases} 0 & \text{, if event } A^c \text{ occurs,} \\ 1 & \text{, if event } A \text{ occurs,} \end{cases}$$

then we obtain

$$\mathbb{E}\big[A \,\big|\, Y = y\big] = \begin{cases} \sum_w \mathbb{E}\big[A \,\big|\, W = w, Y = y\big] p_{W|Y}(w \mid y) & \text{, if } W \text{ is discrete,} \\ \int_{-\infty}^{\infty} \mathbb{E}\big[A \,\big|\, W = w, Y = y\big] f_{W|Y}(w \mid y)\,\mathrm{d}w & \text{, if } W \text{ is continuous.} \end{cases}$$

Furthermore, if we now define

$$\mathbb{E}\big[g(X) \,\big|\, W, Y\big] = \mathbb{E}\big[g(X) \,\big|\, W = w, Y = y\big]\big|_{w=W, y=Y},$$

then the law of total expectation extends to become

$$\mathbb{E}\big[g(X)\big] = \mathbb{E}\big[\mathbb{E}[g(X) \mid Y]\big] = \mathbb{E}\Big[\mathbb{E}\big[\mathbb{E}[g(X) \mid W, Y] \,\big|\, Y\big]\Big].$$

Example 2.13. Consider an experiment in which independent trials, each having success probability $p \in (0, 1)$, are performed until k consecutive successes are achieved where $k \in \mathbb{Z}^+$. Determine the expected number of trials needed to achieve k consecutive successes.

Solution: Let N_k represent the number of trials needed to get k consecutive successes. We wish to determine $\mathbb{E}[N_k]$. For $k = 1$, note that $N_1 \sim \text{GEO}_t(p)$, therefore $\mathbb{E}[N_k] = \frac{1}{p}$. For arbitrary $k \geq 2$, let us consider conditioning on the outcome of the first trial, represented by W, such that

$$W = \begin{cases} 0 & \text{, if first trial is a failure,} \\ 1 & \text{, if first trial is a success.} \end{cases}$$

Thus,

$$\begin{aligned} \mathbb{E}[N_k] &= \mathbb{E}\big[\mathbb{E}[N_k \mid W]\big] \\ &= \mathbb{P}(W = 0)\,\mathbb{E}[N_k \mid W = 0] + \mathbb{P}(W = 1)\,\mathbb{E}[N_k \mid W = 1] \\ &= (1-p)\,\mathbb{E}[N_k \mid W = 0] + p\,\mathbb{E}[N_k \mid W = 1] \end{aligned}$$

Now, it is clear $N_k \mid (W = 0) \sim 1 + N_k$, but unfortunately we <u>do not</u> have a nice corresponding result for $N_k \mid (W = 1)$. It <u>does not</u> hold true that $N_k \mid (W = 0) \sim 1 + N_{k-1}$. What else can we try?
<u>Idea</u>: Let's try $\mathbb{E}[N_k] = \mathbb{E}\big[\mathbb{E}[N_k \mid N_{k-1}]\big]$, i.e., to get k in a row, we must first get $k-1$ in a row. Define

$$Y \mid (N_{k-1} = n) = \begin{cases} 0 & \text{, if } (n+1)^{\text{th}} \text{ trial is a failure,} \\ 1 & \text{, if } (n+1)^{\text{th}} \text{ trial is a success.} \end{cases}$$

By independence of the trials,

$$\begin{aligned} \mathbb{P}(Y = 0 \mid N_{k-1} = n) &= 1 - p, \\ \mathbb{P}(Y = 1 \mid N_{k-1} = n) &= p. \end{aligned}$$

As a result, we get:

$$\begin{aligned} \mathbb{E}[N_k \mid N_{k-1} = n] &= \sum_{y=0}^{1} \mathbb{E}[N_k \mid N_{k-1} = n, Y = y]\,\mathbb{P}(Y = y \mid N_{k-1} = n) \\ &= (1-p)\,\mathbb{E}[N_k \mid N_{k-1} = n, Y = 0] + p\,\mathbb{E}[N_k \mid N_{k-1} = n, Y = 1]. \end{aligned}$$

Note that $N_k \mid (N_{k-1} = n, Y = 0) \sim n + 1 + N_k$ (i.e., given that we know it took n trials to get $k-1$ consecutive successes, and then on the next trial we got a failure, what happens?). Also, $N_k \mid (N_{k-1} = n, Y = 1)$ is equal to $n+1$ with probability 1. Therefore,

$$\begin{aligned} \mathbb{E}[N_k \mid N_{k-1} = n] &= (1-p)(n + 1 + \mathbb{E}[N_k]) + p(n+1) \\ &= n + 1 + (1-p)\,\mathbb{E}[N_k]. \end{aligned}$$

Therefore,

$$\mathbb{E}[N_k \mid N_{k-1}] = \mathbb{E}[N_k \mid N_{k-1} = n]\big|_{n = N_{k-1}} = N_{k-1} + 1 + (1-p)\,\mathbb{E}[N_k].$$

Now, our whole idea was to apply $\mathbb{E}[N_k] = \mathbb{E}\big[\mathbb{E}[N_k \mid N_{k-1}]\big]$, and now we have the inner piece, so

$$\begin{aligned} \mathbb{E}[N_k] &= \mathbb{E}\big[N_{k-1} + 1 + (1-p)\,\mathbb{E}[N_k]\big] \\ &= \mathbb{E}[N_{k-1}] + 1 + (1-p)\,\mathbb{E}[N_k] \end{aligned}$$

Therefore,

$$\big(1 - (1-p)\big)\,\mathbb{E}[N_k] = 1 + \mathbb{E}[N_{k-1}] \implies \mathbb{E}[N_k] = \frac{1}{p} + \frac{\mathbb{E}[N_{k-1}]}{p}, \; k \geq 2,$$

which is a recursive equation for $\mathbb{E}[N_k]$. Take $k = 2$:

$$\mathbb{E}[N_2] = \frac{1}{p} + \frac{\mathbb{E}[N_1]}{p} = \frac{1}{p} + \frac{(1/p)}{p} = \frac{1}{p} + \frac{1}{p^2}.$$

Take $k = 3$:

$$\mathbb{E}[N_3] = \frac{1}{p} + \frac{\mathbb{E}[N_2]}{p} = \frac{1}{p} + \frac{1}{p^2} + \frac{1}{p^3}.$$

Take $k = 4$:

$$\mathbb{E}[N_4] = \frac{1}{p} + \frac{\mathbb{E}[N_3]}{p} = \frac{1}{p} + \frac{1}{p^2} + \frac{1}{p^3} + \frac{1}{p^4}.$$

Continuing inductively, we actually have

$$\mathbb{E}[N_k] = \frac{1}{p} + \frac{1}{p^2} + \cdots + \frac{1}{p^k},$$

which is a finite geometric series, therefore,

$$\mathbb{E}[N_k] = \frac{(1/p) - (1/p^{k+1})}{1 - (1/p)} = \frac{p^{-k} - 1}{1 - p},\ k \geq 2.$$

Actually, this holds true for $k \in \mathbb{Z}^+$ (try it).

Chapter 3

Discrete-time Markov Chains

WEEK 4
0929 to 6th October

3.1 Definitions and Basic Concepts

Stochastic Process

Definition: $\{X(t), t \in \mathcal{T}\}$ is called a *stochastic process* if $X(t)$ is a rv (or possibly a random vector) for any given $t \in \mathcal{T}$. $\mathcal{T}$ is referred to as the *index set* and is often interpreted in the context of time. As such, $X(t)$ is often called the *state of the process at time* t. We note that:

$$\text{Index set } \mathcal{T} \begin{cases} \text{can be a \underline{continuum} of values such as } \mathcal{T} = \{t : t \geq 0\}, \\ \text{can be a set of \underline{discrete} points such as } \mathcal{T} = \{t_0, t_1, t_2, \ldots\}. \end{cases}$$

Since there is a one-to-one correspondence between the sets $\mathcal{T} = \{t_0, t_1, t_2, \ldots\}$ and $\mathbb{N} = \{0, 1, 2, \ldots\}$, we will use $\mathcal{T} = \mathbb{N}$ as the general index set for a discrete-time stochastic process (unless otherwise stated). In other words, $\{X(n), n \in \mathbb{N}\}$ or $\{X_n, n \in \mathbb{N}\}$ will represent a general discrete-time stochastic process.

Discrete-time Stochastic Process

Some examples of a discrete-time stochastic process $\{X_n, n \in \mathbb{N}\}$ might include:

(1) X_n represents the outcome of the n^{th} toss of a die,

(2) X_n represents the price of a stock at the end of day n trading,

(3) X_n represents the maximum temperature in Waterloo during the n^{th} month,

(4) X_n represents the number of goals scored in game n by the varsity hockey team,

(5) X_n represents the number of STAT 333 students in class for the n^{th} lecture.

Discrete-time Markov Chain

Definition: A stochastic process $\{X_n, n \in \mathbb{N}\}$ is said to be a *discrete-time Markov chain* (DTMC) if the following two conditions hold true:

(1) For $n \in \mathbb{N}$, X_n is a <u>discrete</u> rv (i.e., the state space $\mathcal{S}$ of X_n is of discrete type).

(2) For $n \in \mathbb{N}$ and all states $x_0, x_1, \ldots, x_{n+1} \in \mathcal{S}$, the *Markov property* must hold:

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1, X_0 = x_0) = \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n).$$

In mathematical terms, this property states that the conditional distribution of any *future* state X_{n+1} given the past states $X_0, X_1, \ldots, X_{n-1}$ and the *present* state X_n is independent of the past states.

In a more informal way, the Markov property tells us, for a random process, that if we know the value taken by the process at a given time, we will not get any additional information about the future behaviour of the process by gathering more knowledge about the past.

Remarks:

(1) Unless otherwise stated, the state space $\mathcal{S}$ of a DTMC $\{X_n, n \in \mathbb{N}\}$ will be assumed to be $\mathbb{N}$.

(2) In general, the sequence of rvs $\{X_n\}_{n=0}^{\infty}$ are neither independent nor identically distributed.

(3) The Markov property does not require "full" information on the past to ensure independence. For example, consider the following conditional probability:

$$\begin{aligned}&\mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-2} = x_{n-2}, \ldots, X_1 = x_1, X_0 = x_0)\\ &= \frac{\mathbb{P}(X_{n+1} = x_{n+1}, X_n = x_n, X_{n-2} = x_{n-2}, \ldots, X_1 = x_1, X_0 = x_0)}{\mathbb{P}(X_n = x_n, X_{n-2} = x_{n-2}, \ldots, X_1 = x_1, X_0 = x_0)},\end{aligned}$$

which is "missing" the information for X_{n-1}.

However, note that:

$$\begin{aligned}&\mathbb{P}(X_{n+1} = x_{n+1}, X_n = x_n, X_{n-2} = x_{n-2}, \ldots, X_1 = x_1, X_0 = x_0)\\
&= \sum_{x_{n-1}=0}^{\infty} \mathbb{P}(X_{n+1} = x_{n+1}, X_n = x_n, X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots, X_1 = x_1, X_0 = x_0)\\
&= \sum_{x_{n-1}=0}^{\infty} \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots, X_1 = x_1, X_0 = x_0)\\
&\qquad \times \mathbb{P}(X_n = x_n, X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots, X_1 = x_1, X_0 = x_0)\\
&= \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n)\\
&\qquad \times \sum_{x_{n-1}=0}^{\infty} \mathbb{P}(X_n = x_n, X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots, X_1 = x_1, X_0 = x_0) \text{ Markov property}\\
&= \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n)\, \mathbb{P}(X_n = x_n, X_{n-2} = x_{n-2}, \ldots, X_1 = x_1, X_0 = x_0).\end{aligned}$$

We have:

$$\begin{aligned}&\mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-2} = x_{n-2}, \ldots, X_1 = x_1, X_0 = x_0)\\ &= \frac{\mathbb{P}(X_{n+1} = x_{n+1}, X_n = x_n, X_{n-2} = x_{n-2}, \ldots, X_1 = x_1, X_0 = x_0)}{\mathbb{P}(X_n = x_n, X_{n-2} = x_{n-2}, \ldots, X_1 = x_1, X_0 = x_0)}\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}(X_{n+1} = x_{n+1}, X_n = x_n, X_{n-2} = x_{n-2}, \ldots, X_1 = x_1, X_0 = x_0)\\ &= \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n)\, \mathbb{P}(X_n = x_n, X_{n-2} = x_{n-2}, \ldots, X_1 = x_1, X_0 = x_0).\end{aligned}$$

Substituting this latter expression into the numerator of the top equation yields

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-2} = x_{n-2}, \ldots, X_1 = x_1, X_0 = x_0) = \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n).$$

It is straightforward to extend the above result to any number of previous time points from $0, 1, \ldots, n-1$ with "missing" information. This is the essence of the Markov property.

One-step Transition Probability Matrix

Definition: For any pair of states i and j, the *transition probability* from state i at time n to state j at time $n+1$ is given by

$$P_{n,i,j} = \mathbb{P}(X_{n+1} = j \mid X_n = i), \ n \in \mathbb{N}.$$

Let P_n be the associated matrix containing all these transition probabilities, referred to as the *one-step transition probability matrix* (TPM) from time n to time $n+1$. It looks like

$$P_n = [P_{n,i,j}] = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ i \\ \vdots \end{array} \!\! \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \cdots & j & \cdots \end{array} \\ \begin{bmatrix} P_{n,0,0} & P_{n,0,1} & P_{n,0,2} & \cdots & P_{n,0,j} & \cdots \\ P_{n,1,0} & P_{n,1,1} & P_{n,1,2} & \cdots & P_{n,1,j} & \cdots \\ \vdots & \vdots & \vdots & & \vdots & \\ P_{n,i,0} & P_{n,i,1} & P_{n,i,2} & \cdots & P_{n,i,j} & \cdots \\ \vdots & \vdots & \vdots & & \vdots & \end{bmatrix} \end{array},$$

where, for convenience, the states of the DTMC are labelled along the margins of the matrix.
For each pair of states i and j, if $P_{n,i,j} = P_{i,j} \ \forall n \in \mathbb{N}$, then we say that the DTMC is *stationary* or *homogenous*. In this situation, the one-step TPM becomes:

$$P = [P_{i,j}] = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ i \\ \vdots \end{array} \!\! \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \cdots & j & \cdots \end{array} \\ \begin{bmatrix} P_{0,0} & P_{0,1} & P_{0,2} & \cdots & P_{0,j} & \cdots \\ P_{1,0} & P_{1,1} & P_{1,2} & \cdots & P_{1,j} & \cdots \\ \vdots & \vdots & \vdots & & \vdots & \\ P_{i,0} & P_{i,1} & P_{i,2} & \cdots & P_{i,j} & \cdots \\ \vdots & \vdots & \vdots & & \vdots & \end{bmatrix} \end{array}.$$

Remark: In STAT 333, we only consider stationary DTMCs. Moreover, from the construction of the TPM P, it is clear that $P_{i,j} \geq 0 \ \forall i, j \in \mathbb{N}$ and $\sum_{j=0}^{\infty} P_{i,j} = 1 \ \forall i \in \mathbb{N}$ (i.e., each row sum of P must be 1). Such a matrix whose elements are non-negative and whose row sums are equal to 1 is said to be *stochastic*.

Example 3.1. On a given day, the weather is either clear, overcast, or raining. If the weather is clear today, then it will be clear, overcast, or raining tomorrow with respective probabilities 0.6, 0.3, and 0.1. If the weather is overcast today, then it will be clear, overcast, or raining tomorrow with respective probabilities 0.2, 0.5, and 0.3. If the weather is raining today, then it will be clear, overcast, or raining tomorrow with respective probabilities 0.4, 0.2, and 0.4. Construct the underlying DTMC and determine its TPM.

Solution: Note that the weather tomorrow only depends on the weather today, implying that the Markov property holds true. Thus, letting X_n denote the state of the weather on the n^{th} day, $\{X_n, n \in \mathbb{N}\}$ is a three-state DTMC.
If we let state 0 correspond to clear weather, state 1 correspond to overcast, and state 2 correspond to raining, then the state space of the DTMC is $\mathcal{S} = \{0, 1, 2\}$ and its TPM is given by

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \!\! \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.5 & 0.3 \\ 0.4 & 0.2 & 0.4 \end{bmatrix} \end{array}.$$

n-step Transition Probability Matrix

Definition: For any pair of states i and j, the *n-step transition probability* is given by

$$P_{i,j}^{(n)} = \mathbb{P}(X_{m+n} = j \mid X_m = i), \ m, n \in \mathbb{N}.$$

Due to the stationary assumption, this quantity is actually independent of m (which is why we do not include m in its notation). Thus, $P_{i,j}^{(n)} = \mathbb{P}(X_n = j \mid X_0 = i)$, $n \in \mathbb{N}$. Furthermore, it is evident that

$$P_{i,j}^{(0)} = \mathbb{P}(X_0 = j \mid X_0 = i) = \begin{cases} 0, & \text{if } i \neq j, \\ 1, & \text{if } i = j. \end{cases}$$

Similarly, let $P^{(n)} = \left[P_{i,j}^{(n)}\right]$ represent the associated n-step TPM. Clearly, when $n = 1$, $P^{(1)} = P$. When $n = 0$, $P^{(0)} = I$, where I represents the identity matrix. Just as with the one-step TPM, it follows that the row sums of $P^{(n)}$ must equal 1 as well.

Chapman-Kolmogorov Equations

For $n \in \mathbb{Z}^+$, let us consider

$$\begin{aligned} P_{i,j}^{(n)} &= \mathbb{P}(X_n = j \mid X_0 = i) \\ &= \sum_{k=0}^{\infty} \mathbb{P}(X_n = j \mid X_{n-1} = k, X_0 = i)\, \mathbb{P}(X_{n-1} = k \mid X_0 = i) \\ &= \sum_{k=0}^{\infty} P_{i,k}^{(n-1)}\, \mathbb{P}(X_n = j \mid X_{n-1} = k, X_0 = i) \\ &= \sum_{k=0}^{\infty} P_{i,k}^{(n-1)}\, \mathbb{P}(X_n = j \mid X_{n-1} = k) \text{ due to the Markov property} \\ &= \sum_{k=0}^{\infty} P_{i,k}^{(n-1)} P_{k,j}. \end{aligned} \tag{3.1}$$

We have: $P_{i,j}^{(n)} = \sum_{k=0}^{\infty} P_{i,k}^{(n-1)} P_{k,j} \leftarrow$ (3.1).

Recall: If $A = [a_{i,j}]$, $B = [b_{i,j}]$, and $C = AB$ where $C = [c_{i,j}]$, then $c_{i,j} = \sum_k a_{i,k} b_{k,j}$.
As a result, note that (3.1) implies that $P^{(n)} = P^{(n-1)}P$, $n \in \mathbb{Z}^+$. More generally, $P_{i,j}^{(n)}$ can be expressed as

$$P_{i,j}^{(n)} = \sum_{k=0}^{\infty} P_{i,k}^{(n)} P_{k,j}^{(n-m)} \ \forall i, j \in \mathbb{N} \text{ and } 0 \leq m \leq n,$$

which are referred to as the *Chapman-Kolmogorov* equations for a DTMC. In matrix form, this translates to

$$P^{(n)} = P^{(m)} P^{(n-m)}, \ 0 \leq m \leq n.$$

Coming back to $P^{(n)} = P^{(n-1)}P$, $n \in \mathbb{Z}^+$, let us look at a few values of n:

$$\begin{aligned} &\text{Take } n = 2 \implies P^{(2)} = P^{(1)}P = PP = P^2, \\ &\text{Take } n = 3 \implies P^{(3)} = P^{(2)}P = P^2P = P^3, \\ &\text{Take } n = 4 \implies P^{(4)} = P^{(3)}P = P^3P = P^4. \end{aligned}$$

Clearly, we see that

$$P^{(n)} = P^n,$$

and so the n-step TPM is simply the one-step TPM multiplied by itself n times.

Marginal pmf of X_n

For $n \in \mathbb{N}$, let us now introduce a particular row vector, which we will denote as either

$$\underline{\alpha}_n = (\alpha_{n,0}, \alpha_{n,1}, \ldots, \alpha_{n,k}, \ldots),$$

or

$$\underline{\alpha}_n = \begin{bmatrix} \alpha_{n,0} & \alpha_{n,1} & \cdots & \alpha_{n,k} & \cdots \end{bmatrix},$$

where $\alpha_{n,k} = \mathbb{P}(X_n = k)\ \forall k \in \mathbb{N}$. In other words, $\alpha_{n,k}$ represents the marginal pmf of X_n, $n \in \mathbb{N}$. As a consequence, it follows that $\sum_{k=0}^{\infty} \alpha_{n,k} = 1\ \forall n \in \mathbb{N}$.
If we focus on the case when $n = 0$, then $\underline{\alpha}_0$ is referred to as the *initial probability row vector* of the DTMC, or simply the *initial conditions* of the DTMC.
For $n \in \mathbb{Z}^+$, note that

$$\begin{aligned}
\alpha_{n,k} &= \mathbb{P}(X_n = k) \\
&= \sum_{i=0}^{\infty} \mathbb{P}(X_n = k \mid X_m = i)\, \mathbb{P}(X_m = i) \text{ where } 0 \leq m \leq n \\
&= \sum_{i=0}^{\infty} \alpha_{m,i}\, \mathbb{P}(X_{n-m} = k \mid X_0 = i) \text{ due to the stationary assumption} \\
&= \sum_{i=0}^{\infty} \alpha_{m,i} P_{i,k}^{(n-m)}.
\end{aligned}$$

In matrix form, the above relation implies that

$$\underline{\alpha}_n = \underline{\alpha}_m P^{(n-m)} = \underline{\alpha}_m P^{n-m},\ 0 \leq m \leq n,$$

which subsequently leads to

$$\underline{\alpha}_n = \underline{\alpha}_0 P^{(n)} = \underline{\alpha}_0 P^n,\ n \in \mathbb{N}.$$

Probabilities of Interest

Having knowledge of the initial conditions and the one-step transition probabilities, one can readily calculate various probabilities of possible interest such as

$$\begin{aligned}
&\mathbb{P}(X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1, X_0 = x_0) \\
&= \mathbb{P}(X_0 = x_0)\, \mathbb{P}(X_1 = x_1 \mid X_0 = x_0)\, \mathbb{P}(X_2 = x_2 \mid X_1 = x_1, X_0 = x_0) \times \cdots \\
&\quad \times \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots, X_0 = x_0) \\
&= \mathbb{P}(X_0 = x_0)\, \mathbb{P}(X_1 = x_1 \mid X_0 = x_0)\, \mathbb{P}(X_2 = x_2 \mid X_1 = x_1) \cdots \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}) \\
&= \alpha_{0,x_0} P_{x_0,x_1} P_{x_1,x_2} \cdots P_{x_{n-1},x_n},
\end{aligned}$$

where the second last equality follows due to the Markov property.

Similarly,

$$\begin{aligned}
&\mathbb{P}(X_{n+m} = x_{n+m}, X_{n+m-1} = x_{n+m-1}, \dots, X_{n+1} = x_{n+1} \mid X_n = x_n) \\
&= \frac{\mathbb{P}(X_{n+m} = x_{n+m}, X_{n+m-1} = x_{n+m-1}, \dots, X_{n+1} = x_{n+1}, X_n = x_n)}{\mathbb{P}(X_n = x_n)} \\
&= \frac{\mathbb{P}(X_n=x_n)\,\mathbb{P}(X_{n+1}=x_{n+1} \mid X_n=x_n)\cdots \mathbb{P}(X_{n+m}=x_{n+m} \mid X_{n+m-1}=x_{n+m-1}, X_{n+m-2}=x_{n+m-2},\dots,X_n=x_n)}{\mathbb{P}(X_n=x_n)} \\
&= P_{x_n,x_{n+1}} P_{x_{n+1},x_{n+2}} \cdots P_{x_{n+m-1},x_{n+m}}.
\end{aligned}$$

The key observation here is that the DTMC is *completely characterized* by its one-step TPM P and the initial conditions $\underline{\alpha}_0$.

Example 3.2. A particle moves along the states $\{0, 1, 2\}$ according to a DTMC whose TPM is given by

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0 & 0.6 & 0.4 \\ 0.5 & 0 & 0.5 \end{bmatrix} \end{array}.$$

Let X_n denote the position of the particle after the n^{th} move. Suppose that the particle is equally likely to start in any of the three positions.

(a) Calculate $\mathbb{P}(X_3 = 1 \mid X_0 = 0)$.

Solution: We wish to determine $P^{(3)}_{0,1}$. To get this, we proceed to calculate $P^{(3)} = P^3$. First,

$$P^{(2)} = P^2 = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0 & 0.6 & 0.4 \\ 0.5 & 0 & 0.5 \end{bmatrix} \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0 & 0.6 & 0.4 \\ 0.5 & 0 & 0.5 \end{bmatrix} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 0.54 & 0.26 & 0.2 \\ 0.2 & 0.36 & 0.44 \\ 0.6 & 0.1 & 0.3 \end{bmatrix} \end{array}.$$

Then,

$$P^{(3)} = P^{(2)}P = \begin{bmatrix} 0.54 & 0.26 & 0.2 \\ 0.2 & 0.36 & 0.44 \\ 0.6 & 0.1 & 0.3 \end{bmatrix} \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0 & 0.6 & 0.4 \\ 0.5 & 0 & 0.5 \end{bmatrix} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 0.478 & 0.264 & 0.258 \\ 0.36 & 0.256 & 0.384 \\ 0.57 & 0.18 & 0.25 \end{bmatrix} \end{array}.$$

Thus, $\mathbb{P}(X_3 = 1 \mid X_0 = 0) = P^{(3)}_{0,1} = 0.264$.

(b) Calculate $\mathbb{P}(X_4 = 2)$.

Solution: We wish to calculate $\alpha_{4,2} = \mathbb{P}(X_4 = 2)$. Note that

$$\begin{aligned}
\underline{\alpha}_4 &= \begin{bmatrix} \alpha_{4,0} & \alpha_{4,1} & \alpha_{4,2} \end{bmatrix} \\
&= \underline{\alpha}_0 P^{(4)} \\
&= \underline{\alpha}_0 P^{(3)} P \\
&= \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 0.478 & 0.264 & 0.258 \\ 0.36 & 0.256 & 0.384 \\ 0.57 & 0.18 & 0.25 \end{bmatrix} \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0 & 0.6 & 0.4 \\ 0.5 & 0 & 0.5 \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 0.4636 & 0.254 & 0.2824 \\ 0.444 & 0.2256 & 0.3304 \\ 0.524 & 0.222 & 0.254 \end{bmatrix} \\
&= \begin{bmatrix} 0.4772 & 0.233867 & 0.288933 \end{bmatrix}.
\end{aligned}$$

Thus, $\mathbb{P}(X_4 = 2) = 0.288933 \approx 0.289$.

(c) Calculate $\mathbb{P}(X_6 = 0, X_4 = 2)$.

Solution: We have

$$\begin{aligned}
\mathbb{P}(X_6 = 0, X_4 = 2) &= \mathbb{P}(X_4 = 2)\,\mathbb{P}(X_6 = 0 \mid X_4 = 2) \\
&= \alpha_{4,2} P_{2,0}^{(2)} \\
&= (0.288933)(0.6) \\
&= 0.17336 \\
&\approx 0.173.
\end{aligned}$$

(d) Calculate $\mathbb{P}(X_9 = 0, X_7 = 2 \mid X_5 = 1, X_2 = 0)$.

Solution: We have

$$\begin{aligned}
&\mathbb{P}(X_9 = 0, X_7 = 2 \mid X_5 = 1, X_2 = 0) \\
&= \mathbb{P}(X_7 = 2 \mid X_5 = 1, X_2 = 0)\,\mathbb{P}(X_9 = 0 \mid X_7 = 2, X_5 = 1, X_2 = 0) \\
&= \mathbb{P}(X_7 = 2 \mid X_5 = 1)\,\mathbb{P}(X_9 = 0 \mid X_7 = 2) && \text{Markov property} \\
&= P_{1,2}^{(2)} P_{2,0}^{(2)} \\
&= (0.44)(0.6) \\
&= 0.264.
\end{aligned}$$

Accessibility and Communication

With these basic results in place, we next consider the classification of states in a DTMC.

Definition: State j is *accessible* from state i (denoted by $i \to j$) if $\exists n \in \mathbb{N}$ such that $P_{i,j}^{(n)} > 0$. If states i and j are accessible from each other, then the states i and j *communicate* (denoted by $i \leftrightarrow j$). In other words, $i \leftrightarrow j$ iff $\exists m, n \in \mathbb{N}$ such that $P_{i,j}^{(n)} > 0$ and $P_{j,i}^{(m)} > 0$.

In terms of accessibility, note that the size of the components of P do not matter. All that matters is which are positive and which are 0. In particular, if state j is not accessible from state i, then $P_{i,j}^{(n)} = 0 \ \forall n \in \mathbb{N}$ and

$$\begin{aligned}
&\mathbb{P}(\text{DTMC ever visits state } j \mid X_0 = i) \\
&= \mathbb{P}\big(\cup_{n=0}^{\infty} \{X_n = j\} \bigm| X_0 = i\big) \\
&\le \sum_{n=0}^{\infty} \mathbb{P}(X_n = j \mid X_0 = i) \text{ due to Boole's inequality (see Exercise 1.1.1)} \\
&= \sum_{n=0}^{\infty} P_{i,j}^{(n)} \\
&= 0,
\end{aligned}$$

implying that $\mathbb{P}(\text{DTMC ever visits state } j \mid X_0 = i) = 0$.

Equivalence Relation

The concept of communication forms what is known as an equivalence relation, satisfying the following criteria:

(i) **Reflexivity**: $i \leftrightarrow i$.

Clearly true since $P_{i,i}^{(0)} = 1 > 0$.

(ii) **Symmetry**: $i \leftrightarrow j \implies j \leftrightarrow i$.

This is obviously true by definition.

(iii) **Transitivity**: $i \leftrightarrow j$ and $j \leftrightarrow k \implies i \leftrightarrow k$.

To see this holds formally, we know that $\exists n \in \mathbb{N}$ such that $P_{i,j}^{(n)} > 0$. Also, $\exists m \in \mathbb{N}$ such that $P_{j,k}^{(m)} > 0$. Using the Chapman-Kolmogorov equations, we have that

$$P_{i,k}^{(n+m)} = \sum_{\ell=0}^{\infty} P_{i,\ell}^{(n)} P_{\ell,k}^{(m)} \geq P_{i,j}^{(n)} P_{j,k}^{(m)} > 0.$$

Therefore, $P_{i,k}^{(n+m)} > 0$, implying that $i \to k$. Using precisely the same logic, it is straightforward to show that $k \to i$. Thus, by definition, $i \leftrightarrow k$

Communication Classes

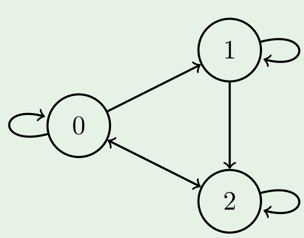
The fact that communication forms an equivalence relation allows us to *partition* all the states of a DTMC into various communication classes, so that within each class, all states communicate. However, if states i and j belong to *different classes*, then $i \leftrightarrow j$ is not true (i.e., at most one of $i \to j$ or $j \to i$ can be true).

Definition: A DTMC that has only one communication class is said to be *irreducible*. On the other hand, a DTMC is called *reducible* if there are two or more communication classes.

Example 3.2. (*continued*) What are the communication classes of the DTMC?

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \!\!\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0 & 0.6 & 0.4 \\ 0.5 & 0 & 0.5 \end{bmatrix} \end{array}.$$

Solution: To answer questions of this nature, it is often useful to draw a state transition diagram.



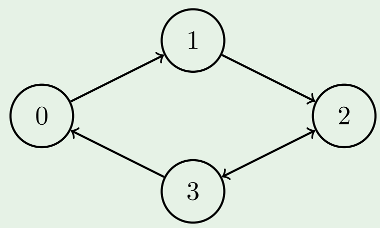
It is clear from this diagram that there is only one class for this DTMC, namely $\{0, 1, 2\}$. Therefore, this DTMC is irreducible.

Example 3.3. Consider a DTMC with TPM

$$P = \begin{array}{c c} & \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \end{array} & \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0.5 & 0 & 0.5 & 0 \end{bmatrix} \end{array}.$$

What are the communication classes of this DTMC?

Solution: State Transition Diagram



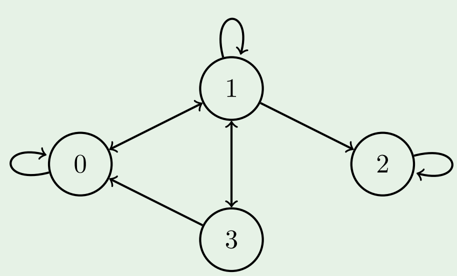
The above diagram indicates that there is only one communication class for this DTMC, namely $\{0, 1, 2, 3\}$. Therefore, this DTMC is irreducible.

Example 3.4. Consider a DTMC with TPM

$$P = \begin{array}{c c} & \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \end{array} & \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \\ 0 & 0 & 1 & 0 \\ \frac{3}{4} & \frac{1}{4} & 0 & 0 \end{bmatrix} \end{array}.$$

What are the communication classes of this DTMC?

Solution: State Transition Diagram



The above diagram indicates that the communication classes are $\{0, 1, 3\}$ and $\{2\}$. Thus, this DTMC is reducible.

Periodicity

Definition: The *period* of state i is given by $d(i) = \gcd\{n \in \mathbb{Z}^+ : P_{i,i}^{(n)} > 0\}$, where $\gcd\{\,\cdot\,\}$ denotes the greatest common divisor of a set of positive integers.

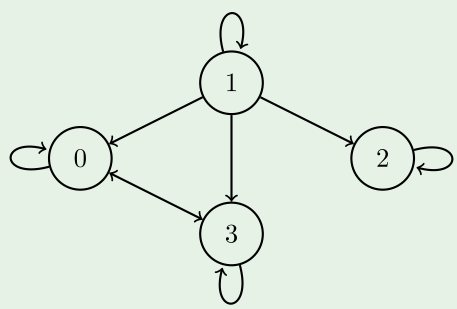
Remark: If $d(i) = 1$, then state i is said to be *aperiodic*. In fact, a DTMC is said to be *aperiodic* if $d(i) = 1 \ \forall i \in \mathbb{N}$. Furthermore, if $P_{i,i}^{(n)} = 0 \ \forall n \in \mathbb{Z}^+$, then we set $d(i) = \infty$.

Example 3.5. Consider a DTMC with TPM

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} \frac{1}{3} & 0 & 0 & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \\ 0 & 0 & 1 & 0 \\ \frac{3}{4} & 0 & 0 & \frac{1}{4} \end{bmatrix} \end{array}.$$

Determine the communication classes of this DTMC and the period of each state.

Solution: State Transition Diagram



Communications are $\{0, 3\}$, $\{1\}$, $\{2\}$. Next, we note that

$$d(0) = \gcd\{n \in \mathbb{Z}^+ : P_{0,0}^{(n)} > 0\} = \gcd\{1, 2, 3, \ldots\} = 1,$$

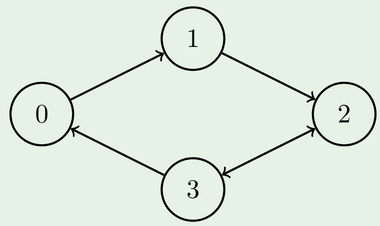
as a consequence of a fact that $P_{0,0} > 0$ (implying $P_{0,0}^{(n)} \geq (P_{0,0})^n > 0$). In fact, since every term on the main diagonal of P is positive, this same argument holds for every state. Thus, $d(1) = d(2) = d(3) = 1$. Note that this DTMC is aperiodic, but not irreducible.

Example 3.3. (*continued*) Recall that $\{0, 1, 2, 3\}$ is the only communication class for the DTMC with TPM

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0.5 & 0 & 0.5 & 0 \end{bmatrix} \end{array}.$$

Determine the period of each state.

Solution: State Transition Diagram



Examining the state transition diagram, the shortest amount of steps that the DTMC can take to arrive at state 0, after leaving state 0 is 4 (corresponding to the path $0 \to 1 \to 2 \to 3 \to 0$), so that $P_{0,0}^{(n)} > 0$ for $n = 4, 8, 12, \ldots$. We also note that since the DTMC can return to state 2 immediately after visiting state 3 (thereby revisiting state 3 again in a total of 2 steps), $P_{0,0}^{(n)} > 0 \ \forall n = 4 + 2k, k \in \mathbb{N}$. Thus,

$$d(0) = \gcd\{n \in \mathbb{Z}^+ : P_{0,0}^{(n)} > 0\} = \gcd\{4, 6, 8, 10, 12, 14, \ldots\} = 2.$$

Following a similar line of logic, we find that

$$\begin{aligned} d(1) &= \gcd\{n \in \mathbb{Z}^+ : P_{1,1}^{(n)} > 0\} = \gcd\{4, 6, 8, 10, \ldots\} = 2, \\ d(2) &= \gcd\{n \in \mathbb{Z}^+ : P_{2,2}^{(n)} > 0\} = \gcd\{2, 4, 6, 8, 10, \ldots\} = 2, \\ d(3) &= \gcd\{n \in \mathbb{Z}^+ : P_{3,3}^{(n)} > 0\} = \gcd\{2, 4, 6, 8, 10, \ldots\} = 2. \end{aligned}$$

Example 3.6. Consider the DTMC with TPM

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \end{array}.$$

Find the communication classes of this DTMC and determine the period of each state.

Solution: The communication classes are clearly $\{0, 1\}$ and $\{2, 3\}$. As in Example 3.5, the main diagonal components $P_{0,0}$ and $P_{1,1}$ are positive, and so $d(0) = d(1) = 1$. For states 2 and 3, the DTMC will continually alternate (with probability 1) between each other at every step, i.e., $2 \to 3 \to 2 \to 3 \to 2 \to \cdots$. Therefore, it is clear that

$$\begin{aligned} d(2) &= \gcd\{n \in \mathbb{Z}^+ : P_{2,2} > 0\} = \gcd\{2, 4, 6, 8, \ldots\} = 2, \\ d(3) &= \gcd\{n \in \mathbb{Z}^+ : P_{3,3} > 0\} = \gcd\{2, 4, 6, 8, \ldots\} = 2. \end{aligned}$$

The above examples illustrate some kinds of periodic behaviour that can be exhibited by DTMCs. However, we do observe that among the states within a given communication class, it seems as though the periodic behaviour is consistent. This is not a coincidence, as the next theorem indicates.

Theorem 3.1. If $i \leftrightarrow j$, then $d(i) = d(j)$.

Proof: Since the result is clearly true when $i = j$, let us assume that $i \neq j$. Since $i \leftrightarrow j$, we know by definition that $P_{i,j}^{(n)} > 0$ for some $n \in \mathbb{Z}^+$ and $P_{j,i}^{(m)} > 0$ for some $m \in \mathbb{Z}^+$. Moreover, since state i is accessible from state j and state j is accessible from state i, $\exists s \in \mathbb{Z}^+$ such that $P_{j,j}^{(s)} > 0$. Note that:

$$P_{i,i}^{(n+m)} \geq P_{i,j}^{(n)} P_{j,i}^{(m)} > 0$$

and

$$P_{i,i}^{(n+s+m)} \geq P_{i,j}^{(n)} P_{j,j}^{(s)} P_{j,i}^{(m)} > 0.$$

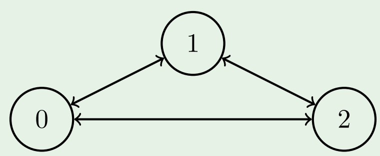
These two inequalities imply that $d(i)$ divides both $n+m$ and $n+s+m$. Therefore, it follows that $d(i)$ also divides their difference $(n+s+m)-(n+m)=s$. Since this holds true for any s which satisfies $P_{j,j}^{(s)} > 0$, it must be the case that $d(i)$ divides $d(j)$. Using the same line of logic, it is straightforward to show that $d(j)$ divides $d(i)$. Thus, $d(i) = d(j)$.

Example 3.7. Consider a DTMC with TPM

$$P = \begin{array}{c c} & \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \end{array} & \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \end{array}.$$

Find the communication classes of this DTMC and determine the period of each state.

Solution: State Transition Diagram



Clearly, there is one communication class, namely $\{0,1,2\}$. This is an irreducible DTMC. Note that

$$\begin{aligned} P_{0,0}^{(1)} &= 0, \\ P_{0,0}^{(2)} &\geq P_{0,1}P_{1,0} = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} > 0, \\ P_{0,0}^{(3)} &\geq P_{0,1}P_{1,2}P_{2,0} = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} > 0, \end{aligned}$$

implying that

$$d(0) = \gcd\{n \in \mathbb{Z}^+ : P_{0,0}^{(n)} > 0\} = \gcd\{2,3,\ldots\} = 1.$$

Thus, by Theorem 3.1, we know that

$$d(2) = d(1) = d(0) = 1.$$

Remark: As the previous example demonstrates, it is still possible to observe aperiodic behaviour even though the main diagonal components of P are all zero. More generally, if $d(i) = k$, then this does not necessarily imply that $P_{i,i}^{(k)} > 0$. Instead, it implies that if the DTMC is in state i at time 0, then it is impossible to observe the DTMC in state i at time $n \in \mathbb{Z}^+$ if n is not a multiple of k (i.e., $P_{i,i}^{(n)} = 0$ for such n).

3.2 Transience and Recurrence

First Visit Probability

We now wish to take a closer look at the likelihood of a DTMC beginning in some state of returning to that particular state. To that end, let us consider the probability that, starting from state i, the first visit of the DTMC to state j occurs at time $n \in \mathbb{Z}^+$, to be denoted by

$$f_{i,j}^{(n)} = \mathbb{P}(X_n = j, X_{n-1} \neq j, \ldots, X_2 \neq j, X_1 \neq j \mid X_0 = i) \; \forall i, j \in \mathbb{N}.$$

Clearly, we see that $f_{i,j}^{(1)} = P_{i,j}$.
For $n \geq 2$, however, the determination of $f_{i,j}^{(n)}$ becomes more complicated, and so we wish to construct a procedure which will enable us to compute $f_{i,j}^{(n)}$ for such n. To do so, we consider the quantity $P_{i,j}^{(n)}$, $n \in \mathbb{Z}^+$, and condition on the time that the first visit to state j is made:

$$\begin{aligned} P_{i,j}^{(n)} &= \mathbb{P}(X_n = j \mid X_0 = i) \\ &= \sum_{k=1}^{n} \mathbb{P}(X_n = j, \text{first visit to state } j \text{ occurs at time } k \mid X_0 = i) \\ &= \sum_{k=1}^{n} \mathbb{P}(X_n = j, X_k = j, X_{k-1} \neq j, \ldots, X_2 \neq j, X_1 \neq j \mid X_0 = i) \\ &= \sum_{k=1}^{n} \mathbb{P}(X_k = j, X_{k-1} \neq j, \ldots, X_2 \neq j, X_1 \neq j \mid X_0 = i) \, \mathbb{P}(X_n = j \mid X_k = j) \\ &= \sum_{k=1}^{n} f_{i,j}^{(k)} P_{j,j}^{(n-k)}, \end{aligned} \tag{3.2}$$

where we applied the Markov property in the second last equality.
We have: $P_{i,j}^{(n)} = \sum_{k=1}^{n} f_{i,j}^{(k)} P_{j,j}^{(n-k)} \leftarrow$ (3.2).
From (3.2), we can write

$$P_{i,j}^{(n)} = f_{i,j}^{(n)} P_{j,j}^{(0)} + \sum_{k=1}^{n-1} f_{i,j}^{(k)} P_{j,j}^{(n-k)} = f_{i,j}^{(n)} + \sum_{k=1}^{n-1} f_{i,j}^{(k)} P_{j,j}^{(n-k)},$$

implying that

$$f_{i,j}^{(n)} = P_{i,j}^{(n)} - \sum_{k=1}^{n-1} f_{i,j}^{(k)} P_{j,j}^{(n-k)}, \; n \in \mathbb{Z}^+. \tag{3.3}$$

For $n \geq 2$, note that (3.3) yields a recursive means to compute $f_{i,j}^{(n)}$.

Transience and Recurrence

Define the related quantity:

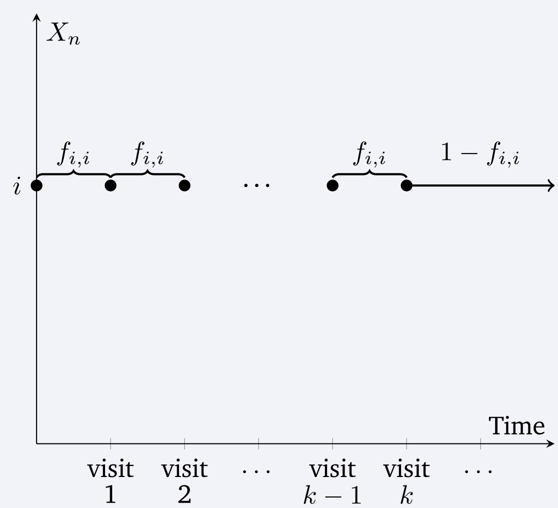
$$f_{i,j} = \mathbb{P}(\text{DTMC ever makes a future visit to state } j \mid X_0 = i).$$

Note that

$$
\begin{aligned}
&f_{i,j} \\
&= \sum_{k=1}^{\infty} \mathbb{P}(\text{DTMC ever makes a future visit to state } j, \text{DTMC visits state } j \text{ for } 1^{\text{th}} \text{ time at time } k \mid X_0 = i) \\
&= \sum_{k=1}^{\infty} \mathbb{P}(\text{DTMC visits state } j \text{ for } 1^{\text{th}} \text{ time at time } k \mid X_0 = i) \\
&= \sum_{k=1}^{\infty} f_{i,j}^{(k)} \leq 1.
\end{aligned}
$$

This leads to the following important concept in the study of Markov chains.

Definition: State i is said to be *transient* if $f_{i,i} < 1$. On the other hand, state i is said to be *recurrent* if $f_{i,i} = 1$.



In what follows, we proceed to look at alternative ways of characterizing the notions of transience and recurrence. As such, let us first define M_i to be a rv which counts the number of (future) times the DTMC visits state i (disregarding the possibility of starting in state i at time 0). If we assume that $f_{i,i} < 1$, then the Markov property and stationary assumption jointly imply that

$$\mathbb{P}(M_i = k \mid X_0 = i) = \left(\prod_{n=1}^{k} f_{i,i}\right)(1 - f_{i,i}) = f_{i,i}^k (1 - f_{i,i}), \; k = 0, 1, 2, \ldots, \tag{3.4}$$

as the DTMC will return to state i, k times with probability $f_{i,i}$, and then never return with probability $1 - f_{i,i}$.

We have: $\mathbb{P}(M_i = k \mid X_0 = i) = f_{i,i}^k (1 - f_{i,i}), \; k = 0, 1, 2, \ldots \leftarrow$ (3.4).

We recognize (3.4) as the pmf of a $\text{GEO}_f(1 - f_{i,i})$ rv, thereby implying that

$$\mathbb{E}[M_i \mid X_0 = i] = \frac{1 - (1 - f_{i,i})}{1 - f_{i,i}} = \frac{f_{i,i}}{1 - f_{i,i}} < \infty \text{ since } f_{i,i} < 1.$$

However, if $f_{i,i} = 1$, then $\mathbb{P}(M_i = \infty \mid X_0 = i) = 1$, immediately implying that $\mathbb{E}[M_i \mid X_0 = i] = \infty$.

Therefore, an equivalent way of viewing transience/recurrence is as follows:

$$\mathbb{E}[M_i \mid X_0 = i] \begin{cases} < \infty, & \text{iff state } i \text{ is transient,} \\ = \infty, & \text{iff state } i \text{ is recurrent.} \end{cases}$$

Following further on the notion of M_i, define a sequence of indicator rvs $\{A_n\}_{n=1}^{\infty}$ such that

$$A_n = \begin{cases} 0, & \text{if } X_n \neq i, \\ 1, & \text{if } X_n = i. \end{cases}$$

With this definition, note that $M_i = \sum_{n=1}^{\infty} A_n$. Now,

$$\begin{aligned} \mathbb{E}[M_i \mid X_0 = i] &= \mathbb{E}\left[\sum_{n=1}^{\infty} A_n \,\middle|\, X_0 = i\right] \\ &= \sum_{n=1}^{\infty} \mathbb{E}[A_n \mid X_0 = i] \\ &= \sum_{n=1}^{\infty} \Big[0\,\mathbb{P}(A_n = 0 \mid X_0 = i) + 1\,\mathbb{P}(A_n = 1 \mid X_0 = i)\Big] \\ &= \sum_{n=1}^{\infty} \mathbb{P}(X_n = i \mid X_0 = i) \\ &= \sum_{n=1}^{\infty} P_{i,i}^{(n)}. \end{aligned}$$

We have: $\mathbb{E}[M_i \mid X_0 = i] = \sum_{n=1}^{\infty} P_{i,i}^{(n)}$.
Therefore, yet another equivalent way of characterizing transience/recurrence is as follows:

$$\sum_{n=1}^{\infty} P_{i,i}^{(n)} \begin{cases} < \infty, & \text{iff state } i \text{ is transient,} \\ = \infty, & \text{iff state } i \text{ is recurrent.} \end{cases}$$

Remark: A simple way of viewing these concepts is as follows: a recurrent state will be visited *infinitely often*, whereas a transient state will only be visited *finitely often*.

As was the case concerning the periodicity of states within the same communication class, the next theorem indicates that transience/recurrence is *also* a class property.

Theorem 3.2. If $i \leftrightarrow j$ and state i is recurrent, then state j is recurrent.

Proof: Since $i \leftrightarrow j$, $\exists m, n \in \mathbb{N}$ such that $P_{i,j}^{(m)} > 0$ and $P_{j,i}^{(n)} > 0$. Also, since state i is recurrent we know that $\sum_{\ell=1}^{\infty} P_{i,i}^{(\ell)} = \infty$. Suppose that $s \in \mathbb{Z}^+$. Note that

$$P_{j,j}^{(n+s+m)} \geq P_{j,i}^{(n)} P_{i,i}^{(s)} P_{i,j}^{(m)}.$$

Look at

$$\begin{aligned}
\sum_{k=1}^{\infty} P_{j,j}^{(k)} &\geq \sum_{k=n+m+1}^{\infty} P_{j,j}^{(k)} \\
&= \sum_{s=1}^{\infty} P_{j,j}^{(n+s+m)} && \text{let } s = k-n-m \\
&\geq \sum_{s=1}^{\infty} P_{j,i}^{(n)} P_{i,i}^{(s)} P_{i,j}^{(m)} \\
&= \underbrace{P_{j,i}^{(n)}}_{>0} \underbrace{P_{i,j}^{(m)}}_{>0} \underbrace{\sum_{s=1}^{\infty} P_{i,i}^{(s)}}_{=\infty} \\
&= \infty.
\end{aligned}$$

Therefore, $\sum_{k=1}^{\infty} P_{j,j}^{(k)} = \infty$ and state j is recurrent.

Remark: An obvious by-product of this theorem is that if $i \leftrightarrow j$ and state i is transient, then state j must also be transient.

The following theorem serves as a companion result to Theorem 3.2.

Theorem 3.3. If $i \leftrightarrow j$ and state i is recurrent, then

$$f_{i,j} = \mathbb{P}(\text{DTMC ever makes a future visit to state } j \mid X_0 = i) = 1.$$

Proof: Clearly, the result is true if $i = j$. Therefore, suppose that $i \neq j$. Since $i \leftrightarrow j$, the fact that state i is recurrent implies that state j is recurrent by Theorem 3.2, and $f_{j,j} = 1$. To prove that $f_{i,j} = 1$, suppose that $f_{i,j} < 1$ and try to get a contradiction. Since $i \leftrightarrow j$, $\exists n \in \mathbb{Z}^+$, such that $P_{j,i}^{(n)} > 0$. Let n_i be the smallest such n satisfying $P_{j,i}^{(n)} > 0$. Thus, each time the DTMC visits state j, there is the possibility of being in state i, n_i time units later (with probability $P_{j,i}^{(n_i)} > 0$). If we suppose that $f_{i,j} < 1$, then this implies that the probability of returning to state j after visiting state i in the future is not guaranteed (as $1 - f_{i,j} > 0$). Therefore, we have:

$$\begin{aligned}
1 - f_{j,j} &= \mathbb{P}(\text{DTMC } \underline{\text{never}} \text{ makes a future visit to state } j \mid X_0 = j) \\
&\geq \underbrace{P_{j,i}^{(n_i)}}_{>0} \underbrace{(1 - f_{i,j})}_{>0} \\
&> 0 \implies 1 - f_{j,j} > 0 \implies f_{j,j} < 1,
\end{aligned}$$

which is a contradiction since $f_{j,j} = 1$. Thus, it must hold true that $f_{i,j} = 1$ when state i is recurrent and $i \leftrightarrow j$.

Remark: Based on the above result, we know that, starting from any state of a recurrent class, a DTMC will visit each state of that class infinitely many times.

At this stage, a natural question to ask is "*What do the results that we have accumulated thus far tell us about the behaviour of states within the same communication class?*" The answer would be:

(i) these states communicate with each other,

(ii) these states all have the same period,

(iii) these states are all either recurrent or all transient.

In fact, in an irreducible DTMC, there is only one communication class and so all the states are either recurrent or transient. When the assumption that the DTMC has a finite number of states is included, we obtain the following important result.

Theorem 3.4. A finite-state DTMC has at least one recurrent state.

Proof: We wish to prove the existence of at least one recurrent state in a finite-state DTMC, or equivalently, that not all states can be transient. Suppose that $\{0, 1, \ldots, N\}$ represents the states of the DTMC where $N < \infty$. To prove that not all states can be transient, suppose they are all transient and try to get a contradiction. Now, for each $i = 0, 1, 2, \ldots N$, if state i is assumed to be transient, then we know that after a finite amount of time (denoted by T_i), state i will never be visited again. As a result, after a finite amount of time

$$T = \max\{T_0, T_1, \ldots, T_N\}$$

has gone by, none of these states will be visited ever again. However, the DTMC must be in some state after time T, but we have exhausted all states for the DTMC to be in. This is a contradiction. Thus, not all states can be transient in a finite-state DTMC.

Remarks:

(1) Looking at the above result, it is useful to think of it in the following way. There must be at least one recurrent state. After all, the DTMC has to spend its time somewhere, and if it visits each of its *finitely many states finitely many times*, then where else could it possibly go?

(2) As an immediate consequence of Theorem 3.4, an irreducible, finite-state DTMC must be recurrent (i.e., all states of the DTMC are recurrent).

Example 3.3. (*continued*) Recall that we considered the irreducible DTMC with TPM

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0.5 & 0 & 0.5 & 0 \end{bmatrix} \end{array}.$$

Determine whether each state is transient or recurrent.

Solution: Since this is a finite-state DTMC as well, each of states 0, 1, 2, and 3 is therefore recurrent.

Another interesting property concerning recurrence can also be deduced.

Theorem 3.5. If state i is recurrent and state i does not communicate with state j, then $P_{i,j}^{(k)} = 0 \ \forall k \in \mathbb{Z}^+$.

Proof: Let us assume that $P_{i,j}^{(k)} > 0$ for some $k \in \mathbb{Z}^+$. Let k_i be the smallest k that satisfies $P_{i,j}^{(k_i)} > 0$. Then, $P_{j,i}^{(n)}$ would be equal to $0 \ \forall n \in \mathbb{Z}^+$, since otherwise, states i and j would communicate. However, the DTMC, starting in state i, would be a positive probability of at least $P_{i,j}^{(k_i)}$ of never returning to state i (by the nature of how k_i was chosen). This contradicts the recurrence of state i. Hence, we must have $P_{i,j}^{(k)} = 0 \ \forall k \in \mathbb{Z}^+$.

Example 3.4. (*continued*) Recall our earlier DTMC with TPM

$$P = \begin{array}{c c} & \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \end{array} & \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \\ 0 & 0 & 1 & 0 \\ \frac{3}{4} & \frac{1}{4} & 0 & 0 \end{bmatrix} \end{array}.$$

Determine whether each state is transient or recurrent.

Solution: We previously found the communication classes for this DTMC were $\{0, 1, 3\}$ and $\{2\}$.

$$\sum_{n=1}^{\infty} P_{2,2}^{(n)} = \sum_{n=1}^{\infty} 1 = \infty \implies \text{state 2 is recurrent.}$$

Looking at the possible transitions that can take place among states 0, 1, and 3, we strongly suspect state 1 to be transient (since there is a positive probability of never returning to state 1 if a transition to state 2 occurs). To prove this formally, assume instead that state 1 is recurrent and try to get a contradiction. Assuming that state 1 is recurrent, note that state 1 does not communicate with state 2. By Theorem 3.5, we have $P_{1,2}$ must be equal to 0. But in fact, we have $P_{1,2} = 1/8 \neq 0$. This is a contradiction. Thus, state 1 must indeed be transient. Thus, state 1 must be transient, and so $\{0, 1, 3\}$ is a transient class.

Remark: As the previous example illustrates, the contrapositive of Theorem 3.5 also provides a test for transience, in that if $\exists k \in \mathbb{Z}^+$ such that $P_{i,j}^{(k)} > 0$ and states i and j do not communicate, then state i must be transient. Moreover, this result implies that once a process enters a recurrent class of states, it can never leave that class. For this reason, a recurrent class is often referred to as a *closed class*.

Example 3.8. Consider a DTMC with TPM

$$P = \begin{array}{c c} & \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \end{array} & \begin{bmatrix} \frac{1}{4} & 0 & \frac{3}{4} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{2}{3} \\ 0 & 1 & 0 & 0 \\ 0 & \frac{2}{5} & 0 & \frac{3}{5} \end{bmatrix} \end{array}.$$

Determine whether each state is transient or recurrent.

Solution: There are three communication classes, namely $\{0\}$, $\{1, 3\}$, and $\{2\}$.

$$\sum_{n=1}^{\infty} P_{0,0}^{(n)} = \sum_{n=1}^{\infty} \left(\frac{1}{4}\right)^n = \frac{1/4}{1 - 1/4} = \frac{1}{3} < \infty,$$

and hence state 0 is transient.

$$\sum_{n=1}^{\infty} P_{2,2}^{(n)} = \sum_{n=1}^{\infty} 0 = 0 < \infty,$$

and hence state 2 is transient.
On the other hand, concerning $\{1, 3\}$, we observe that

$$f_{1,1}^{(1)} = P_{1,1} = \frac{1}{3},$$

and

$$f_{1,1}^{(n)} = \frac{2}{3}\left(\frac{3}{5}\right)^{n-2}\left(\frac{2}{5}\right),\ n \geq 2.$$

Thus,

$$\begin{aligned} f_{1,1} &= \sum_{n=1}^{\infty} f_{1,1}^{(n)} \\ &= \frac{1}{3} + \sum_{n=2}^{\infty} \frac{2}{3}\left(\frac{3}{5}\right)^{n-2}\left(\frac{2}{5}\right) \\ &= \frac{1}{3} + \frac{2}{3}\left(\frac{2}{5}\right)\frac{1}{1-3/5} \\ &= \frac{1}{3} + \frac{2}{3}\left(\frac{2}{5}\right)\left(\frac{5}{2}\right) \\ &= 1. \end{aligned}$$

By definition, state 1 is recurrent, and hence the class $\{1, 3\}$ is recurrent.

Remark: Instead of showing that $f_{1,1} = 1$ in the previous example, we could have used an even simpler argument to conclude that $\{1, 3\}$ is a recurrent class. In particular, after establishing that $\{0\}$ and $\{2\}$ are transient classes, this DTMC has a finite number of states, and so $\{1, 3\}$ must be recurrent due to Theorem 3.4.

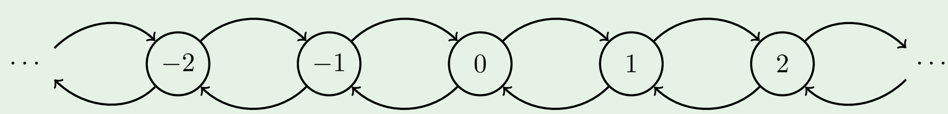
Random Walk

Example 3.9. Consider a DTMC $\{X_n, n \in \mathbb{N}\}$ whose state space $\mathcal{S}$ is the set of all integers (i.e., $\mathcal{S} = \mathbb{Z}$). Furthermore, suppose that the TPM for this DTMC satisfies

$$P_{i,i-1} = 1 - p \text{ and } P_{i,i+1} = p\ \forall i \in \mathbb{Z} \text{ where } 0 < p < 1.$$

In other words, from any state, either a jump up by one unit or a jump down by one unit takes place in the next transition. As such, X_n is expressible as $X_n = \sum_{k=0}^{n} Y_k$ where $\{Y_k\}_{k=0}^{\infty}$ is an independent sequence of rvs with $Y_0 = X_0$ and $\mathbb{P}(Y_k = -1) = 1 - p$ and $\mathbb{P}(Y_k = 1) = p$, $k \in \mathbb{Z}^+$. This DTMC is well-studied in the literature and is the basis for many applications in a variety of areas (particularly in finance). It is often referred to as the *Random Walk* or *Drunkard's Walk*. Characterize the behaviour of this DTMC in terms of its communication classes, periodicity, and transience/recurrence.

Solution: State Transition Diagram



Since $0 < p < 1$, all states clearly communicate with each other. This implies that $\{X_n, n \in \mathbb{N}\}$ is an irreducible DTMC. Hence, we can determine its periodicity (and likewise transience/recurrence) by analysing any state we wish. Let us select state 0. Starting from state 0, note that we cannot possibly be visited in an odd number of transitions, since we are guaranteed to have the number of up (down) jumps exceed the number of down (up) jumps. Thus,

$$P_{0,0}^{(1)} = P_{0,0}^{(3)} = \cdots = 0,$$

or equivalently

$$P_{0,0}^{(2n-1)} = 0\ \forall n \in \mathbb{Z}^+.$$

However, since it is clearly possible to return to state 0 in an <u>even</u> number of transitions, it immediately follows that

$$P_{0,0}^{(2n)} > 0 \; \forall n \in \mathbb{Z}^+.$$

Hence,

$$\begin{aligned} d(0) &= \gcd\{n \in \mathbb{Z}^+ : P_{0,0}^{(n)} > 0\} \\ &= \gcd\{2, 4, 6, \ldots\} \\ &= 2. \end{aligned}$$

Finally, to determine whether state 0 is transient <u>or</u> recurrent, let us consider

$$\begin{aligned} \sum_{n=1}^{\infty} P_{0,0}^{(n)} &= \underbrace{P_{0,0}^{(1)}}_{=0} + P_{0,0}^{(2)} + \underbrace{P_{0,0}^{(3)}}_{=0} + P_{0,0}^{(4)} + \cdots \\ &= \sum_{n=1}^{\infty} P_{0,0}^{(2n)} \\ &= \sum_{n=1}^{\infty} \binom{2n}{n} p^n (1-p)^n, \end{aligned}$$

where the last equality follows from the fact that in order for the DTMC to return to state 0 from state 0 in $2n$ steps, there must be an equal number (n) of up and down jumps, and $\binom{2n}{n}$ represents the number of ways these jumps could be arranged among the $2n$ steps.
<u>Recall</u>: (*Ratio Test for Series*). Suppose that $\sum_{n=1}^{\infty} a_n$ is a series of positive terms and $L = \lim_{n\to\infty} \frac{a_{n+1}}{a_n}$.

(i) If $L < 1$, the series converges.

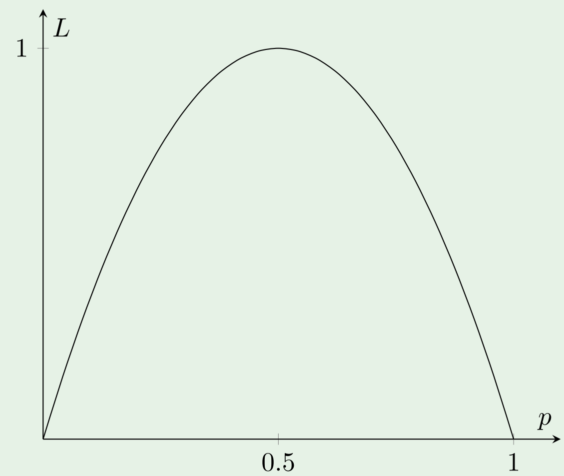
(ii) If $L > 1$, the series diverges.

(iii) If $L = 1$, the test is inconclusive.

In our case, $a_n = \binom{2n}{n} p^n (1-p)^n$.

$$\begin{aligned} L &= \lim_{n\to\infty} \frac{a_{n+1}}{a_n} \\ &= \lim_{n\to\infty} \frac{\frac{(2(n+1))!}{(n+1)!(n+1)!} p^{n+1} (1-p)^{n+1}}{\frac{(2n)!}{n!n!} p^n (1-p)^n} \\ &= \lim_{n\to\infty} \frac{(2n+2)(2n+1)}{(n+1)(n+1)} p(1-p) \\ &= 4p(1-p). \end{aligned}$$

A plot of $L = 4p(1-p)$ reveals the following shape:



Note that if $p \neq 1/2$, then $L < 1$. By the ratio test, this implies that

$$\sum_{n=1}^{\infty} \binom{2n}{n} p^n (1-p)^n < \infty,$$

and so state 0 is therefore transient. Thus, the entire DTMC is transient when $p \neq 1/2$. On the other hand, if $p = 1/2$, then $L = 1$, and so the ratio test is inconclusive. To determine what is happening when $p = 1/2$, we consider an alternative approach in which $p = 1/2$ and $p \neq 1/2$ can both be handled. First, recall that

$$f_{i,j} = \mathbb{P}(\text{DTMC ever makes a future visit to state } j \mid X_0 = i).$$

Let $q = 1 - p$. Condition on the state of the DTMC at time 1 to get

$$\begin{aligned} f_{0,0} &= \mathbb{P}(\text{DTMC ever makes a future visit to state } 0 \mid X_0 = 0) \\ &= \mathbb{P}(X_1 = -1 \mid X_0 = 0)\, \mathbb{P}(\text{DTMC ever makes a future visit to state } 0 \mid X_1 = -1, X_0 = 0) \\ &\quad + \mathbb{P}(X_1 = 1 \mid X_0 = 0)\, \mathbb{P}(\text{DTMC ever makes a future visit to state } 0 \mid X_1 = 1, X_0 = 0) \\ &= q f_{-1,0} + p f_{1,0} \end{aligned}$$

We have

$$\boxed{f_{0,0} = q f_{-1,0} + p f_{1,0}}. \tag{3.5}$$

If we let $\mathcal{F}_0$ represent the event that the DTMC ever makes a future visit to state 0, then

$$\mathcal{F}_0 = \bigcup_{i=1}^{\infty} \{X_i = 0\}.$$

So,

$$\begin{aligned}
f_{1,0} &= \mathbb{P}(\mathcal{F}_0 \mid X_0 = 1) \\
&= \mathbb{P}\big(\mathcal{F}_0 \cap \{X_1 = 0\} \mid X_0 = 1\big) + \mathbb{P}\big(\mathcal{F}_0 \cap \{X_1 = 2\} \mid X_0 = 1\big) \\
&= \mathbb{P}(\mathcal{F}_0 \mid X_1 = 0, X_0 = 1)\,\mathbb{P}(X_1 = 0 \mid X_0 = 1) + \mathbb{P}(\mathcal{F}_0 \mid X_1 = 2, X_0 = 1)\,\mathbb{P}(X_1 = 2 \mid X_0 = 1) \\
&= \mathbb{P}(X_1 = 0 \mid X_0 = 1) + \mathbb{P}(X_1 = 2 \mid X_0 = 1)\,\mathbb{P}(\mathcal{F}_0 \mid X_1 = 2, X_0 = 1) \\
&= q + p\,\mathbb{P}(\mathcal{F}_0 \mid X_1 = 2) \text{ due to the Markov property} \\
&= q + p\,\mathbb{P}\left(\bigcup_{i=2}^{\infty}\{X_i = 0\} \cup \{X_1 = 0\} \,\middle|\, X_1 = 2\right) \\
&= q + p\,\mathbb{P}\left(\bigcup_{i=2}^{\infty}\{X_i = 0\} \,\middle|\, X_1 = 2\right) \\
&= q + p\,\mathbb{P}(\mathcal{F}_0 \mid X_0 = 2) \text{ due to the stationary assumption} \\
&= q + pf_{2,0} \text{ by definition.}
\end{aligned}$$

Moreover, it follows that

$$\begin{aligned}
f_{1,0} &= q + pf_{2,0} \\
&= q + pf_{2,1}f_{1,0} \\
&= q + pf_{1,0}f_{1,0} \\
&= q + pf_{1,0}^2. \qquad (3.6)
\end{aligned}$$

Rewriting (3.6), we end up with

$$pf_{1,0}^2 - f_{1,0} + q = 0,$$

which is a quadratic equation in $f_{1,0}$. Applying the quadratic formula, we get that

$$\begin{aligned}
f_{1,0} &= \frac{1 \pm \sqrt{1 - 4pq}}{2p} \\
&= \frac{1 \pm \sqrt{(p+q)^2 - 4pq}}{2p} && \text{since } p + q = 1 \\
&= \frac{1 \pm \sqrt{p^2 + 2pq + q^2 - 4pq}}{2p} \\
&= \frac{1 \pm \sqrt{p^2 + 2pq + q^2}}{2p} \\
&= \frac{1 \pm \sqrt{(p-q)^2}}{2p} \\
&= \frac{1 \pm |p-q|}{2p}.
\end{aligned}$$

Let

$$r_1 = \frac{1 + |p-q|}{2p}, \text{ and } r_2 = \frac{1 - |p-q|}{2p}$$

denote the two roots, and let us consider the case we are mostly interested in, which is when $p = q$ (i.e., $p = 1/2$). In this case, r_1 and r_2 yield the same value, namely

$$\frac{1 \pm |1/2 - 1/2|}{2(1/2)} = 1,$$

and so it must be that $f_{1,0} = 1$. Similarly, it follows (via a symmetry argument) that $f_{-1,0} = 1$ when $p = 1/2$. Therefore, for $p = 1/2$, (3.5) simplifies to become

$$f_{0,0} = \frac{1}{2}(1) + \frac{1}{2}(1) = 1,$$

implying that state 0 is recurrent by definition. Thus, we conclude that the DTMC is recurrent only when $p = 1/2$. Out of mathematical interest, let us now attempt to determine $f_{0,0}$ for $p \neq q$. Consider the special case when $p < q$. Then $|p - q| = -(p - q)$ and the roots r_1 and r_2 simplify to become

$$r_1 = \frac{1 - (p - q)}{2p} = \frac{1 - p + q}{2p} = \frac{2q}{2p} = \frac{q}{p} > 1,$$

and

$$r_2 = \frac{1 + (p - q)}{2p} = \frac{1 - q + p}{2p} = \frac{2p}{2p} = 1.$$

Since $0 \leq f_{1,0} \leq 1$, the root r_1 must be inadmissible, thereby implying that r_2 is the correct root to use in this case. Moreover, by interchanging the up and down jump probabilities and applying the same symmetry argument above, it readily follows that

$$f_{-1,0} = 1 \text{ for } p > q.$$

Remark: For $p > q$, it can be shown that r_2 is again the admissible root for $f_{1,0}$, ultimately leading to $f_{1,0} = q/p$ (see Exercise 3.2.3). As an immediate consequence, we also have $f_{-1,0} = p/q$ for $p < q$. Combining our findings for $p < q$ and $p > q$, we have that

$$f_{1,0} = \frac{1 - |p - q|}{2p} \text{ and } f_{-1,0} = \frac{1 - |q - p|}{2q}.$$

Therefore, (3.5) gives rise to

$$\begin{aligned} f_{0,0} &= q\left(\frac{1 - |q - p|}{2q}\right) + p\left(\frac{1 - |p - q|}{2p}\right) \\ &= 1 - \frac{1}{2}|q - p| - \frac{1}{2}|p - q| \\ &= 1 - \frac{1}{2}\big(|q - p| + |p - q|\big). \end{aligned}$$

If $p > q$ (i.e., $2q < 1$), then

$$\begin{aligned} f_{0,0} &= 1 - \frac{1}{2}\big(-(q - p) + (p - q)\big) \\ &= 1 - \frac{1}{2}(2p - 2q) \\ &= 1 - p + q \\ &= 2q < 1. \end{aligned} \tag{3.7}$$

Therefore, state 0 is transient. On the other hand, if $p < q$ (i.e., $2p < 1$), it can be shown that $f_{0,0} = 2p < 1$ (see Exercise 3.2.4), implying also that state 0 is transient. Thus, if both cases are combined, we end up obtaining

$$f_{0,0} = 2\min\{p, q\} < 1 \text{ for } p \neq q \text{ (i.e., } p \neq 1/2).$$

However, this formula even gives the correct result when $p = q = 1/2$. In general,

$$f_{0,0} = 2\min\{p, 1 - p\},\ 0 < p < 1.$$

WEEK 6
20th to 27th October

3.3 Limiting Behaviour of DTMCs

The concepts of periodicity and transience/recurrence play an important role in characterizing the limiting behaviour of a DTMC. To demonstrate their influence, let us consider three examples with varying forms of limiting behaviour.

Example 3.10. Consider a DTMC with TPM

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \end{array}.$$

Determine if $\lim_{n\to\infty} P^{(n)}$ exists.

Solution: There are obviously two communication classes, namely $\{0, 2\}$ and $\{1\}$. Each class is recurrent with periods 2 and 1. Moreover, $n \in \mathbb{Z}^+$, note that

$$P^{(2n)} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{array}, \text{ and } P^{(2n-1)} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \end{array}.$$

As a result, $\lim_{n\to\infty} P^{(n)}$ does not exist since $P(n)$ alternates between these two matrices. For example, $\lim_{n\to\infty} P^{(n)}_{0,0}$ and $\lim_{n\to\infty} P^{(n)}_{0,2}$ do not exist. However, note that some limits do exist such as $\lim_{n\to\infty} P^{(n)}_{0,1} = 0$ and $\lim_{n\to\infty} P^{(n)}_{1,1} = 1$.

Example 3.11. Consider a DTMC with TPM

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{bmatrix} \end{array}.$$

Determine if $\lim_{n\to\infty} P^{(n)}$ exists.

Solution: There is clearly only one communication class, and so the DTMC is irreducible. It is also straightforward to verify the DTMC is aperiodic and recurrent. As we will soon learn, it can be shown that

$$\lim_{n\to\infty} P^{(n)} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} \frac{4}{11} & \frac{4}{11} & \frac{3}{11} \\ \frac{4}{11} & \frac{4}{11} & \frac{3}{11} \\ \frac{4}{11} & \frac{4}{11} & \frac{3}{11} \end{bmatrix} \end{array}.$$

Note that this limiting matrix has <u>identical rows</u>. This implies that $P^{(n)}_{i,j}$ converges to a value as $n \to \infty$ which is the same for all initial states i. In other words, there is a limiting probability that the process will be in state j as $n \to \infty$, and this probability is <u>independent</u> of the initial state i.

Example 3.12. Consider a DTMC with TPM

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \\ 0 & 0 & 1 \end{bmatrix} \end{array}.$$

Examine $\lim_{n\to\infty} P^{(n)}$, and explain why the limiting probability of being in a state can depend on the initial state of this DTMC.

Solution: Clearly, there are 3 communication classes: $\{0\}$, $\{1\}$, $\{2\}$. As all the main diagonal elements of P are positive, each state is aperiodic. Moreover, states 0 and 2 are recurrent states. In fact, for obvious reasons states 0 and 2 are examples of what are known as absorbing states (i.e., $P_{0,0} = P_{2,2} = 1$). Finally, since $P_{1,0} = 1/3 > 0$ and states 0 and 1 do not communicate, we can conclude that state 1 is transient by the contrapositive statement of Theorem 3.5. It can be shown that (see Exercise 3.3.3)

$$\lim_{n\to\infty} P^{(n)} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ \frac{2}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 \end{bmatrix} \end{array}.$$

In looking at the form of $P^{(n)}$ as $n \to \infty$, we would say that unlike the previous example, it can matter from which state one begins in this DTMC.

In the above example, note that the second column of the limiting matrix contains all zeros. Not surprisingly, this is indicative of transient behaviour, implying that one will never end up in state 1 in the long run. This property can be proven more formally in the next theorem.

Theorem 3.6. For any state i and transient state j of a DTMC, $\lim_{n\to\infty} P^{(n)}_{i,j} = 0$.

Proof: Recall that

$$f^{(n)}_{i,j} = \mathbb{P}(X_n = j, X_{n-1} \neq j, \ldots, X_2 \neq j, X_1 \neq j \mid X_0 = i),$$

and

$$f_{i,j} = \mathbb{P}(\text{DTMC ever makes a future visit to state } j \mid X_0 = i) = \sum_{n=1}^{\infty} f^{(n)}_{i,j}.$$

Recall (3.2) from the previous section:

$$P^{(n)}_{i,j} = \sum_{k=1}^{n} f^{(k)}_{i,j} P^{(n-k)}_{j,j}.$$

Using (3.2) as our starting point, note that

$$\begin{aligned}
\sum_{n=1}^{\infty} P_{i,j}^{(n)} &= \sum_{n=1}^{\infty}\sum_{k=1}^{n} f_{i,j}^{(k)} P_{j,j}^{(n-k)} \\
&= \sum_{k=1}^{\infty}\sum_{n=k}^{\infty} f_{i,j}^{(k)} P_{j,j}^{(n-k)} \\
&= \sum_{k=1}^{\infty} f_{i,j}^{(k)} \sum_{n=k}^{\infty} P_{j,j}^{(n-k)} \\
&= \sum_{k=1}^{\infty} f_{i,j}^{(k)} \sum_{\ell=0}^{\infty} P_{j,j}^{(\ell)} && \text{let } \ell = n-k \\
&= f_{i,j}\left(1 + \sum_{\ell=1}^{\infty} P_{j,j}^{(\ell)}\right) \\
&\le 1\left(1 + \sum_{\ell=1}^{\infty} P_{j,j}^{(\ell)}\right) \\
&= 1 + \underbrace{\sum_{\ell=1}^{\infty} P_{j,j}^{(\ell)}}_{<\infty} && \text{state } j \text{ is transient} \\
&< \infty.
\end{aligned}$$

By the n^{th} term test for infinite series, we must have

$$\lim_{n\to\infty} P_{i,j}^{(n)} = 0.$$

Mean Recurrent Time

As the previous three examples show, there is variation in the limiting behaviour of a DTMC. In particular, it is worthwhile to determine a set of conditions which ensure the "nice" limiting behaviour witnessed in Example 3.11. To ascertain when such conditions exist, we need to distinguish between two kinds of recurrence. Let

$$N_i = \min\{n \in \mathbb{Z}^+ : X_n = i\},$$

where state i is assumed to be recurrent. Clearly, the conditional rv $N_i \mid (X_0 = i)$ takes on values in $\mathbb{Z}^+$. Moreover, its conditional pmf is given by

$$\mathbb{P}(N_i = n \mid X_0 = i) = f_{i,i}^{(n)},\ n = 1, 2, 3, \ldots.$$

We observe that this is indeed a pmf since $\sum_{n=1}^{\infty} f_{i,i}^{(n)} = f_{i,i} = 1$, as state i is recurrent. This leads to the introduction of the following important quantity.

Definition: If state i is recurrent, then its *mean recurrent time* is given by

$$m_i = \mathbb{E}[N_i \mid X_0 = i] = \sum_{n=1}^{\infty} n f_{i,i}^{(n)}.$$

Positive and Null Recurrence

In words, m_i represents the average time it takes the DTMC to make successive visits to state i. Two notions of recurrence can now be defined based on the value of m_i.

Definition: Suppose that state i is recurrent. State i is said to be *positive recurrent* if $m_i < \infty$. On the other hand, state i is said to be *null recurrent* if $m_i = \infty$.

Remark: A fair question to ask is whether it is even possible for a discrete probability distribution on $\mathbb{Z}^+$ to have an **undefined** mean (i.e., a mean of ∞). To show that this is indeed possible, consider a rv X with pmf

$$\mathbb{P}(X = x) = \frac{1}{x(x+1)}, \ x = 1, 2, 3, \ldots.$$

Let us first confirm that this is indeed a pmf:

$$\begin{aligned}
\sum_{x=1}^{\infty} \frac{1}{x(x+1)} &= \lim_{n\to\infty} \sum_{x=1}^{n} \frac{1}{x(x+1)} \\
&= \lim_{n\to\infty} \sum_{x=1}^{n} \left(\frac{1}{x} - \frac{1}{x+1}\right) \\
&= \lim_{n\to\infty} \left\{\left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \cdots + \left(\frac{1}{n} - \frac{1}{n+1}\right)\right\} \\
&= \lim_{n\to\infty} \left(1 - \frac{1}{n+1}\right) \\
&= 1.
\end{aligned}$$

However,

$$\mathbb{E}[X] = \sum_{x=1}^{\infty} x \cdot \frac{1}{x(x+1)} = \sum_{x=1}^{\infty} \frac{1}{x+1} = \infty,$$

since the above harmonic series is known to diverge. In other words, a finite mean does not exist!

Some Facts About Positive and Null Recurrence:

1. If $i \leftrightarrow j$ and state i is positive recurrent, then state j is also positive recurrent. This means that positive recurrence is also a class property. An obvious by-product of this result is that null recurrence is a class property too.
2. In a finite-state DTMC, there can **never** be any null recurrent states.

Remarks:

(1) The above facts are provided without formal justification, as their proofs are rather lengthy and depend on material beyond the scope of STAT 333.

(2) Positive recurrent, aperiodic states are referred to as *ergodic* states.

Stationary Distribution

Before stating the main result governing the "nice" limiting behaviour demonstrated in Example 3.11, we introduce a special type of probability distribution.

Definition: A probability distribution $\{p_i\}_{i=0}^{\infty}$ is called a *stationary distribution* of a DTMC if $\{p_i\}_{i=0}^{\infty}$ satisfies the conditions $\sum_{i=0}^{\infty} p_i = 1$ and $p_j = \sum_{i=0}^{\infty} p_i P_{i,j}$ $\forall j \in \mathbb{N}$.

Remark: If we define the row vector

$$\underline{p} = (p_0, p_1, \ldots, p_j, \ldots),$$

then the above conditions can be represented in matrix form as

$$\underline{p}\underline{e}^\top = 1 \text{ and } \underline{p} = \underline{p}P,$$

where $\underline{e}^\top = (1, 1, \ldots, 1, \ldots)^\top$ denotes a column vector of ones (in general, the $^\top$ notation will be used to represent column vectors).

A logical question to ask is "*Why is such a distribution called stationary*?"
To answer this question, suppose that the initial conditions of the DTMC are given by $\underline{\alpha}_0 = \underline{p}$. As a result, we have that $\alpha_{0,j} = \mathbb{P}(X_0 = j) = p_j \ \forall j \in \mathbb{N}$. Now, for any $j \in \mathbb{N}$, note that

$$\alpha_{1,j} = \mathbb{P}(X_1 = j) = \sum_{i=1}^{\infty} \alpha_{0,i} P_{i,j} = \sum_{i=0}^{\infty} p_i P_{i,j} = p_j = \alpha_{0,j}.$$

The above equation indicates that X_1 has the same probability distribution as X_0 when $\underline{\alpha}_0 = \underline{p}$. More generally, it is straightforward to show (using mathematical induction) that each X_i, $i \in \mathbb{Z}^+$, is *identically distributed* to X_0, provided that $\underline{\alpha}_0 = \underline{p}$.
In other words, if a DTMC is started according to a stationary distribution, then the probability of being in a given state remains *unchanged* (i.e., stationary) over time.

Remarks:

(1) In some texts, the stationary probability distribution is sometimes called the *invariant probability distribution* or *steady-state probability distribution*.

(2) A known fact (which again we do not prove formally) is that a stationary distribution will not exist if all the states of the DTMC are either null recurrent or transient. On the other hand, an irreducible DTMC is positive recurrent iff a stationary distribution exists.

(3) Stationary distributions are not necessarily unique. This happens when a DTMC has more than one positive recurrent communication class. For instance, it is not difficult to verify that the DTMC in Example 3.10 has an **infinite** number of stationary distributions (left as an upcoming exercise).

The Basic Limit Theorem

We are now in position to state the fundamental limiting theorem for DTMCs, generally referred to as the Basic Limit Theorem (BLT).

Basic Limit Theorem: For an irreducible, recurrent, and aperiodic DTMC, $\lim_{n\to\infty} P_{i,j}^{(n)}$ exists and is independent of state i, satisfying

$$\lim_{n\to\infty} P_{i,j}^{(n)} = \pi_j = \frac{1}{m_j} \ \forall i, j \in \mathbb{N}.$$

If the DTMC also happens to be positive recurrent, then $\{\pi_j\}_{j=0}^{\infty}$ is the unique, positive solution to the system of linear equations defined by

$$\begin{cases} \pi_j = \sum_{i=0}^{\infty} \pi_i P_{i,j} \ \forall j \in \mathbb{N}, \\ \sum_{j=0}^{\infty} \pi_j = 1. \end{cases}$$

Remarks:

(1) A formal proof of the BLT is beyond the scope of STAT 333. However, it is not difficult to understand why $\{\pi_j\}_{j=0}^{\infty}$ (if they exist) satisfies the above system of linear equations. Specifically, recall the Chapman-Kolmogorov equations with $m = n - 1$, namely

$$P_{i,j}^{(n)} = \sum_{k=0}^{\infty} P_{i,k}^{(n-1)} P_{k,j} \ \forall i, j \in \mathbb{N}$$

Taking the limit as $n \to \infty$ of both sides of this equation and assuming that it is permissible to pass the limit through the summation sign, we obtain

$$\lim_{n\to\infty} P_{i,j}^{(n)} = \lim_{n\to\infty} \sum_{k=0}^{\infty} P_{i,k}^{(n-1)} P_{k,j}.$$

$$\pi_j = \sum_{k=0}^{\infty} \lim_{n\to\infty} P_{i,k}^{(n-1)} P_{k,j} = \sum_{k=0}^{\infty} \pi_k P_{k,j} \forall j \in \mathbb{N},$$

which is precisely the above system of equations.

(2) If we define the row vector of limiting probabilities

$$\underline{\pi} = (\pi_0, \pi_1, \ldots, \pi_j, \ldots),$$

then the above system of linear equations can be written succinctly in matrix form as:

$$\begin{cases} \underline{\pi} = \underline{\pi} P, \\ \underline{\pi}\underline{e}^{\top} = 1. \end{cases}$$

Therefore, if a DTMC is irreducible and ergodic, then the BLT states that the limiting probability distribution is the unique stationary distribution.

(3) When a DTMC has a finite number of states (i.e., suppose that the state space is $\{0, 1, \ldots, N\}$ where $N < \infty$), the BLT states that there are $N + 1$ linear equations to consider of the form

$$\pi_j = \sum_{i=0}^{N} \pi_i P_{i,j}, \ j = 0, 1, \ldots, N. \tag{3.8}$$

Along with the condition $\sum_{j=0}^{N} \pi_j = 1$, this leads to $N + 2$ equations in $N + 1$ unknowns, of which a unique solution must exist. In fact, the first $N + 1$ equations given by (3.8) are linearly dependent (implying that there is a redundancy), and so we can drop **any one** of the equations given by (3.8) and solve the remaining $N + 1$ equations to obtain a unique solution.

(4) If the conditions of the BLT are satisfied and state j happens to be null recurrent, then $\pi_j = 0$ which interestingly is similar to the limiting behaviour of a transient state.

Example 3.11. (*continued*) Recall that we previously considered a DTMC with TPM

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{bmatrix} \end{array}.$$

Find the limiting probabilities for this DTMC.

Solution: Clearly, the DTMC is irreducible, aperiodic, and positive recurrent. Therefore, the conditions of the BLT are satisfied and $\underline{\pi} = (\pi_0, \pi_1, \pi_2)$ is known to exist. To find $\underline{\pi}$, we need to solve the system of linear equations defined by

$$\underline{\pi} = \underline{\pi}P$$

$$(\pi_0, \pi_1, \pi_2) = (\pi_0, \pi_1, \pi_2)\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{bmatrix}, \text{ subject to } \underline{\pi}\underline{e}^\top = 1.$$

This leads to:

$$\begin{aligned} \pi_0 &= \tfrac{1}{2}\pi_0 + \tfrac{1}{2}\pi_1, \\ \pi_1 &= \tfrac{1}{2}\pi_0 + \tfrac{1}{4}\pi_1 + \tfrac{1}{3}\pi_2, \\ \pi_2 &= \tfrac{1}{4}\pi_1 + \tfrac{2}{3}\pi_2, \\ 1 &= \pi_0 + \pi_1 + \pi_2. \end{aligned}$$

We may disregard any one of the first three equations, and so we select the equation for π_1, as it involves the most terms.

$$\pi_0 = \tfrac{1}{2}\pi_0 + \tfrac{1}{2}\pi_1 \implies \tfrac{1}{2}\pi_0 = \tfrac{1}{2}\pi_1 \implies \pi_0 = \pi_1$$

$$\pi_2 = \tfrac{1}{4}\pi_1 + \tfrac{2}{3}\pi_2 \implies \tfrac{1}{3}\pi_2 = \tfrac{1}{4}\pi_1 \implies \pi_2 = \tfrac{3}{4}\pi_1$$

$$1 = \pi_0 + \pi_1 + \pi_2 \implies 1 = \pi_1 + \pi_1 + \tfrac{3}{4}\pi_1 \implies \pi_1 = \tfrac{4}{11}$$

It immediately follows that $\pi_0 = \frac{4}{11}$, and $\pi_2 = \frac{3}{4} \cdot \frac{4}{11} = \frac{3}{11}$. Thus,

$$\underline{\pi} = (\tfrac{4}{11}, \tfrac{4}{11}, \tfrac{3}{11}).$$

Recall from earlier that

$$\lim_{n\to\infty} P^{(n)} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} \frac{4}{11} & \frac{4}{11} & \frac{3}{11} \\ \frac{4}{11} & \frac{4}{11} & \frac{3}{11} \\ \frac{4}{11} & \frac{4}{11} & \frac{3}{11} \end{bmatrix} \end{array} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} \pi_0 & \pi_1 & \pi_2 \\ \pi_0 & \pi_1 & \pi_2 \\ \pi_0 & \pi_1 & \pi_2 \end{bmatrix} \end{array}.$$

Hence, we observe that each row of $P^{(n)}$ converges to $\underline{\pi}$ as $n \to \infty$.

Doubly Stochastic TPM

Recall that the TPM of a DTMC is stochastic, with all row sums of P being equal to 1. However, a TPM is said to be *doubly stochastic* if all column sums of P are also equal to 1 (i.e., $\sum_{i=0}^{\infty} P_{i,j} = 1 \ \forall j \in \mathbb{N}$). The following theorem provides an interesting result concerning the limiting behaviour of a class of such DTMCs.

Theorem 3.7. Suppose that a finite-state DTMC with state space $\mathcal{S} = \{0, 1, \ldots, N-1\}$ is irreducible and aperiodic. If the associated TPM is doubly stochastic, then the limiting probabilities $\{\pi_j\}_{j=0}^{N-1}$ exist and are given by

$$\pi_j = \frac{1}{N}, \ j = 0, 1, \ldots, N-1.$$

Proof: If the DTMC is irreducible and aperiodic, then a unique limiting probability distribution exists by the BLT, as each state must also be positive recurrent due to there being a finite number of states. To determine the limiting distribution, let us propose that

$$\pi_j = \frac{1}{N}, \ j = 0, 1, \ldots, N-1.$$

Clearly,

$$\sum_{j=0}^{N-1} \pi_j = \sum_{j=0}^{N-1} \frac{1}{N} = \frac{1}{N} \cdot N = 1.$$

Moreover, for $j = 0, 1, \ldots, N-1$, note that

$$\begin{aligned}
\sum_{k=0}^{N-1} \pi_k P_{k,j} &= \sum_{k=0}^{N-1} \frac{1}{N} P_{k,j} \\
&= \frac{1}{N} \underbrace{\sum_{k=0}^{N-1} P_{k,j}}_{=1} && \text{where } \sum_{k=0}^{N-1} P_{k,j} \text{ is sum of } j^{\text{th}} \text{ column of } P \\
&= \frac{1}{N} && \text{since the TPM is doubly stochastic} \\
&= \pi_j.
\end{aligned}$$

Thus, we conclude that $\pi_j = \frac{1}{N}, \ j = 0, 1, \ldots, N-1$ is the unique limiting distribution.

Alternative Interpretation

The primary interpretation of the limiting distribution of a DTMC is that after the process has been in operation for a "long" period of time, the probability of finding the process in state j is π_j (assuming the conditions of the BLT are met). In such situations, however, another interpretation exists for π_j. Specifically, π_j also represents the "long-run" mean fraction of time that the process spends in state j.
To see that this interpretation is valid, define the sequence of indicator random variables $\{A_k\}_{k=1}^{\infty}$ as follows:

$$A_k = \begin{cases} 0, & \text{if } X_k \neq j, \\ 1, & \text{if } X_k = j. \end{cases}$$

The fraction of time the DTMC visits state j during the time interval from 1 to n inclusive is therefore given by

$$\frac{1}{n} \sum_{k=1}^{n} A_k.$$

Looking at the quantity

$$\mathbb{E}\left[\frac{1}{n} \sum_{k=1}^{n} A_k \,\middle|\, X_0 = i \right],$$

which is interpreted as the mean fraction of time spent in state j during the time interval from 1 to n inclusive, given that the process starts in state i, note that

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} A_k \,\middle|\, X_0 = i\right] &= \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[A_k \mid X_0 = i] \\
&= \frac{1}{n}\sum_{k=1}^{n} \big(0 \cdot \mathbb{P}(A_k = 0 \mid X_0 = i) + 1 \cdot \mathbb{P}(A_k = 1 \mid X_0 = i)\big) \\
&= \frac{1}{n}\sum_{k=1}^{n} \mathbb{P}(X_k = j \mid X_0 = i) \\
&= \frac{1}{n}\sum_{k=1}^{n} P_{i,j}^{(k)}.
\end{aligned}$$

We have: $\mathbb{E}\big[\frac{1}{n}\sum_{k=1}^{n} A_k \mid X_0 = i\big] = \frac{1}{n}\sum_{k=1}^{n} P_{i,j}^{(k)}$.
Recall: If $\{a_n\}_{n=1}^{\infty}$ is a real sequence such that $a_n \to a$ as $n \to \infty$, then $\frac{1}{n}\sum_{k=1}^{n} a_k \to a$ as $n \to \infty$.
Thus, if the conditions of the BLT are satisfied, then $P_{i,j}^{(n)} \to \pi_j$ as $n \to \infty$. Therefore, applying the above result with $a_n = P_{i,j}^{(n)}$ and $a = \pi_j$, we obtain

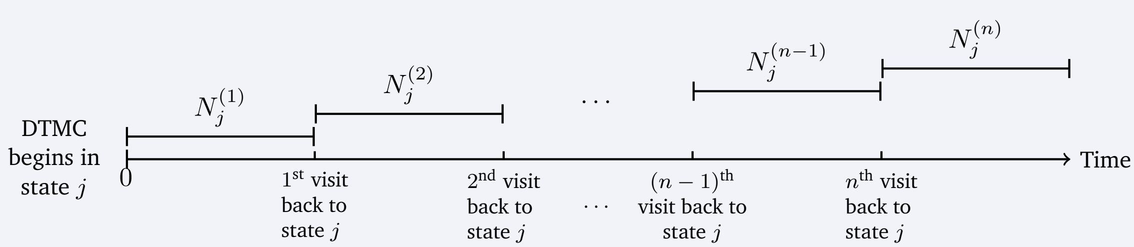
$$\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} A_k \,\middle|\, X_0 = i\right] \to \pi_j \text{ as } n \to \infty,$$

implying that the long-run mean fraction of time spent in state j is also equal to π_j.

Remark: If one begins in recurrent state j, we realize that the process spends one unit of time in state j every N_j time units. On average, this amounts to one unit of time in state j every $\mathbb{E}[N_j \mid X_0 = j] = m_j$ time units. If the conditions of the BLT are satisfied, then it makes sense intuitively that $\pi_j = 1/m_j$, as the BLT specifies. For a more formal justification in the positive recurrent case, let $\{N_j^{(n)}\}_{n=1}^{\infty}$ be a sequence of rvs where $N_j^{(n)}$ represents the number of transitions between the $(n-1)^{\text{th}}$ and n^{th} visits into state j, as illustrated in the diagram below. By the Markov property and the stationary assumption of the DTMC, $\{N_j^{(n)}\}_{n=1}^{\infty}$ is actually an iid sequence of rvs with common mean $m_j < \infty$. Therefore, the long-run fraction of time spend in state j can be viewed as

$$\pi_j = \lim_{n\to\infty} \frac{n}{\sum_{i=1}^{n} N_j^{(i)}} = \lim_{n\to\infty} \frac{1}{\frac{1}{n}\sum_{i=1}^{n} N_j^{(i)}} = \frac{1}{m_j},$$

where the last equality follows from the SLLN.



3.4 Two Interesting Applications

3.4.1 Interesting Application #1: The Galton-Watson Branching Process

- The famous *Galton-Watson Branching Process* was first introduced by Francis Galton in 1889 as a simple mathematical model for the propagation of family names.
- They were reinvented by Leo Szilard in the late 1930s as models for the proliferation of free neutrons in a nuclear fission reaction.
- Such mathematical models (and their generalizations) continue to play an important role in both the theory and applications of stochastic processes.

The Galton-Watson Branching Process

- In what follows, we assume that a population of individuals (which may represent people, organisms, free neutrons, etc.) evolves in discrete time. Specifically, we define:

$$\begin{aligned} X_0 &\equiv \text{population of the } 0^{\text{th}} \text{ (original) generation}, \\ X_1 &\equiv \text{population of the } 1^{\text{st}} \text{ generation}, \\ &\vdots \\ X_n &\equiv \text{population of the } n^{\text{th}} \text{ generation}, \\ &\vdots \end{aligned}$$

- We assume that each individual in a generation produces a random number (possibly 0) of individuals, called offspring, which go on and become part of the very next generation.
- In other words, it is always the offspring of a current generation which go on to form the next generation.
- We further assume that individuals produce offspring independently of all others according to the same probability distribution, namely

$$\alpha_m = \mathbb{P}(\text{an individual produces } m \text{ offspring}),\ m = 0, 1, 2, \ldots.$$

- In addition, for purposes we will see later, we assume that $\alpha_0 \in (0, 1)$ and $\alpha_0 + \alpha_1 < 1$.
- For $j \in \mathbb{N}$, let $Z_i^{(j)}$ be the number of offspring produced from individual i in the j^{th} generation.
- Due to the earlier independence assumptions, $\{Z_i^{(j)}\}_{i=1}^{\infty}$ is an iid sequence of rvs with $\alpha_m = \mathbb{P}(Z_i^{(j)} = m)$ for any $j \in \mathbb{N}$. Moreover, let $\mu = \mathbb{E}[Z_i^{(j)}]$ and $\sigma^2 = \mathrm{Var}(Z_i^{(j)})$ represent the (common) mean and variance, respectively, of the number of offspring produced by a single individual.
- Based on the above assumptions, the Galton-Watson process $\{X_n, n \in \mathbb{N}\}$ is actually a DTMC taking values in the state space $\mathcal{S} = \mathbb{N}$, since it follows that

$$X_n = \sum_{i=1}^{X_{n-1}} Z_i^{(n-1)}, \tag{3.9}$$

  implying that the Markov property and stationarity assumption are both satisfied.
- In this DTMC, we remark that $P_{0,0} = 1$, since state 0 is obviously an absorbing state.
- If we now consider state i, $i \in \mathbb{Z}^+$, then we can easily show that state i is transient as follows:
  - Clearly, states 0 and i do not communicate.

  - Note that $P_{i,0} = \alpha_0^i > 0$ (since $\alpha_0 > 0$).
  - By the contrapositive of Theorem 3.5, state i must therefore be transient.

- Thus, since state 0 is recurrent and states $1, 2, 3, \ldots$ are transient, the following conclusion can be drawn: *The population will either die out completely or its size will grow indefinitely (to ∞).*

- **We have**: $X_n = \sum_{i=1}^{X_{n-1}} Z_i^{(n-1)} \leftarrow$ (3.9).

- Since (3.9) infers that X_n is expressible as a *random sum*, we can apply the results of Example 2.9 to obtain

$$\mathbb{E}[X_n] = \mu\,\mathbb{E}[X_{n-1}]$$

  and

$$\mathrm{Var}(X_n) = \sigma^2\,\mathbb{E}[X_{n-1}] + \mu^2\,\mathrm{Var}(X_{n-1}).$$

- For convenience, let us henceforth assume that $X_0 = 1$ (with probability 1). As it is understood that $X_0 = 1$, for ease of notation, we will suppress writing the condition "$X_0 = 1$" in all expectations and probabilities which follow.

The Galton-Watson Branching Process: Mean

- **We have**: $\mathbb{E}[X_n] = \mu\,\mathbb{E}[X_{n-1}]$, $X_0 = 1$ with probability 1.

- Now, let us consider $\mathbb{E}[X_n]$ for several values of n:

$$\begin{aligned}
\text{Take } n = 1 &\implies \mathbb{E}[X_1] = \mu\,\mathbb{E}[X_0] = \mu, \\
\text{Take } n = 2 &\implies \mathbb{E}[X_2] = \mu\,\mathbb{E}[X_1] = \mu^2, \\
\text{Take } n = 3 &\implies \mathbb{E}[X_3] = \mu\,\mathbb{E}[X_2] = \mu^3.
\end{aligned}$$

  Based on the above findings, it is straightforward to deduce that $\mathbb{E}[X_n] = \mu^n$, $n \in \mathbb{N}$.

The Galton-Watson Branching Process: Variance

- **We have**: $\mathrm{Var}(X_n) = \sigma^2\,\mathbb{E}[X_{n-1}] + \mu^2\,\mathrm{Var}(X_{n-1})$, $X_0 = 1$ with probability 1. Similarly, we have

$$\begin{aligned}
\text{Take } n = 1 &\implies \mathrm{Var}(X_1) = \sigma^2\,\mathbb{E}[X_0] + \underbrace{\mu^2\,\mathrm{Var}(X_0)}_{0} = \sigma^2, \\
\text{Take } n = 2 &\implies \mathrm{Var}(X_2) = \sigma^2\,\mathbb{E}[X_1] + \mu^2\,\mathrm{Var}(X_1) = \sigma^2\mu + \sigma^2\mu^2, \\
\text{Take } n = 3 &\implies \mathrm{Var}(X_3) = \sigma^2\,\mathbb{E}[X_2] + \mu^2\,\mathrm{Var}(X_3) = \sigma^2\mu^2 + \sigma^2\mu^3 + \sigma^2\mu^4.
\end{aligned}$$

  Continuing inductively, we find that

$$\mathrm{Var}(X_n) = \sigma^2\mu^{n-1}\sum_{i=0}^{n-1}\mu^i,\ n \in \mathbb{N},$$

  which simplifies to give

$$\mathrm{Var}(X_n) = \begin{cases} n\sigma^2, & \text{if } \mu = 1, \\ \sigma^2\mu^{n-1}\left(\frac{1-\mu^n}{1-\mu}\right), & \text{if } \mu \neq 1. \end{cases}$$

The Galton-Watson Branching Process: Extinction Probability

- Under the assumption that $X_0 = 1$, let π_0 denote the limiting probability that the population dies out. In other words,
$$\pi_0 = \lim_{n\to\infty} \mathbb{P}(X_n = 0).$$

- Let us first consider the situation when $\mu < 1$, which is referred to as the *subcritical* case. Clearly, as $n \to \infty$, $\mathbb{E}[X_n] = \mu^n$ and $\mathrm{Var}(X_n) = \sigma^2\mu^{n-1}\left(\frac{1-\mu^n}{1-\mu}\right)$ both converge to 0. Therefore, we would expect that if $\mu < 1$, then $\pi_0 = 1$. To prove this formally, note that
$$\mu^n = \mathbb{E}[X_n] = \sum_{j=1}^{\infty} j\,\mathbb{P}(X_n = j) \geq \sum_{j=1}^{\infty} 1 \cdot \mathbb{P}(X_n = j) = \mathbb{P}(X_n \geq 1) = 1 - \mathbb{P}(X_n = 0).$$
This implies that $1 - \mu^n \leq \mathbb{P}(X_n = 0) \leq 1$. Taking the limit as $n \to \infty$ leads to:
$$\lim_{n\to\infty}(1-\mu^n) \leq \lim_{n\to\infty} \mathbb{P}(X_n = 0) \leq \lim_{n\to\infty} 1 \implies 1 \leq \pi_0 \leq 1,$$
or simply, $\pi_0 = 1$.

- On the other hand, suppose that $\mu \geq 1$. By conditioning on the number of offspring produced by the single individual present in the population at time 0, we obtain
$$\pi_0 = \mathbb{P}(\text{population dies out}) = \sum_{j=0}^{\infty} \mathbb{P}(\text{population dies out} \mid X_1 = j)\alpha_j.$$
However, with $X_1 = j$, the population will eventually die out iff each of the j families started by the members of the first generation eventually dies out.

  As each family is assumed to act independently, and since the probability that any particular family dies out is simply π_0, it follows that $\mathbb{P}(\text{population dies out} \mid X_1 = j) = \pi_0^j$ and our above equation becomes
$$\pi_0 = \sum_{j=0}^{\infty} \pi_0^j \alpha_j. \tag{3.10}$$

- **We have**: $\pi_0 = \sum_{j=0}^{\infty} \pi_0^j \alpha_j \leftarrow$ (3.10). Equivalently, $z = \pi_0$ satisfies the equation
$$z = \tilde{\alpha}(z), \tag{3.11}$$
where $\tilde{\alpha}(z) = \sum_{j=0}^{\infty} z^j \alpha_j$. Note that $\tilde{\alpha}(0) = \alpha_0 > 0$ and $\tilde{\alpha}(1) = \sum_{j=0}^{\infty} \alpha_j = 1$. Consequently, $z = 0$ is **not** a solution to (3.11), whereas $z = 1$ is a solution to (3.11).

- The important question which needs to be addressed now is this: *Is $z = 1$ the only solution to (3.11) on $[0,1]$?*

- To answer this question, we observe the following facts concerning the function $\tilde{\alpha}(z)$:

  (i) $\tilde{\alpha}(z)$ is clearly a continuous function of z on the interval $[0,1]$,

  (ii) $\tilde{\alpha}'(z) = \frac{\mathrm{d}}{\mathrm{d}z}\tilde{\alpha}(z) = \sum_{j=1}^{\infty} jz^{j-1}\alpha_j$ and $\tilde{\alpha}'(1) = j\alpha_j = \mu$,

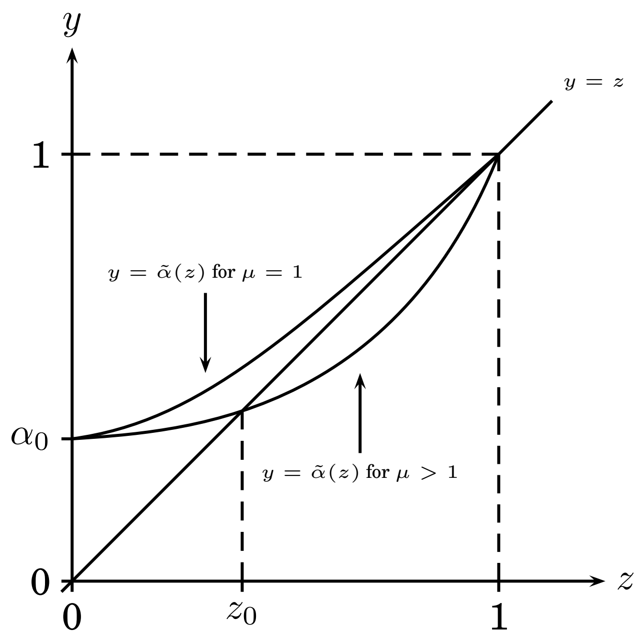
  (iii) $\tilde{\alpha}'(z) > 0$ for $z > 0 \implies \tilde{\alpha}(z)$ is an increasing function of z on $(0,1]$.

  (iv) $\tilde{\alpha}''(z) = \frac{\mathrm{d}}{\mathrm{d}z}\tilde{\alpha}'(z) = \sum_{j=2}^{\infty} j(j-1)z^{j-2}\alpha_j$,

  (v) $\tilde{\alpha}''(z) > 0$ for $z > 0$ since $\alpha_0 + \alpha_1 < 1 \implies \tilde{\alpha}(z)$ is concave up for $z \in (0,1]$.

- As a result of these facts, the following diagram depicts the behaviour of the function $y = \tilde{\alpha}(z)$ in relation to $y = z$:

  In other words, when $\mu = 1$ (i.e., the so-called *critical case*), there is only one root in $[0,1]$, namely $z = 1$. However, when $\mu > 1$, there is a second root $z = z_0 \in (0,1)$ which satisfies (3.11). Therefore, this now raises the question: *Is $\pi_0 = z_0$ or $\pi_0 = 1$ when $\mu > 1$?*



- **We have**: $z = \tilde{\alpha}(z) = \sum_{j=0}^{\infty} z^j \alpha_j \leftarrow$ (3.11). Let $z = z^\star$ be any non-negative solution satisfying (3.11). It is straightforward to show by mathematical induction that $z^\star \geq \mathbb{P}(X_n = 0)$, $n \in \mathbb{N}$ (left as an upcoming exercise). As a result, it follows that

$$\lim_{n\to\infty} z^\star \geq \lim_{n\to\infty} \mathbb{P}(X_n = 0) \implies z^\star \geq \pi_0,$$

  which implies that $z = \pi_0$ is the *smallest* positive number satisfying (3.11). It is only when $\mu > 1$ (referred to as the *supercritical* case) that π_0 is known to exist in the interval $(0, 1)$.

Example 3.13. Given the following offspring probabilities, what is the probability that the population dies out in the long run assuming that $X_0 = 1$?

(a) $\alpha_0 = 3/4$, $\alpha_1 = 1/8$, $\alpha_2 = 1/8$.

Solution: First, we calculate

$$\mu = 0\left(\frac{3}{4}\right) + 1\left(\frac{1}{8}\right) + 2\left(\frac{1}{8}\right) = \frac{3}{8}.$$

Since $\mu < 1$, the population will die out with probability 1.

(b) $\alpha_0 = 1/5$, $\alpha_1 = 1/10$, $\alpha_2 = 7/10$.

Solution: Again, we begin by calculating

$$\mu = 0\left(\frac{1}{5}\right) + 1\left(\frac{1}{10}\right) + 2\left(\frac{7}{10}\right) = 1.5.$$

Since $\mu > 1$, $\pi_0 \in (0, 1)$ is known to exist. To find π_0, we solve (3.11):

$$\begin{aligned} z &= \tilde{\alpha}(z) \\ z &= \sum_{j=0}^{\infty} z^j \alpha^j \\ z &= \frac{1}{5} + z\left(\frac{1}{10}\right) + z^2\left(\frac{7}{10}\right), \end{aligned}$$

giving rise to a quadratic equation

$$7z^2 - 9z + 2 = 0,$$

or equivalently (since we know $z = 1$ is always a solution to the above equation)

$$(7z - 2)(z - 1) = 0.$$

The two roots are $z = 2/7$ and $z = 1$. Thus, $\pi_0 = 2/7$.

Remarks:

(1) In the case when $X_0 = n$, $n \in \mathbb{Z}^+$, the population will die out iff the families of each of the n members of the initial generation die out. As a result, it immediately follows that the extinction probability is simply π_0^n.

(2) For certain choices of the offspring distribution, the Galton-Watson branching process is not very interesting to analyse. For example, with $X_0 = 1$ and $\alpha_r = 1$ for some $r \in \mathbb{N}$, the evolution of the process is purely deterministic (i.e., $\mathbb{P}(X_n = r^n) = 1$, $n \in \mathbb{N}$). Another uninteresting case occurs when $\alpha_0, \alpha_1 > 0$ and $\alpha_0 + \alpha_1 = 1$. In this situation, the population remains at its initial size $X_0 = 1$ for a random number of generations (according to a geometric distribution), before dying out completely.

3.4.2 Interesting Application #2: The Gambler's Ruin Problem

- One of the most powerful ideas in the theory of DTMCs is that many fundamental probabilities and expectations can be computed as the solutions of systems of linear equations.
- We have already seen one such example of this through the application of the BLT.
- In what follows, we will continue to illustrate this idea by deriving appropriate linear systems for a number of key probabilities and expectations that arise in certain settings.
- To motivate this idea, let us consider a famous problem in stochastic processes known as the *Gambler's Ruin Problem*.

The Gambler's Ruin Problem

Example $\boldsymbol{\pi} \approx$ 3.14. Consider a gambler who, at each play of a game, has probability $p \in (0, 1)$ of winning one unit and probability $q = 1 - p$ of losing one unit. Assume that successive plays of the game are independent. If the gambler initially begins with i units, what is the probability that the gambler's fortune will reach N units ($N < \infty$) before reaching 0 units? This problem is often referred to as the *Gambler's Ruin Problem*, with state 0 representing bankruptcy and state N representing the jackpot.

Solution: For $n \in \mathbb{N}$, define X_n as the gambler's fortune after the n^{th} play of the game, with $X_0 = i$. Clearly, $\{X_n, n \in \mathbb{N}\}$ is a DTMC with TPM

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ N-1 \\ N \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & \cdots & N-2 & N-1 & N \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ q & 0 & p & 0 & \cdots & 0 & 0 & 0 \\ 0 & q & 0 & p & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & q & 0 & p \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix} \end{array}.$$

Note that states 0 and N are treated as absorbing states, implying that they are both recurrent. States $\{1, 2, \ldots, N-1\}$ belong to the same communication class, and it is straightforward to verify that it is a transient class (left as an upcoming exercise). Our goal is to determine $G(i)$, $i = 0, 1, \ldots, N$, which represents the probability that starting with i units, the gambler's fortune will eventually reach N units. As a consequence, we can deduce that

$$\lim_{n\to\infty} P^{(n)} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ N-1 \\ N \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & \cdots & N-2 & N-1 & N \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 1-G(1) & 0 & 0 & 0 & \cdots & 0 & 0 & G(1) \\ 1-G(2) & 0 & 0 & 0 & \cdots & 0 & 0 & G(2) \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 1-G(N-1) & 0 & 0 & 0 & \cdots & 0 & 0 & G(N-1) \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix} \end{array}.$$

From above, note that $G(0) = 0$ and $G(N) = 1$ (initial conditions). Moreover, by conditioning on the outcome of the very first game, we readily obtain for $i = 1, 2, \ldots, N-1$:

$$\begin{aligned} G(i) &= pG(i+1) + qG(i-1) \\ 1 \cdot G(i) &= pG(i+1) + qG(i-1) \\ (p+q)G(i) &= pG(i+1) + qG(i-1) \\ pG(i) + qG(i) &= pG(i+1) + qG(i-1) \\ p\big[G(i+1) - G(i)\big] &= q\big[G(i) - G(i-1)\big] \\ G(i+1) - G(i) &= \frac{q}{p}\big[G(i) - G(i-1)\big]. \end{aligned}$$

To determine whether an explicit solution is possible, consider several values of i as follows:

$$\begin{aligned} &\text{Take } i = 1 \implies G(2) - G(1) = \frac{q}{p}\big[G(1) - G(0)\big] = \frac{q}{p}G(1), \\ &\text{Take } i = 2 \implies G(3) - G(2) = \frac{q}{p}\big[G(2) - G(1)\big] = \left(\frac{q}{p}\right)^2 G(1), \\ &\text{Take } i = 3 \implies G(4) - G(3) = \frac{q}{p}\big[G(3) - G(2)\big] = \left(\frac{q}{p}\right)^3 G(1). \end{aligned}$$

Therefore, if we take $i = k$ we get:

$$G(k+1) - G(k) = \frac{q}{p}\big[G(k) - G(k-1)\big] = \left(\frac{q}{p}\right)^k G(1).$$

Note that the above k equations are linear in terms of $G(1), G(2), \ldots, G(k+1)$. Summing these k equations yields:

$$\begin{aligned} G(k+1) - G(1) &= \sum_{i=1}^{k} \left(\frac{q}{p}\right)^i G(1) \\ G(k+1) &= \sum_{i=0}^{k} \left(\frac{q}{p}\right)^i G(1), \text{ for } k = 0, 1, \ldots, N-1, \end{aligned}$$

or equivalently,

$$G(k) = \sum_{i=0}^{k-1} \left(\frac{q}{p}\right)^i G(1), \text{ for } k = 1, 2, \ldots, N.$$

Applying the formula for a finite geometric series, we end up with

$$G(k) = \begin{cases} \dfrac{1-(q/p)^k}{1-(q/p)} G(1), & \text{if } p \neq 1/2, \\ kG(1), & \text{if } p = 1/2. \end{cases}$$

Plugging in $k = N$, we obtain for $p \neq 1/2$:

$$1 = G(N) = \frac{1-(q/p)^N}{1-(q/p)} G(1) \implies G(1) = \frac{1-(q/p)}{1-(q/p)^N}.$$

Similarly, for $p = 1/2$, we obtain:

$$G(1) = \frac{1}{N}.$$

Combining both cases, we ultimately get for $k = 1, 2, \ldots, N$.

$$G(k) = \begin{cases} \dfrac{1-(q/p)^k}{1-(q/p)^N}, & \text{if } p \neq 1/2, \\ \dfrac{k}{N}, & \text{if } p = 1/2. \end{cases}$$

In fact, this holds for $k = 0, 1, 2, \ldots, N$.

Remarks:

(1) An interesting question to ask is what happens to the gambler's probability of winning the jackpot, given an initial fortune of i units, as N grows "larger" (i.e., $N \to \infty$)? In other words, what happens to the limit of $G(i)$ as $N \to \infty$. Looking at three cases based on the value of p, we see:

(i) When $p = 1/2$, $G(i) = \dfrac{i}{N} \to 0$ as $N \to \infty$,

(ii) When $p < 1/2$, $G(i) = \dfrac{1-(q/p)^i}{1-(q/p)^N} \to 0$ as $N \to \infty$,

(iii) When $p > 1/2$, $G(i) = \dfrac{1-(q/p)^i}{1-(q/p)^N} \to 1 - \left(\dfrac{q}{p}\right)^i$ as $N \to \infty$,

since $q/p > (<)1$ when $p < (>)1/2$. Simply put, only when $p > 1/2$ does a positive probability exist that the gambler's fortune will increase indefinitely. Otherwise, the gambler is sure to go broke.

(2) In our study of the *Random Walk* in Example 3.9 featuring a DTMC on the state space $\mathbb{Z}$ with transition probabilities analogous to those in the *Gambler's Ruin Problem*, we previously showed that

$$f_{0,0} = (1-p)f_{-1,0} + pf_{1,0}.$$

Suppose that $p > 1/2$. First, note that

$$\begin{aligned} f_{1,0} &= \mathbb{P}(\text{Random Walk DTMC ever makes a future visit to state } 0 \text{ starting from state } 1) \\ &= \lim_{N\to\infty} \mathbb{P}(\text{Gambler's Ruin DTMC ends up in bankruptcy} \mid X_0 = 1) \\ &= 1 - \left(1 - \left(\frac{1-p}{p}\right)^1\right) \text{ from case (iii) of Remark (1)} \\ &= \frac{1-p}{p}. \end{aligned}$$

Similarly, we also have that

$$\begin{aligned} f_{-1,0} &= \mathbb{P}(\text{Random Walk DTMC ever makes a future visit to state } 0 \text{ starting from state } -1) \\ &= \lim_{N\to\infty} \mathbb{P}(\text{Gambler's Ruin DTMC with "up" probability } 1-p \text{ ends up in bankruptcy} \mid X_0 = 1) \\ &= 1 - 0 \text{ from case (ii) of Remark (1) since } 1-p < 1/2 \\ &= 1. \end{aligned}$$

Therefore, we end up ultimately obtaining

$$f_{0,0} = (1-p) \cdot 1 + p\left(\frac{1-p}{p}\right) = 2(1-p),$$

which agrees with our earlier result. The same essential procedure can be adapted to verify that $f_{0,0} = 2p$ if $p < 1/2$.

WEEK 8
3rd to 10th November

3.5 Absorbing DTMCs

The *Gambler's Ruin Problem* is actually an example of a more general problem which we turn our attention to next. In particular, consider a DTMC $\{X_n, n \in \mathbb{N}\}$ with a finite number of states arranged specifically as follows:

$$\underbrace{0, 1, \ldots, M-1}_{\text{transient states}}, \underbrace{M, M+1, \ldots, N}_{\text{absorbing states}}.$$

The TPM for this DTMC can be expressed as:

$$P = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ M-1 \\ M \\ M+1 \\ \vdots \\ N \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & \cdots & M-1 & M & M+1 & \cdots & N \end{array} \\ \left[\begin{array}{c|c} Q & R \\ \hline 0 & I \end{array}\right] \end{array},$$

where

$$Q = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ M-1 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & \cdots & M-1 \end{array} \\ \begin{bmatrix} Q_{0,0} & Q_{0,1} & \cdots & Q_{0,M-1} \\ Q_{1,0} & Q_{1,1} & \cdots & Q_{1,M-1} \\ \vdots & \vdots & & \vdots \\ Q_{M-1,0} & Q_{M-1,1} & \cdots & Q_{M-1,M-1} \end{bmatrix} \end{array},$$

$$R = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ M-1 \end{array} \begin{array}{c} \begin{array}{cccc} M & M+1 & \cdots & N \end{array} \\ \begin{bmatrix} R_{0,M} & R_{0,M+1} & \cdots & R_{0,N} \\ R_{1,M} & R_{1,M+1} & \cdots & R_{1,N} \\ \vdots & \vdots & & \vdots \\ R_{M-1,M} & R_{M-1,M+1} & \cdots & R_{M-1,N} \end{bmatrix} \end{array},$$

0 is a matrix of zero dimension $(N-M+1) \times M$, and I is an identity matrix of dimension $(N-M+1) \times (N-M+1)$.

Absorbing DTMCs: Absorption Probability

- In what follows, let i be a transient state (i.e., $0 \le i \le M-1$) and assume that $X_0 = i$.
- Let $T = \min\{n \in \mathbb{Z}^+ : M \le X_n \le N\}$ be the *absorption time* rv.
- For $M \le k \le N$ (i.e., k is an absorbing state), consider the *absorption probability into state k from state i* defined by

$$U_{i,k} = \mathbb{P}(X_T = k \mid X_0 = i).$$

Conditioning on the state of the DTMC at time 1, note that

$$\begin{aligned}
U_{i,k} &= \mathbb{P}(X_T = k \mid X_0 = i) \\
&= \sum_{j=0}^{N} \mathbb{P}(X_T = k \mid X_1 = j, X_0 = i)\,\mathbb{P}(X_1 = j \mid X_0 = i) \\
&= \sum_{j=0}^{M-1} \mathbb{P}(X_T = k \mid X_1 = j, X_0 = i)P_{i,j} + \sum_{j=M}^{N} \mathbb{P}(X_T = k \mid X_1 = j, X_0 = i)P_{i,j} \\
&= \sum_{j=0}^{M-1} P_{i,j}\,\mathbb{P}(X_T = k \mid X_1 = j, X_0 = i) + P_{i,k},
\end{aligned}$$

since it follows that for $M \le j \le N$, $T \mid (X_1 = j, X_0 = i)$ is degenerate at 1, and so $\mathbb{P}(X_T = k \mid X_1 = j, X_0 = i) = \delta_{j,k}$, where $\delta_{j,k}$ denotes the *Kronecker delta* function given

$$\delta_{j,k} = \begin{cases} 0, & \text{if } j \ne k, \\ 1, & \text{if } j = k. \end{cases}$$

For $0 \le j \le M-1$, however, note that

$$\mathbb{P}(X_T = k \mid X_1 = j, X_0 = i) = \mathbb{P}(X_T = k \mid X_0 = j) = U_{j,k}.$$

To see this, let T_i denote the remaining number of transitions until absorption given that the DTMC is currently in transient state i. Clearly, $T \mid (X_0 = i) \sim T_i$. Moreover, for transient state j, we have

$$T \mid (X_1 = j, X_0 = i) \sim (1 + T_j) \mid (X_1 = j),$$

due to the Markov property.

- **We have**: $T \mid (X_1 = j, X_0 = i) \sim (1 + T_j) \mid (X_1 = j)$. Therefore,

$$\begin{aligned}
\mathbb{P}(X_T = k \mid X_1 = j, X_0 = i) &= \mathbb{P}(X_{1+T_j} = k \mid X_1 = j) \\
&= \mathbb{P}(X_{T_j} = k \mid X_0 = j) \text{ due to the stationary assumption} \\
&= \mathbb{P}(X_T = k \mid X_0 = j) \\
&= U_{j,k},
\end{aligned}$$

where the second last equality holds due to T_j and T being equivalent in distribution under the condition $X_0 = j$. As a result, we ultimately end up with

$$U_{i,k} = P_{i,k} + \sum_{j=0}^{M-1} P_{i,j}U_{j,k} = R_{i,k} + \sum_{j=0}^{M-1} Q_{i,j}U_{j,k} \;\forall 0 \le i \le M-1,\; M \le k \le N \tag{3.12}$$

In other words, to determine $U_{i,k}$ for a particular pair of values for i and k, the system of M linear equations given by (3.12) must be solved, yielding solutions for $U_{0,k}, U_{1,k}, \ldots, U_{M-1,k}$.

Example 3.12. (*continued*) Recall the earlier DTMC we considered having TPM

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \\ 0 & 0 & 1 \end{bmatrix} \end{array}.$$

We previously claimed that

$$\lim_{n\to\infty} P^{(n)} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ \frac{2}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 \end{bmatrix} \end{array}.$$

Show that the absorption probabilities from transient state 1 into states 0 and 2 are equal to $\lim_{n\to\infty} P_{1,0}^{(n)}$ and $\lim_{n\to\infty} P_{1,2}^{(n)}$, respectively.

Solution: First, relabel the states of this DTMC as follows:

$$\begin{aligned} 0^\star &= \text{state } 1 \text{ in the original DTMC}, \\ 1^\star &= \text{state } 0 \text{ in the original DTMC}, \\ 2^\star &= \text{state } 2 \text{ in the original DTMC}. \end{aligned}$$

As a result, the "new" TPM corresponding to states $\{0^\star, 1^\star, 2^\star\}$ looks like:

$$P = \begin{array}{c} \\ 0^\star \\ 1^\star \\ 2^\star \end{array}\begin{array}{c} \begin{array}{ccc} 0^\star & 1^\star & 2^\star \end{array} \\ \left[\begin{array}{c|cc} \frac{1}{2} & \frac{1}{3} & \frac{1}{6} \\ \hline 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right] \end{array},$$

so that

$$Q = \left[\tfrac{1}{2}\right], \qquad R = \left[\tfrac{1}{3} \quad \tfrac{1}{6}\right].$$

Using (3.12) we find that $U_{0^\star,1^\star}$ and $U_{0^\star,2^\star}$ to be

$$U_{0^\star,1^\star} = R_{0^\star,1^\star} + Q_{0^\star,0^\star}U_{0^\star,1^\star} = \frac{1}{3} + \frac{1}{2}U_{0^\star,1^\star} \implies U_{0^\star,1^\star} = \frac{2}{3} = \lim_{n\to\infty} P_{1,0}^{(n)},$$

and

$$U_{0^\star,2^\star} = R_{0^\star,2^\star} + Q_{0^\star,0^\star}U_{0^\star,2^\star} = \frac{1}{6} + \frac{1}{2}U_{0^\star,2^\star} \implies U_{0^\star,2^\star} = \frac{1}{3} = \lim_{n\to\infty} P_{1,2}^{(n)}.$$

Remarks:

(1) If we define $U = [U_{i,k}]$ to be the $M \times (N - M + 1)$ matrix of absorption probabilities, then (3.12) can be expressed more succinctly in matrix form as

$$U = R + QU,$$

or equivalently,

$$(I - Q)U = R.$$

A known mathematical fact is that the matrix $I - Q$ is invertible, and so an explicit (matrix) solution for U is given by

$$U = (I - Q)^{-1}R. \tag{3.13}$$

In the previous example, note that

$$U = \left[U_{0^\star,1^\star} \quad U_{0^\star,2^\star}\right] = \left(1 - \frac{1}{2}\right)^{-1}\left[\tfrac{1}{3} \quad \tfrac{1}{6}\right] = 2\left[\tfrac{1}{3} \quad \tfrac{1}{6}\right] = \left[\tfrac{2}{3} \quad \tfrac{1}{3}\right]$$

Absorbing DTMCs: Limiting Behaviour

(2) As demonstrated in the previous example, the limiting behaviour of the general DTMC we are considering can be characterized as follows:

$$\lim_{n\to\infty} P^{(n)} = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ M-1 \\ M \\ M+1 \\ \vdots \\ N \end{array} \begin{array}{c} \begin{array}{cccc|cccc} 0 & 1 & \cdots & M-1 & M & M+1 & \cdots & N \end{array} \\ \left[\begin{array}{cccc|cccc} 0 & 0 & \cdots & 0 & U_{0,M} & U_{0,M+1} & \cdots & U_{0,N} \\ 0 & 0 & \cdots & 0 & U_{1,M} & U_{1,M+1} & \cdots & U_{1,N} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & U_{M-1,M} & U_{M-1,M+1} & \cdots & U_{M-1,N} \\ \hline 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{array}\right] \end{array}.$$

Another way to see this is through the use of Exercise 3.5.1, which states that

$$P^{(n)} = \begin{bmatrix} Q^n & \sum_{i=0}^{n-1} Q^i R \\ 0 & I \end{bmatrix}, \; n \in \mathbb{Z}^+.$$

Therefore, it follows that

$$\lim_{n\to\infty} P^{(n)} = \begin{bmatrix} \lim_{n\to\infty} Q^n & \left(\lim_{n\to\infty} \sum_{i=0}^{n-1} Q^i\right) R \\ 0 & I \end{bmatrix}.$$

However, $\lim_{n\to\infty} Q^n = 0$ (as Q has only transient states). Moreover, note that

$$\begin{aligned}(I-Q)\left(\lim_{n\to\infty}\sum_{i=0}^{n-1} Q^i\right) &= \lim_{n\to\infty}\sum_{i=0}^{n-1}(I-Q)Q^i \\ &= \lim_{n\to\infty}\sum_{i=0}^{n-1}(Q^i - Q^{i+1}) \\ &= \lim_{n\to\infty}\left((Q^0-Q^1)+(Q^1-Q^2)+\cdots+(Q^{n-1}-Q^n)\right) \\ &= \lim_{n\to\infty}(I-Q^n) \\ &= I - \lim_{n\to\infty} Q^n \\ &= I.\end{aligned}$$

We have: $U = (I-Q)^{-1}R \leftarrow$ (3.13). Therefore, we have that

$$\sum_{i=0}^{\infty} Q^i = \lim_{n\to\infty}\sum_{i=0}^{n-1} Q^i = (I-Q)^{-1} I = (I-Q)^{-1},$$

which yields a formula for an infinite geometric series of matrices. From (3.13), we obtain

$$\lim_{n\to\infty} P^{(n)} = \begin{bmatrix} 0 & U \\ 0 & I \end{bmatrix}.$$

Absorbing DTMCs: Gambler's Ruin Problem

(3) In the *Gambler's Ruin Problem*, if we reorder the states $0, 1, 2, \ldots, N-1, N$ as

$$\underbrace{1, 2, \ldots, N-1}_{\text{transient}}, \underbrace{N, 0}_{\text{absorbing}},$$

then $U_{i,N} = G(i)$ and $U_{i,0} = 1 - G(i)$, $i = 1, 2, \ldots, N-1$.

Absorbing DTMCs: Absorption Probability

Example 3.15. Consider a DTMC with TPM

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \left[\begin{array}{cc|cc} 0.4 & 0.3 & 0.2 & 0.1 \\ 0.1 & 0.3 & 0.3 & 0.3 \\ \hline 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right] \end{array}.$$

Suppose that the DTMC begins in state 1. What is the probability that the DTMC ultimately ends up in state 3? How would this probability change if the DTMC begins in state 0 with probability $3/4$ and in state 1 with probability $1/4$?

Solution: First, we wish to calculate $U_{1,3}$. In this example,

$$Q = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 0.4 & 0.3 \\ 0.1 & 0.3 \end{bmatrix} \end{array}, \qquad R = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 2 & 3 \end{array} \\ \begin{bmatrix} 0.2 & 0.1 \\ 0.3 & 0.3 \end{bmatrix} \end{array}, \qquad U = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 2 & 3 \end{array} \\ \begin{bmatrix} U_{0,2} & U_{0,3} \\ U_{1,2} & U_{1,3} \end{bmatrix} \end{array}.$$

Since $U = (I-Q)^{-1}R$, we need to find the inverse of

$$I - Q = \begin{bmatrix} 0.6 & -0.3 \\ -0.1 & 0.7 \end{bmatrix}.$$

<u>Recall</u>: For a 2×2 matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \implies A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}, \text{ provided that } ad - bc \neq 0.$$

Applying this formula, we get:

$$(I - Q)^{-1} = \begin{bmatrix} \frac{70}{39} & \frac{10}{13} \\ \frac{10}{39} & \frac{20}{13} \end{bmatrix}.$$

Therefore,

$$U = (I - Q)^{-1}R = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 2 & 3 \end{array} \\ \begin{bmatrix} \frac{23}{39} & \frac{16}{39} \\ \frac{20}{39} & \frac{19}{39} \end{bmatrix} \end{array}.$$

Thus, $U_{1,3} = 19/39 \simeq 0.487$. Under the alternative set of conditions, we should calculate:

$$\begin{aligned} \mathbb{P}(\text{DTMC ultimately ends up in state 3}) &= \mathbb{P}(X_0 = 0)\,\mathbb{P}(\text{DTMC ultimately ends up in state } 3 \mid X_0 = 0) \\ &\quad + \mathbb{P}(X_0 = 1)\,\mathbb{P}(\text{DTMC ultimately ends up in state } 3 \mid X_0 = 1) \\ &= \frac{3}{4}U_{0,3} + \frac{1}{4}U_{1,3} \\ &= \frac{3}{4} \cdot \frac{16}{39} + \frac{1}{4} \cdot \frac{19}{39} \\ &= \frac{67}{156} \simeq 0.429. \end{aligned}$$

Exercise: Use (3.12) to solve the linear system of equations for $U_{0,3}$ and $U_{1,3}$.

An interesting feature of the above methodology is that it can even be used for DTMCs in which the set of absorbing states are replaced by one or more recurrent classes. The following example demonstrates the basic idea.

Example 3.16. Consider a DTMC with TPM

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \left[\begin{array}{cc|ccc} 0.4 & 0.3 & 0.2 & 0.1 & 0 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0.7 & 0.3 \\ 0 & 0 & 0 & 0.4 & 0.6 \end{array}\right] \end{array}.$$

Suppose that the DTMC begins in state 1. What is the probability that the DTMC ultimately ends up in state 3?

Solution: We wish to determine $\lim_{n\to\infty} P_{1,3}^{(n)}$. To do this, let us group states 3 and 4 together as a single state, which will be denoted by $3^\star$. As a result of this grouping, our revised TPM has the form:

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3^\star \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3^\star \end{array} \\ \left[\begin{array}{cc|cc} 0.4 & 0.3 & 0.2 & 0.1 \\ 0.1 & 0.3 & 0.3 & 0.3 \\ \hline 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right] \end{array},$$

which is identical to the TPM from Example 3.15. Using the results of Example 3.15, we know that $U_{1,3^\star} = \frac{19}{39}$. However, once in state $3^\star$, the DTMC will remain in recurrent class $\{3, 4\}$ with associated TPM

$$U = \begin{array}{c} \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cc} 3 & 4 \end{array} \\ \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \end{array}.$$

For this "smaller" DTMC, the conditions of the BLT are satisfied, meaning that we can determine the limiting probabilities π_3 and π_4 by solving:

$$(\pi_3, \pi_4) = (\pi_3, \pi_4) \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}, \text{ subject to } \pi_3 + \pi_4 = 1.$$

This leads to $\pi_3 = 4/7$ and $\pi_4 = 3/7$. Thus, it ultimately follows that

$$\lim_{n\to\infty} P_{1,3}^{(n)} = U_{1,3^\star} \cdot \pi_3 = \frac{19}{39} \cdot \frac{4}{7} = \frac{76}{273} \simeq 0.278.$$

Absorbing DTMCs: Mean Absorption Time

Next, for $0 \le i \le M-1$, let $v_i = \mathbb{E}[T \mid X_0 = i]$ be the *mean absorption time from state* i. Once again, conditioning on the state of the DTMC at time 1, we get

$$\begin{aligned} v_i &= \mathbb{E}[T \mid X_0 = i] \\ &= \sum_{j=0}^{M-1} \mathbb{E}[T \mid X_1 = j, X_0 = i]P_{i,j} + \sum_{j=M}^{N} \mathbb{E}[T \mid X_1 = j, X_0 = i]P_{i,j} \\ &= \sum_{j=0}^{M-1} \mathbb{E}[T \mid X_1 = j, X_0 = i]P_{i,j} + \sum_{j=M}^{N} 1 \cdot P_{i,j}, \end{aligned}$$

since $T \mid (X_1 = j, X_0 = i)$ is degenerate at 1 for $M \le j \le N$. For $0 \le j \le M-1$, our previous arguments lead to

$$T \mid (X_1 = j, X_0 = i) \sim (1 + T_j) \mid (X_1 = j) \sim 1 + T \mid (X_0 = j),$$

which implies that

$$\begin{aligned} v_i &= \sum_{j=0}^{M-1} \big(1 + \mathbb{E}[T \mid X_0 = j]\big)P_{i,j} + \sum_{j=M}^{N} 1 \cdot P_{i,j} \\ &= 1 + \sum_{j=0}^{M-1} P_{i,j}v_j \\ &= 1 + \sum_{j=0}^{M-1} Q_{i,j}v_j. \end{aligned} \tag{3.14}$$

We have: $v_i = 1 + \sum_{j=0}^{M-1} Q_{i,j}v_j \leftarrow$ (3.14). If we now define the $M \times 1$ column vector of mean absorption times

$$\underline{v}^\top = \begin{bmatrix} v_0 \\ v_1 \\ \vdots \\ v_{M-1} \end{bmatrix},$$

then (3.14) yields in matrix form

$$\underline{v}^\top = \underline{e}^\top + Q\underline{v}^\top.$$

Therefore, an explicit (matrix) solution for $\underline{v}^\top$ is

$$\underline{v}^\top = (I - Q)^{-1}\underline{e}^\top.$$

Example 3.12. (*continued*) Recall the modified TPM

$$P^\star = \begin{array}{c} \\ 0^\star \\ 1^\star \\ 2^\star \end{array} \begin{array}{c} \begin{array}{ccc} 0^\star & 1^\star & 2^\star \end{array} \\ \left[\begin{array}{cc|c} 1/2 & 1/3 & 1/6 \\ \hline 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right] \end{array}.$$

What is the mean absorption time for this DTMC, given that it begins in state $0^\star$?

Solution: Using the matrix equation for $\underline{v}^\top = [v_{0^\star}]$, we have

$$v_{0^\star} = (I - Q)^{-1}\underline{e}^\top = \left(1 - \frac{1}{2}\right)^{-1}(1) = 2.$$

Looking at this particular TPM, given that the DTMC initially begins in state $0^\star$, each transition will return to state $0^\star$ with probability of $1/2$, or become absorbed into one of the two absorbing states with probability $1/3 + 1/6 = 1/2$. Therefore, the number of transitions required for absorption to occur simply follows a geometric distribution, namely

$$T \mid (X_0 = 0^\star) \sim \text{GEO}_t\left(\frac{1}{2}\right) \implies \mathbb{E}[T \mid X_0 = 0^\star] = \frac{1}{1/2} = 2.$$

Example 3.15. (*continued*) If the DTMC with TPM

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \left[\begin{array}{cc|cc} 0.4 & 0.3 & 0.2 & 0.1 \\ 0.1 & 0.3 & 0.3 & 0.3 \\ \hline 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right] \end{array},$$

begins in state 1, how long, on average, does it take to end up in either of states 2 or 3?

Solution: Since we have two transient states, we wish to find v_1 where

$$\underline{v}^\top = \begin{bmatrix} v_0 \\ v_1 \end{bmatrix}.$$

Note that

$$\begin{aligned} \underline{v}^\top &= (I - Q)^{-1} \underline{e}^\top \\ &= \begin{bmatrix} \frac{70}{39} & \frac{10}{13} \\ \frac{10}{39} & \frac{20}{13} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \end{aligned}$$

and so

$$v_1 = \frac{10}{39} + \frac{20}{13} = \frac{70}{39} \simeq 1.79.$$

Absorbing DTMCs: Mean Number of Transient State Visits

As before, assume that $X_0 = i$ where $0 \le i \le M - 1$. Let ℓ be a transient state as well, so that $0 \le \ell \le M - 1$. Define the following sequence of indicator rvs $\{A_n\}_{n=0}^\infty$ such that

$$A_n = \begin{cases} 0, & \text{if } X_n \neq \ell, \\ 1, & \text{if } X_n = \ell. \end{cases}$$

We are interested in computing the quantity

$$W_{i,\ell} = \mathbb{E}\left[\sum_{n=0}^{T-1} A_n \,\middle|\, X_0 = i\right],$$

which represents the *mean number of times that state ℓ is visited (including time 0) prior to absorption given* *that* $X_0 = i$. To begin, note that

$$\begin{aligned}
W_{i,\ell} &= \mathbb{E}\left[\sum_{n=0}^{T-1} A_n \,\middle|\, X_0 = i\right] \\
&= \mathbb{E}\left[A_0 + \sum_{n=1}^{T-1} A_n \,\middle|\, X_0 = i\right] \\
&= \mathbb{E}[A_0 \mid X_0 = i] + \mathbb{E}\left[\sum_{n=1}^{T-1} A_n \,\middle|\, X_0 = i\right] \\
&= 0 \cdot \mathbb{P}(A_0 = 0 \mid X_0 = i) + 1 \cdot \mathbb{P}(\underbrace{A_0 = 1}_{X_0=\ell} \mid X_0 = i) + \mathbb{E}\left[\sum_{n=1}^{T-1} A_n \,\middle|\, X_0 = i\right] \\
&= \delta_{i,\ell} + \mathbb{E}\left[\sum_{n=1}^{T-1} A_n \,\middle|\, X_0 = i\right].
\end{aligned}$$

To find $\mathbb{E}\left[\sum_{n=1}^{T-1} A_n \,\middle|\, X_0 = i\right]$, we condition on the state of the DTMC at time 1 to obtain

$$\mathbb{E}\left[\sum_{n=1}^{T-1} A_n \,\middle|\, X_0 = i\right] = \sum_{j=0}^{M-1} \mathbb{E}\left[\sum_{n=1}^{T-1} A_n \,\middle|\, X_1 = j, X_0 = i\right] P_{i,j} + \sum_{j=M}^{N} \mathbb{E}\left[\sum_{n=1}^{T-1} A_n \,\middle|\, X_1 = j, X_0 = i\right] P_{i,j}.$$

When $M \leq j \leq N$, note that

$$\mathbb{E}\left[\sum_{n=1}^{T-1} A_n \,\middle|\, X_1 = j, X_0 = i\right] = \mathbb{E}\left[\sum_{n=1}^{1-1} A_n \,\middle|\, X_1 = j, X_0 = i\right] = 0,$$

since $T \mid (X_1 = j, X_0 = i)$ is degenerate at 1. When $0 \leq j \leq M - 1$, recall that

$$T \mid (X_1 = j, X_0 = i) \sim (1 + T_j) \mid (X_1 = j)$$

and

$$T_j \sim T \mid (X_0 = j),$$

where T_j denotes the remaining number of transitions until absorption given that the DTMC is currently in transient state j. Therefore,

$$\begin{aligned}
\sum_{j=0}^{M-1} \mathbb{E}\left[\sum_{n=1}^{T-1} A_n \,\middle|\, X_1 = j, X_0 = i\right] P_{i,j} &= \sum_{j=0}^{M-1} \mathbb{E}\left[\sum_{n=1}^{1+T_j-1} A_n \,\middle|\, X_1 = j\right] P_{i,j} \\
&= \sum_{j=0}^{M-1} \mathbb{E}\left[\sum_{n=1}^{T_j} A_n \,\middle|\, X_1 = j\right] P_{i,j} \\
&= \sum_{j=0}^{M-1} \mathbb{E}\left[\sum_{m=0}^{T_j-1} A_{m+1} \,\middle|\, X_1 = j\right] P_{i,j}.
\end{aligned}$$

$$
\begin{aligned}
&\sum_{j=0}^{M-1} \mathbb{E}\left[\sum_{n=1}^{T-1} A_n \,\middle|\, X_1 = j, X_0 = i\right] P_{i,j} \\
&= \sum_{j=0}^{M-1} \mathbb{E}\left[\sum_{m=0}^{T_j-1} A_{m+1} \,\middle|\, X_1 = j\right] P_{i,j} \\
&= \sum_{j=0}^{M-1} P_{i,j}\,\mathbb{E}\left[\sum_{m=0}^{T_j-1} A_m \,\middle|\, X_0 = j\right] \text{ due to the stationary assumption} \\
&= \sum_{j=0}^{M-1} P_{i,j}\,\mathbb{E}\left[\sum_{m=0}^{T-1} A_m \,\middle|\, X_0 = j\right] \text{ since } T_j \mid (X_0 = j) \sim T \mid (X_0 = j) \\
&= \sum_{j=0}^{M-1} Q_{i,j} W_{j,\ell}.
\end{aligned}
$$

Therefore, we ultimately end up with

$$
W_{i,\ell} = \delta_{i,\ell} + \sum_{j=0}^{M-1} Q_{i,j} W_{j,\ell}, \; 0 \le i, \ell \le M-1. \tag{3.15}
$$

In matrix notation, define the $M \times M$ matrix $W = [W_{i,\ell}]$, so that (3.15) in matrix form then becomes

$$
W = I + QW.
$$

Solving for W leads to

$$
\begin{aligned}
W - QW &= I \\
(I - Q)W &= I \\
W &= (I - Q)^{-1}.
\end{aligned}
$$

Remark: From our earlier result, we recognize that

$$
\underline{v}^\top = (I - Q)^{-1} \underline{e}^\top = W \underline{e}^\top.
$$

In other words, by summing the rows of W (i.e., adding up the mean number of times each of the individual transient states is visited), we actually obtain the mean number of transitions to reach absorption.

Absorbing DTMCs: Visitation Probability

Finally, let us again consider $0 \le i, \ell \le M-1$ and recall that

$$
\begin{aligned}
f_{i,\ell} &= \mathbb{P}(\text{DTMC ever makes a future visit to state } \ell \mid X_0 = i) \\
&= \mathbb{P}(X_n = \ell \text{ for some } 1 \le n \le T-1 \mid X_0 = i).
\end{aligned}
$$

Using an elementary conditioning argument (as well as the Markov and stationary properties of the

DTMC), note that

$$\begin{aligned}
W_{i,\ell} &= \mathbb{E}\left[\sum_{n=0}^{T-1} A_n \,\middle|\, X_0 = i\right] \\
&= \underbrace{\mathbb{E}\left[\sum_{n=0}^{T-1} A_n \,\middle|\, X_0 = i, X_n = \ell \text{ for some } 1 \le n \le T-1\right] \cdot f_{i,\ell}}_{\text{state } \ell \text{ is visited possibly at time 0 and then at some point later on}} \\
&\quad + \underbrace{\mathbb{E}\left[\sum_{n=0}^{T-1} A_n \,\middle|\, A_n X_0 = i, X_n \neq \ell \text{ for some } 1 \le n \le T-1\right] \cdot (1 - f_{i,\ell})}_{\text{only possible visit to state } \ell \text{ is at time 0}} \\
&= (\delta_{i,\ell} + W_{\ell,\ell}) \cdot f_{i,\ell} + \delta_{i,\ell} \cdot (1 - f_{i,\ell}) \\
&= \delta_{i,\ell} + W_{\ell,\ell} f_{i,\ell}.
\end{aligned}$$

This immediately leads to

$$f_{i,\ell} = \frac{W_{i,\ell} - \delta_{i,\ell}}{W_{\ell,\ell}}.$$

<u>Remark</u>: For $\ell = i$, the probability of the DTMC visiting state i in the future (given that $X_0 = i$) is

$$f_{i,i} = \frac{W_{i,i} - \delta_{i,i}}{W_{i,i}} = 1 - \frac{1}{W_{i,i}},$$

where $W_{i,i}$ is the expected number of visits to state i (including time 0) before absorption. From the above relation,

$$W_{i,i} = \frac{1}{1 - f_{i,i}}.$$

However, recall that for transient state i, the random number M_i of **future** visits to state i, given $X_0 = i$, has conditional pmf given by (3.4), namely

$$\mathbb{P}(M_i = k \mid X_0 = i) = f_{i,i}^k (1 - f_{i,i}), k = 0, 1, 2, \ldots,$$

which we recognize as the pmf of a GEO$_f(1 - f_{i,i})$ rv. Therefore, the expected number of **future** visits to state i is

$$\mathbb{E}[M_i \mid X_0 = i] = \frac{1 - (1 - f_{i,i})}{1 - f_{i,i}} = \frac{1}{1 - f_{i,i}} - 1 = W_{i,i} - 1,$$

which is in agreement with the definition of $W_{i,i}$, since we count the initial visit to state i occurring at time 0.

Example 3.15. (*continued*) Recall the DTMC with TPM

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \left[\begin{array}{cc|cc} 0.4 & 0.3 & 0.2 & 0.1 \\ 0.1 & 0.3 & 0.3 & 0.3 \\ \hline 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right] \end{array}.$$

Given $X_0 = 1$, what is the average number of visits to state 0 prior to absorption? Also, what is the probability that the DTMC ever makes a visit to state 0?

Solution: First, we wish to find $W_{1,0}$ where

$$W = \begin{array}{c c} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} W_{0,0} & W_{0,1} \\ W_{1,0} & W_{1,1} \end{bmatrix} \end{array}.$$

From our earlier calculations,

$$\begin{aligned} W &= (I - Q)^{-1} \\ &= \begin{bmatrix} \frac{70}{39} & \frac{10}{13} \\ \frac{10}{39} & \frac{20}{13} \end{bmatrix}. \end{aligned}$$

Thus, $W_{1,0} = 10/39 \simeq 0.257$. Lastly, we calculate

$$f_{1,0} = \frac{W_{1,0} - \delta_{1,0}}{W_{0,0}} = \frac{(10/39) - 0}{(70/39)} = \frac{10}{39} \times \frac{39}{70} = \frac{1}{7} \simeq 0.143.$$

Chapter 4

The Exponential Distribution and the Poisson Process

WEEK 9
10th to 17th November

4.1 Properties of the Exponential Distribution

Basic Distributional Results

If a rv X has an exponential distribution with parameter $\lambda > 0$ (i.e., $X \sim \text{EXP}(\lambda)$ where λ is often referred to as the "rate"), then we have the following basic distributional results in place:

- **pdf**: $f(x) = \lambda e^{-\lambda x}$, $x > 0$,
- **cdf**: $F(x) = \mathbb{P}(X \leq x) = \int_0^x \lambda e^{-\lambda y} \, \mathrm{d}y = 1 - e^{-\lambda x}$, $x \geq 0$,
- **tpf**: $\bar{F}(x) = \mathbb{P}(X > x) = 1 - F(x) = e^{-\lambda x}$, $x \geq 0$,
- **mgf**: $\phi_X(t) = \mathbb{E}[e^{tX}] = \int_0^\infty e^{tx} \lambda e^{-\lambda x} \, \mathrm{d}x = \frac{\lambda}{\lambda - t} \int_0^\infty (\lambda - t) e^{-(\lambda - t)x} \, \mathrm{d}x = \frac{\lambda}{\lambda - t}$, $t < \lambda$,
- **mean**: $\mathbb{E}[X] = 1/\lambda$,
- **variance**: $\mathrm{Var}(X) = 1/\lambda^2$.

Minimum of Independent Exponentials

Minimum of Independent Exponentials: Let $\{X_i\}_{i=1}^n$ be a sequence of independent rvs where $X_i \sim \text{EXP}(\lambda_i)$, $i = 1, 2, \ldots, n$. Define $Y = \min\{X_1, X_2, \ldots, X_n\}$ to be the smallest *order statistic* of $\{X_1, X_2, \ldots, X_n\}$. Clearly, Y takes on possible values in the state space $\mathcal{S} = (0, \infty)$. To determine the distribution of Y, consider its tpf:

$$\begin{aligned}
\bar{F}_Y(y) &= \mathbb{P}(Y > y) \\
&= \mathbb{P}\big(\min\{X_1, X_2, \ldots, X_n\} > y\big) \\
&= \mathbb{P}(X_1 > y, X_2 > y, \ldots, X_n > y) \\
&= \mathbb{P}(X_1 > y)\, \mathbb{P}(X_2 > y) \cdots \mathbb{P}(X_n > y) && \text{by independence} \\
&= e^{-\lambda_1 y} e^{-\lambda_2 y} \cdots e^{-\lambda_n y} && \text{provided that } y \geq 0 \\
&= \underbrace{e^{-(-\lambda_1 + \lambda_2 + \cdots + \lambda_n) y}}_{\text{tpf of an EXP}\left(\sum_{i=1}^n \lambda_i\right) \text{ rv}}, && y \geq 0.
\end{aligned}$$

Therefore, $Y = \min\{X_1, X_2, \dots, X_n\} \sim \text{EXP}(\sum_{i=1}^{n} \lambda_i)$.

Remark: As a special case of this result, if we additionally assume that $X_1, X_2, \dots, X_n$ are iid EXP(λ) rvs, then $Y = \min\{X_1, X_2, \dots, X_n\} \sim \text{EXP}(n\lambda)$.

Example 4.1. Let $\{X_i\}_{i=1}^{n}$ be a sequence of independent rvs where $X_i \sim \text{EXP}(\lambda_i)$, $i = 1, 2, \dots, n$.

(a) For $j \in \{1, 2, \dots, n\}$, determine $\mathbb{P}\big(X_j = \min\{X_1, X_2, \dots, X_n\}\big)$.

Solution: We wish to determine

$$\begin{aligned}
&\mathbb{P}\big(X_j = \min\{X_1, X_2, \dots, X_n\}\big) \\
&= \underbrace{\mathbb{P}(X_j < X_1, X_j < X_2, \dots, X_j < X_n)}_{(n-1)\text{-fold intersection}} \\
&= \int_0^\infty \mathbb{P}(X_j < X_1, X_j < X_2, \dots, X_j < X_n \mid X_j = x)\lambda_j e^{-\lambda_j x}\,\mathrm{d}x \\
&= \int_0^\infty \mathbb{P}(X_1 > x, X_2 > x, \dots, X_n > x)\lambda_j e^{-\lambda_j x}\,\mathrm{d}x \text{ since } X_j \text{ is independent of } \{X_i\}_{i=1}^{n}\ i \neq j \\
&= \int_0^\infty \mathbb{P}(X_1 > x)\,\mathbb{P}(X_2 > x)\cdots\mathbb{P}(X_n > x)\lambda_j e^{-\lambda_j x}\,\mathrm{d}x \\
&= \int_0^\infty e^{-\lambda_1 x} e^{-\lambda_2 x}\cdots e^{-\lambda_n x}\lambda_j e^{-\lambda_j x}\,\mathrm{d}x \\
&= \frac{\lambda_j}{\sum_{i=1}^{n}\lambda_i}\int_0^\infty \underbrace{\left(\sum_{i=1}^{n}\lambda_i\right) e^{-\left(\sum_{i=1}^{n}\lambda_i\right)x}}_{\text{EXP}(\lambda_1+\lambda_2+\cdots+\lambda_n)\text{ pdf}}\,\mathrm{d}x \\
&= \frac{\lambda_j}{\sum_{i=1}^{n}\lambda_i}.
\end{aligned} \tag{4.1}$$

(b) Show that the condition rv $X_1 \mid (X_1 < X_2 < \cdots < X_n)$ is identically distributed to the rv $\min\{X_1, X_2, \dots, X_n\}$.

Solution: Let $Y = X_1 \mid (X_1 < X_2 < \cdots < X_n)$. The tpf of Y is given by

$$\begin{aligned}
\bar{F}_Y(y) &= \mathbb{P}(X_1 > y \mid X_1 < X_2 < \cdots < X_n) \\
&= \frac{\mathbb{P}(y < X_1 < X_2 < \cdots < X_n)}{\mathbb{P}(X_1 < X_2 < \cdots < X_n)},\ y \geq 0.
\end{aligned} \tag{4.2}$$

Note that

$$\begin{aligned}
&\mathbb{P}(y < X_1 < X_2 < \cdots < X_n) \\
&= \int_y^\infty \int_{x_1}^\infty \int_{x_2}^\infty \cdots \int_{x_{n-1}}^\infty \left(\prod_{i=1}^{n}\lambda_i e^{-\lambda_i x_i}\right)\mathrm{d}x_n \cdots \mathrm{d}x_3\,\mathrm{d}x_2\,\mathrm{d}x_1 \\
&= \prod_{i=1}^{n-1}\lambda_i \int_y^\infty e^{-\lambda_1 x_1}\int_{x_1}^\infty e^{-\lambda_2 x_2} \times \\
&\quad \times \int_{x_2}^\infty e^{-\lambda_3 x_3}\cdots\int_{x_{n-2}}^\infty e^{-\lambda_{n-1}x_{n-1}}\int_{x_{n-1}}^\infty \lambda_n e^{-\lambda_n x_n}\,\mathrm{d}x_n\,\mathrm{d}x_{n-1}\cdots\mathrm{d}x_3\,\mathrm{d}x_2\,\mathrm{d}x_1 \\
&= \frac{\prod_{i=1}^{n-1}\lambda_i}{\prod_{i=1}^{n-1}(\sum_{j=i}^{n}\lambda_j)} e^{-\left(\sum_{i=1}^{n}\lambda_i\right)y}.
\end{aligned} \tag{4.3}$$

Using (4.3), we immediately obtain:

$$\begin{aligned}\mathbb{P}(X_1 < X_2 < \cdots < X_n) &= \mathbb{P}(0 < X_1 < X_2 < \cdots < X_n) \\ &= \frac{\prod_{i=1}^{n-1} \lambda_i}{\prod_{i=1}^{n-1} (\sum_{j=i}^{n} \lambda_j)}.\end{aligned} \tag{4.4}$$

Therefore, substituting (4.3) and (4.4) into (4.2) yields

$$\bar{F}_Y(y) = e^{-\left(\sum_{i=1}^{n} \lambda_i\right)y}, \ y \geq 0,$$

which is the tpf of an $\text{EXP}\left(\sum_{i=1}^{n} \lambda_i\right)$ rv. Since

$$\min\{X_1, X_2, \ldots, X_n\} \sim \text{EXP}\left(\sum_{i=1}^{n} \lambda_i\right),$$

it follows that

$$Y = X_1 \mid (X_1 < X_2 < \cdots < X_n) \sim \min\{X_1, X_2 \ldots, X_n\}.$$

Remarks:

(1) In the case when $n = 2$, note that the result from part (a) simplifies to become

$$\mathbb{P}\big(X_1 = \min\{X_1, X_2\}\big) = \mathbb{P}(X_1 < X_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2},$$

which agrees with the result of Example 2.11.

(2) Interestingly, looking at the derivation in part (b), we see that

$$\begin{aligned}&\mathbb{P}(X_1 < X_2 < \cdots < X_n) \\ &= \frac{\lambda_1}{\lambda_1 + \lambda_2 + \cdots + \lambda_n} \cdot \frac{\lambda_2}{\lambda_2 + \lambda_3 + \cdots + \lambda_n} \cdots \frac{\lambda_{n-2}}{\lambda_{n-2} + \lambda_{n-1} + \lambda_n} \cdot \frac{\lambda_{n-1}}{\lambda_{n-1} + \lambda_n} \\ &= \prod_{i=1}^{n-1} \mathbb{P}\big(X_i = \min\{X_i, X_{i+1}, \ldots, X_n\}\big).\end{aligned}$$

Memoryless Property

Memoryless Property: A rv X is *memoryless* iff

$$\mathbb{P}(X > y + z \mid X > y) = \mathbb{P}(X > z) \ \forall y, z \geq 0.$$

Note that if we express $\mathbb{P}(X > y + z \mid X > y)$ as $\mathbb{P}(X - y > z \mid X > y)$ and think of X as being the lifetime of some component, then the memoryless property (or sometimes referred to as the *forgetfulness property*) states that the distribution of the remaining lifetime is independent of the time the component has already lasted.
In other words, such a probability distribution is independent of its history.

An equivalent way to define the memoryless property is given by the following theorem.

Theorem 4.1. X is memoryless iff $\mathbb{P}(X > y + z) = \mathbb{P}(X > y)\,\mathbb{P}(X > z)\ \forall y, z \geq 0$.

Proof: ($\Longrightarrow$) Note

$$\begin{aligned}\mathbb{P}(X > y + z \mid X > y) &= \frac{\mathbb{P}(X > y + z, X > y)}{\mathbb{P}(X > y)} \\ &= \frac{\mathbb{P}(X > y + z)}{\mathbb{P}(X > y)}.\end{aligned}$$

If X is memoryless, then

$$\mathbb{P}(X > y + z \mid X > y) = \mathbb{P}(X > z)\ \forall y, z \geq 0,$$

and so

$$\begin{aligned}\mathbb{P}(X > z) &= \frac{\mathbb{P}(X > y + z)}{\mathbb{P}(X > y)} \\ \mathbb{P}(X > y + z) &= \mathbb{P}(X > z)\,\mathbb{P}(X > y).\end{aligned}$$

($\Longleftarrow$) Conversely, if $\mathbb{P}(X > y + z) = \mathbb{P}(X > y)\,\mathbb{P}(X > z)\ \forall y, z \geq 0$, then

$$\begin{aligned}\mathbb{P}(X > y + z \mid X > y) &= \frac{\mathbb{P}(X > y + z)}{\mathbb{P}(X > y)} \\ &= \frac{\mathbb{P}(X > y)\,\mathbb{P}(X > z)}{\mathbb{P}(X > y)} \\ &= \mathbb{P}(X > z).\end{aligned}$$

By definition, X is memoryless.

This leads to the main result concerning the exponential distribution.

Theorem 4.2. An exponential distribution is memoryless.

Proof: Suppose that $X \sim \text{EXP}(\lambda)$. For $y, z \geq 0$, we have

$$\begin{aligned}\mathbb{P}(X > y + z) &= e^{-\lambda(y+z)} \\ &= e^{-\lambda y} e^{-\lambda z} \\ &= \mathbb{P}(X > y)\,\mathbb{P}(X > z).\end{aligned}$$

Thus, by Theorem 4.1, X is memoryless.

The Exponential Distribution

Example 4.2. Suppose that a computer has 3 switches which govern the transfer of electronic impulses. These switches operate simultaneously and independently of one another, with lifetimes that are exponentially distributed with mean lifetimes of 10, 5, and 4 years, respectively.

(a) What is the probability that the time until the very first switch breakdown exceeds 6 years?

Solution: Let X_i represent the lifetime of switch i, $i = 1, 2, 3$. We know that $X_i \sim \text{EXP}(\lambda_i)$ where $\lambda_1 = 1/10$, $\lambda_2 = 1/5$, and $\lambda_3 = 1/4$. The time until the 1^{st} breakdown is defined by the rv

$Y = \min\{X_1, X_2, X_3\}$. Since the lifetimes are independent of each other,

$$Y \sim \text{EXP}(\lambda),\ \lambda = \frac{1}{10} + \frac{1}{5} + \frac{1}{4} = \frac{11}{20}.$$

We wish to calculate:

$$\mathbb{P}(Y > 6) = e^{-(11/20)(6)} = e^{-3.3} \simeq 0.0369.$$

(b) What is the probability that switch 2 outlives switch 1?

Solution: We simply want to compute

$$\mathbb{P}(X_1 < X_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{(1/10)}{(1/10) + (1/5)} = \frac{1}{3} \simeq 0.3\bar{3}.$$

(c) What is the probability that switch 1 has the longest lifetime, followed next by switch 3 and then switch 2?

Solution: We wish to calculate

$$\mathbb{P}(X_2 < X_3 < X_1).$$

To do so, let $Y_1 = X_2$, $Y_2 = X_3$, $Y_3 = X_1$, so that

$$Y_i \sim \text{EXP}(\lambda_i^\star),\ i = 1, 2, 3,$$

with $\lambda_1^\star = 1/5$, $\lambda_2^\star = 1/4$, and $\lambda_3^\star = 1/10$. Therefore,

$$\begin{aligned}
\mathbb{P}(X_2 < X_3 < X_1) &= \mathbb{P}(Y_1 < Y_2 < Y_3) \\
&= \frac{\prod_{i=1}^{3-1} \lambda_i^\star}{\prod_{i=1}^{3-1}(\sum_{j=i}^{3} \lambda_j^\star)} && \text{by (4.4)} \\
&= \frac{(1/5)(1/4)}{(1/5 + 1/4 + 1/10)(1/4 + 1/10)} \\
&= \frac{1/20}{(11/20)(7/20)} \\
&= \frac{20}{77} \simeq 0.26.
\end{aligned}$$

(d) If switch 3 is known to have lasted 2 years, what is the probability it will last at most 3 more years?

Solution: We wish to calculate

$$\begin{aligned}
\mathbb{P}(X_3 \leq 5 \mid X_3 > 2) &= 1 - \mathbb{P}(X_3 > 5 \mid X_3 > 2) \\
&= 1 - \mathbb{P}(X_3 > 2 + 3 \mid X_3 > 2) \\
&= 1 - \mathbb{P}(X_3 > 3) && \text{due to the memoryless property} \\
&= 1 - e^{-(1/4)(3)} \\
&= 1 - e^{-0.75} \simeq 0.528.
\end{aligned}$$

(e) Considering only switches 1 and 2, what is the expected amount of time until they have both suffered a breakdown?

Solution: We wish to solve for

$$\mathbb{E}\big[\max\{X_1, X_2\}\big].$$

We note the following useful identity:

$$\min\{X_1, X_2\} + \max\{X_1, X_2\} = X_1 + X_2.$$

Taking expectations of the above equality, we ultimately obtain:

$$\begin{aligned}\mathbb{E}\big[\max\{X_1, X_2\}\big] &= \mathbb{E}[X_1] + \mathbb{E}[X_2] - \mathbb{E}\big[\min\{X_1, X_2\}\big] \\ &= 10 + 5 - \frac{1}{1/10 + 1/5} \\ &= 15 - \frac{10}{3} \\ &= \frac{35}{3} \simeq 11.6\bar{6}.\end{aligned}$$

Memoryless Property

Remarks:

(1) The exponential distribution is the *unique* continuous distribution possessing the memoryless property (incidentally, the geometric distribution is the *unique* discrete distribution which is memoryless, which is not all that surprising in light of Exercise 2.2.3).

To prove this statement, suppose that X is a continuous rv satisfying the memoryless property. Let $\bar{F}(x) = \mathbb{P}(X > x)$, which is a continuous function of x. By Theorem 4.2, it follows that

$$\bar{F}(y+z) = \bar{F}(y)\bar{F}(z) \ \forall y, z \geq 0.$$

Note that

$$\bar{F}\left(\frac{2}{n}\right) = \bar{F}\left(\frac{1}{n} + \frac{1}{n}\right) = \bar{F}^2\left(\frac{1}{n}\right).$$

As a result, it immediately follows that $\bar{F}(m/n) = \bar{F}^m(1/n)$. Furthermore,

$$\bar{F}(1) = \bar{F}\left(\frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n}\right) = \bar{F}^n\left(\frac{1}{n}\right),$$

or equivalently

$$\bar{F}\left(\frac{1}{n}\right) = \big(\bar{F}(1)\big)^{1/n}.$$

Thus, $\bar{F}(x) = \big(\bar{F}(1)\big)^x$ for all rational values of x, and by the continuity of $\bar{F}(x)$, this implies that $\bar{F}(x) = \big(\bar{F}(1)\big)^x \ \forall x \geq 0$. However, note that we can write

$$\bar{F}(x) = e^{\ln\left(\bar{F}(1)\right)^x} = e^{x\ln\left(\bar{F}(x)\right)} = e^{-\lambda x},$$

where $\lambda = -\ln\big(\bar{F}(1)\big) > 0$. In other words,

$$F(x) = \mathbb{P}(X \leq x) = 1 - e^{-\lambda x},$$

which shows that X is exponentially distributed.

(2) The memoryless property of the exponential distribution even holds in a broader setting. Specifically, if $X \sim \text{EXP}(\lambda)$, then

$$\mathbb{P}(X > Y + Z \mid X > Y) = \mathbb{P}(X > Z), \tag{4.5}$$

where Y and Z are independently distributed non-negative valued rvs which are both independent of X. The equality defined by (4.5) is referred to as the *generalized memoryless property*.

To prove that the above result holds, note that

$$\mathbb{P}(X > Y + Z \mid X > Y) = \frac{\mathbb{P}(X > Y + Z, X > Y)}{\mathbb{P}(X > Y)}.$$

Without loss of generality, assume that Y and Z are independent continuous rvs, so that

$$
\begin{aligned}
&\mathbb{P}(X > Y + Z, X > Y) \\
&= \int_0^\infty \mathbb{P}(X > Y + Z, X > Y \mid Y = y) f_Y(y)\,\mathrm{d}y \\
&= \int_0^\infty \mathbb{P}(X > y + Z, X > y) f_Y(y)\,\mathrm{d}y && \text{since } X, Y, \text{ and } Z \text{ are independent rvs} \\
&= \int_0^\infty \mathbb{P}(X > y + Z) f_Y(y)\,\mathrm{d}y \\
&= \int_0^\infty \left(\int_0^\infty \mathbb{P}(X > y + Z \mid Z = z) f_Z(z)\,\mathrm{d}z \right) f_Y(y)\,\mathrm{d}y \\
&= \int_0^\infty \left(\int_0^\infty \mathbb{P}(X > y + z) f_Z(z)\,\mathrm{d}z \right) f_Y(y)\,\mathrm{d}y && \text{since } X \text{ and } Z \text{ are independent rvs} \\
&= \int_0^\infty \left(\int_0^\infty e^{-\lambda(y+z)} f_Z(z)\,\mathrm{d}z \right) f_Y(y)\,\mathrm{d}y \\
&= \int_0^\infty \left(\int_0^\infty e^{-\lambda z} f_Z(z)\,\mathrm{d}z \right) e^{-\lambda y} f_Y(y)\,\mathrm{d}y \\
&= \int_0^\infty \mathbb{P}(X > Z) e^{-\lambda y} f_Y(y)\,\mathrm{d}y \\
&= \mathbb{P}(X > Z) \int_0^\infty e^{-\lambda y} f_Y(y)\,\mathrm{d}y \\
&= \mathbb{P}(X > Z)\, \mathbb{P}(X > Y),
\end{aligned}
$$

since we have for independent continuous rvs Y (and similarly for Z)

$$\mathbb{P}(X > Y) = \int_0^\infty \mathbb{P}(X > y) f_Y(y)\,\mathrm{d}y = \int_0^\infty e^{-\lambda y} f_Y(y)\,\mathrm{d}y.$$

Thus,

$$\mathbb{P}(X > Y + Z \mid X > Y) = \frac{\mathbb{P}(X > Y + Z, X > Y)}{\mathbb{P}(X > Y)} = \frac{\mathbb{P}(X > Z)\,\mathbb{P}(X > Y)}{\mathbb{P}(X > Y)} = \mathbb{P}(X > Z).$$

(3) The generalized memoryless property implies that $(X - Y) \mid (X > Y) \sim$ EXP(λ) regardless of the distribution Y. To see this, let Z be a rv with a degenerate distribution at z. In this case, (4.5) becomes

$$\mathbb{P}(X > Y + z \mid X > Y) = \mathbb{P}(X > z) = e^{-\lambda z},$$

since $X \sim$ EXP(λ). Thus,

$$\mathbb{P}(X - Y > z \mid X > Y) = e^{-\lambda z},$$

and so

$$(X - Y) \mid (X > Y) \sim \text{EXP}(\lambda).$$

The Exponential Distribution

Example 4.3. Let X_1 and X_2 be independent rvs where $X_i \sim$ EXP(λ_i), $i = 1, 2$. Given $X_1 < X_2$, show that X_1 and $X_2 - X_1$ are conditionally independent rvs.

Solution: Consider the following conditional joint cdf:

$$\begin{aligned}\mathbb{P}(X_1 \le x, X_2 - X_1 \le y \mid X_1 < X_2) &= \frac{\mathbb{P}(X_1 \le x, X_2 - X_1 \le y, X_1 < X_2)}{\mathbb{P}(X_1 < X_2)} \\ &= \frac{\mathbb{P}(X_1 \le x, X_1 \ge X_2 - y, X_1 < X_2)}{\mathbb{P}(X_1 < X_2)} \\ &= \frac{\mathbb{P}(X_1 \le x, X_1 \ge X_2 - y, X_1 < X_2)}{\frac{\lambda_1}{\lambda_1+\lambda_2}} \\ &= \frac{\lambda_1 + \lambda_2}{\lambda_1}\,\mathbb{P}(X_1 \le x, X_1 \ge X_2 - y, X_1 < X_2) \\ &= \frac{\lambda_1 + \lambda_2}{\lambda_1}\,\mathbb{P}\big(X_2 - y \le X_1 \le \min\{x, X_2\}\big),\ \forall x, y \ge 0.\end{aligned}$$

Suppose that $x \le y$. It follows that

$$\begin{aligned}&\mathbb{P}\big(X_2 - y \le X_1 \le \min\{x, X_2\}\big) \\ &= \int_0^\infty \mathbb{P}\big(X_2 - y \le X_1 \le \min\{x, X_2\} \bigm| X_2 = w\big) f_{X_2}(w)\,\mathrm{d}w \\ &= \int_0^\infty \mathbb{P}\big(w - y \le X_1 \le \min\{x, w\}\big) f_{X_2}(w)\,\mathrm{d}w \text{ since } X_1 \text{ and } X_2 \text{ are independent} \\ &= \int_0^x \mathbb{P}\big(\underbrace{w-y}_{<0} \le X_1 \le \underbrace{\min\{x,w\}}_{=w}\big) f_{X_2}(w)\,\mathrm{d}w + \int_x^y \mathbb{P}\big(\underbrace{w-y}_{<0} \le X_1 \le \underbrace{\min\{x,w\}}_{=x}\big) f_{X_2}(w)\,\mathrm{d}w \\ &\quad + \int_y^{y+x} \mathbb{P}\big(\underbrace{w-y}_{>0} \le X_1 \le \underbrace{\min\{x,w\}}_{=x}\big) f_{X_2}(w)\,\mathrm{d}w + \int_{y+x}^\infty \underbrace{\mathbb{P}\big(\underbrace{w-y}_{>x} \le X_1 \le \underbrace{\min\{x,w\}}_{=x}\big)}_{=0} f_{X_2}(w)\,\mathrm{d}w \\ &= \int_0^x \mathbb{P}(X_1 \le w)\lambda_2 e^{-\lambda_2 w}\,\mathrm{d}w + \mathbb{P}(X_2 \le x)\int_x^y \lambda_2 e^{-\lambda_2 w}\,\mathrm{d}w \\ &\quad + \int_y^{y+x} \big[\mathbb{P}(X_1 > w - y) - \mathbb{P}(X_1 > x)\big]\lambda_2 e^{-\lambda_2 w}\,\mathrm{d}w \\ &= \int_0^x (1 - e^{-\lambda_1 w})\lambda_2 e^{-\lambda_2 w}\,\mathrm{d}w + (1 - e^{-\lambda_1 x})(e^{-\lambda_2 x} - e^{-\lambda_2 y}) + \int_y^{y+x} (e^{-\lambda_1 (w-y)} - e^{-\lambda_1 x})\lambda_2 e^{-\lambda_2 w}\,\mathrm{d}w \\ &= \frac{\lambda_1}{\lambda_1 + \lambda_2}\big(1 - e^{-\lambda_2 y} - e^{-(\lambda_1+\lambda_2)x} + e^{-\lambda_2 y} e^{-(\lambda_1+\lambda_2)x}\big). \qquad (4.6)\end{aligned}$$

Similarly, it can be shown that (4.6) also holds true in the case when $y \le x$ (Exercise 4.1.2). Therefore, in general, we have:

$$\begin{aligned}\mathbb{P}(X_1 \le x, X_2 - X_1 \le y \mid X_1 < X_2) &= \frac{\lambda_1+\lambda_2}{\lambda_1}\cdot\frac{\lambda_1}{\lambda_1+\lambda_2}\big(1 - e^{-\lambda_2 y} - e^{-(\lambda_1+\lambda_2)x} + e^{-\lambda_2 y} e^{-(\lambda_1+\lambda_2)x}\big) \\ &= 1 - e^{-\lambda_2 y} - e^{-(\lambda_1+\lambda_2)x} + e^{-\lambda_2 y} e^{-(\lambda_1+\lambda_2)x} \\ &= (1 - e^{-(\lambda_1+\lambda_2)x})(1 - e^{-\lambda_2 y}) \\ &= \mathbb{P}(X_1 \le x \mid X_1 < X_2)\,\mathbb{P}(X_2 - X_1 \le y \mid X_1 < X_2),\ \forall x, y \ge 0,\end{aligned}$$

where we applied the result of Example 4.1 (b) (i.e., $X_1 \mid (X_1 < X_2) \sim \min\{X_1, X_2\}$) and the (generalized) memoryless property (i.e., $(X_2 - X_1) \mid (X_2 > X_1) \sim X_2$) to obtain the last equality. Thus, by definition, X_1 and $X_2 - X_1$ are conditionally (given $X_1 < X_2$) independent rvs.

The Erlang Distribution

The Erlang Distribution: Recall that if $X \sim \text{Erlang}(n, \lambda)$ where $n \in \mathbb{Z}^+$ and $\lambda > 0$, then its pdf is of the form

$$f(x) = \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}, \ x > 0.$$

Letting $n = 1$, then above pdf simplifies to become $f(x) = \lambda e^{-\lambda x}$, $x > 0$, which is the EXP(λ) pdf. To obtain the corresponding cdf of an Erlang(n, λ) rv, we consider

$$F(x) = \mathbb{P}(X \le x) = \int_0^x \frac{\lambda^n y^{n-1} e^{-\lambda y}}{(n-1)!} \, \mathrm{d}y = \frac{\lambda^n}{(n-1)!} \int_0^x y^{n-1} e^{-\lambda y} \, \mathrm{d}y, \ x \ge 0.$$

We have: $F(x) = \frac{\lambda^n}{(n-1)!} \int_0^x y^{n-1} e^{-\lambda y} \, \mathrm{d}y$. Assume that $n \ge 2$ and apply integration by parts, that is,

$$\int u \, \mathrm{d}v = uv - \int v \, \mathrm{d}u,$$

to the above integral. In particular, choose

$$u = y^{n-1} \implies \frac{\mathrm{d}u}{\mathrm{d}y} = (n-1)y^{n-2} \implies \mathrm{d}u = (n-1)y^{n-2} \, \mathrm{d}y,$$

and

$$\mathrm{d}v = e^{-\lambda y} \, \mathrm{d}y \implies \int 1 \, \mathrm{d}v = \int e^{-\lambda y} \, \mathrm{d}y \implies v = -\frac{1}{\lambda} e^{-\lambda y},$$

so that

$$\begin{aligned} \int_0^x y^{n-1} e^{-\lambda y} \, \mathrm{d}y &= \left[-\frac{1}{\lambda} y^{n-1} e^{-\lambda y} \right]_{y=0}^{y=x} + \frac{(n-1)}{\lambda} \int_0^x y^{n-2} e^{-\lambda y} \, \mathrm{d}y \\ &= -\frac{1}{\lambda} x^{n-1} e^{-\lambda x} + \frac{(n-1)}{\lambda} \int_0^x y^{n-2} e^{-\lambda y} \, \mathrm{d}y. \end{aligned}$$

Therefore,

$$\begin{aligned} F(x) &= \frac{\lambda^n}{(n-1)!} \left(-\frac{1}{\lambda} x^{n-1} e^{-\lambda x} + \frac{(n-1)}{\lambda} \int_0^x y^{n-2} e^{-\lambda y} \, \mathrm{d}y \right) \\ &= \frac{\lambda^{n-1}}{(n-2)!} \int_0^x y^{n-2} e^{-\lambda y} \, \mathrm{d}y - \frac{(\lambda x)^{n-1}}{(n-1)!} e^{-\lambda x}. \end{aligned}$$

If we continue to apply integration by parts until the "y-term" in the integrand has a power of zero, then it is possible to show that

$$F(x) = 1 - e^{-\lambda x} \sum_{j=0}^{n-1} \frac{(\lambda x)^j}{j!}, \ x \ge 0. \tag{4.7}$$

Remark: Substituting $n = 1$ into (4.7), we immediately obtain

$$F(x) = 1 - e^{-\lambda x} \sum_{j=0}^{1-1} \frac{(\lambda x)^j}{j!} = 1 - e^{-\lambda x}, \ x \ge 0,$$

which is clearly the cdf of an EXP(λ) rv.

To determine the mgf of $X \sim$ Erlang(n, λ), we consider

$$\begin{aligned}
\phi_X(t) &= \int_0^\infty e^{tx} \cdot \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!} \, \mathrm{d}x \\
&= \frac{\lambda^n}{(n-1)!} \int_0^\infty x^{n-1} e^{-\tilde{\lambda} x} \, \mathrm{d}x && \text{where we define } \tilde{\lambda} = \lambda - t \\
&= \frac{\lambda^n}{\tilde{\lambda}^n} \int_0^\infty \underbrace{\frac{\tilde{\lambda}^n x^{n-1} e^{-\tilde{\lambda} x}}{(n-1)!}}_{\text{Erlang}(n, \tilde{\lambda}) \text{ pdf}} \, \mathrm{d}x && \text{provided that } \tilde{\lambda} = \lambda - t > 0 \\
&= \left(\frac{\lambda}{\lambda - t} \right)^n, && t < \lambda.
\end{aligned}$$

However, note that

$$\phi_X(t) = \left(\frac{\lambda}{\lambda - t} \right)^n = \prod_{i=1}^n \left(\frac{\lambda}{\lambda - t} \right), \ t < \lambda,$$

is the product of n terms, where each term is the mgf of an EXP(λ) rv. Let $\{Y_i\}_{i=1}^n$ be the iid sequence of EXP(λ) rvs, with $\phi_{Y_i}(t) = \frac{\lambda}{\lambda - t}$, $t < \lambda$, for $i = 1, 2, \ldots, n$. Since $\phi_X(t) = \prod_{i=1}^n \phi_{Y_i}(t)$, it follows that an Erlang distribution can be viewed as the distribution of a sum of iid exponential rvs. As a result, the mean, and variance of an Erlang(n, λ) rv X are simply obtained as

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^n Y_i \right] = \sum_{i=1}^n \mathbb{E}[Y_i] = \frac{n}{\lambda}$$

and

$$\mathrm{Var}(X) = \mathrm{Var}\left(\sum_{i=1}^n Y_i \right) = \sum_{i=1}^n \mathrm{Var}(Y_i) = \frac{n}{\lambda^2}.$$

WEEK 10
17th to 24th November

4.2 The Poisson Process

Counting Process

Definition: A *counting process* $\{N(t), t \geq 0\}$ is a stochastic process in which $N(t)$ represents the number of events that happen (or occur) by time t, where the index t measures time over a continuous range. Some examples of counting processes $\{N(t), t \geq 0\}$ might include:

(1) $N(t)$ represents the number of automobile accidents at a specified intersection by week t,

(2) $N(t)$ represents the number of births by year t in Canada,

(3) $N(t)$ represents the number of visits to a particular webpage by time t,

(4) $N(t)$ represents the number of customers who enter a store by time t,

(5) $N(t)$ represents the number of accident claims reported to an insurance company by time t.

Basic Properties of Counting Processes:

(1) $N(0) = 0$.

(2) $N(t)$ is a non-negative integer $\forall t \geq 0$ (i.e., $N(t) \in \mathbb{N} \ \forall t \geq 0$).

(3) If $s < t$, then $N(s) \leq N(t)$.

(4) $N(t) - N(s)$ counts the number of events to occur in the time interval $(s, t]$ for $s < t$.

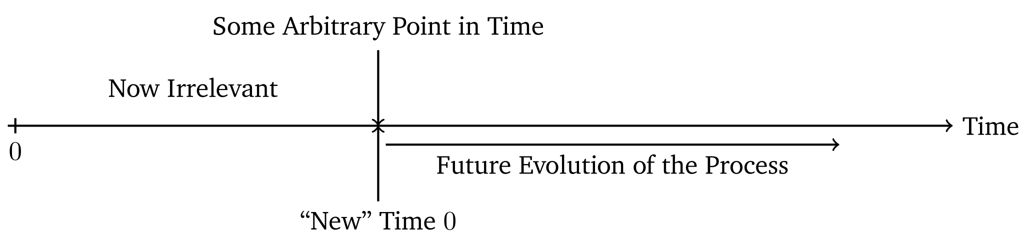
Independent and Stationary Increments

We now introduce two important properties associated with counting processes.

Definition: A counting process $\{N(t), t \geq 0\}$ has *independent increments* if $N(t_1) - N(s_1)$ is independent of $N(t_2) - N(s_2)$ whenever $(s_1, t_1] \cap (s_2, t_2] = \varnothing$ for all choices of s_1, t_1, s_2, t_2 (i.e., the number of events in non-overlapping time intervals are assumed to be independent of each other).

Definition: A counting process $\{N(t), t \geq 0\}$ has *stationary increments* if the distribution of the number of events in $(s, s+t]$ (i.e., $N(s+t) - N(s)$) depends only on t, the length of the time interval. In this case, $N(s+t) - N(s)$ has the same probability distribution as $N(0+t) - N(0) = N(t)$, the number of events occurring in the interval $[0, t]$.

Remark: As the diagram below indicates, the assumption of stationary and independent increments is essentially equivalent to stating that, at any point in time, the process $\{N(t), t \geq 0\}$ *probabilistically restarts itself*.



$\mathrm{o}(h)$ Function

Before introducing the formal definition of a Poisson process, we first introduce a few mathematical tools which are needed.

Definition: A function $y = f(x)$ is said to be “$\mathrm{o}(h)$” (i.e., of order h) if

$$\lim_{h \to 0} \frac{f(h)}{h} = 0.$$

Remark: An $\mathrm{o}(h)$ function $y = f(x)$ is one in which $f(h)$ approaches 0 faster than h does.

Examples:

(1) $y = f(x) = x$. Note that

$$\lim_{h \to 0} \frac{f(h)}{h} = \lim_{h \to 0} \frac{h}{h} = \lim_{h \to 0} 1 \neq 0.$$

Thus, $y = x$ is not of order h.

(2) $y = f(x) = x^2$. Note that

$$\lim_{h\to 0} \frac{f(h)}{h} = \lim_{h\to 0} \frac{h^2}{h} = \lim_{h\to 0} h = 0.$$

Thus, $y = x^2$ is of order h. In fact, the function of the form $y = x^r$ is clearly of order h provided that $r > 1$.

(3) Suppose that $\{f_i(x)\}_{i=1}^n$ is a sequence of $\mathrm{o}(h)$ functions. Consider the linear combination of $\mathrm{o}(h)$ functions, namely $y = \sum_{i=1}^n c_i f_i(x)$, and note that

$$\lim_{h\to 0} \frac{\sum_{i=1}^n c_i f_i}{h} = \sum_{i=1}^n c_i \underbrace{\lim_{h\to 0} \frac{f_i(h)}{h}}_{=0} = 0.$$

Thus, a linear combination of $\mathrm{o}(h)$ functions is still of order h.

Remark: In most cases, this result is actually true when $n = \infty$.

Poisson Process

Definition: A counting process $\{N(t), t \geq 0\}$ is said to be a *Poisson process* at rate λ if the following three conditions hold true:

(1) The process has both independent and stationary increments.

(2) For $h > 0$, $\mathbb{P}\big(N(h) = 1\big) = \lambda h + \mathrm{o}(h)$.

(3) For $h > 0$, $\mathbb{P}\big(N(h) \geq 2\big) = \mathrm{o}(h)$.

Remarks:

(1) Condition (2) in the above definition implies that in a "small" interval of time, the probability of a single event occurring is essentially proportional to the length of the interval.

(2) Condition (3) in the above definition implies that two or more events occurring in a "small" interval of time is rare.

(3) Conditions (2) and (3) yield

$$\begin{aligned}
\mathbb{P}\big(N(h) = 0\big) &= 1 - \mathbb{P}\big(N(h) > 0\big) \\
&= 1 - \mathbb{P}\big(N(h) = 1\big) - \mathbb{P}\big(N(h) \geq 2\big) \\
&= 1 - \big(\lambda h + \mathrm{o}(h)\big) - \mathrm{o}(h) \\
&= 1 - \lambda h - \mathrm{o}(h) - \mathrm{o}(h) \\
&= 1 - \lambda h + \mathrm{o}(h).
\end{aligned}$$

Ultimately, for a Poisson process $\{N(t), t \geq 0\}$ at rate λ, we would like to know the distribution of the rv $N(s+t) - N(s)$, representing the number of events occurring in the interval $(s, s+t]$, $s, t \geq 0$. Since a Poisson process has stationary increments, this rv has the same probability distribution as $N(t)$. The following theorem specifies the distribution of $N(t)$.

Theorem 4.3. If $\{N(t), t \geq 0\}$ is a Poisson process at rate λ, then $N(t) \sim \mathrm{POI}(\lambda t)$.

Proof: For $t \geq 0$, let

$$\underbrace{\phi_t(u) = \mathbb{E}[e^{uN(t)}]}_{\text{mgf of } N(t)}.$$

For $h \geq 0$, consider

$$\begin{aligned}
\phi_{t+h}(u) &= \mathbb{E}[e^{uN(t+h)}] \\
&= \mathbb{E}[e^{u\left(N(t+h)-N(t)+N(t)\right)}] \\
&= \mathbb{E}[e^{u\left(N(t+h)-N(t)\right)}e^{uN(t)}] \\
&= \mathbb{E}[e^{u\left(N(t+h)-N(t)\right)}]\,\mathbb{E}[e^{uN(t)}] && \text{due to independent increments} \\
&= \mathbb{E}[e^{uN(h)}]\,\mathbb{E}[e^{uN(t)}] && \text{due to stationary increments} \\
&= \phi_t(u)\phi_h(u). && (4.8)
\end{aligned}$$

Note that for $j \geq 2$,

$$\begin{aligned}
&0 \leq \mathbb{P}\big(N(h) = j\big) \leq \mathbb{P}\big(N(h) \geq 2\big) \\
\implies &0 \leq \frac{\mathbb{P}\big(N(h) = j\big)}{h} \leq \frac{\mathbb{P}\big(N(h) \geq 2\big)}{h}.
\end{aligned}$$

Letting $h \to 0$, we obtain:

$$0 \leq \lim_{h\to 0} \frac{\mathbb{P}\big(N(h) = j\big)}{h} \leq \lim_{h\to 0} \frac{\mathbb{P}\big(N(h) \geq 2\big)}{h} = 0,$$

since $\mathbb{P}\big(N(h) \geq 2\big)$ is an $\mathrm{o}(h)$ function by condition (3). Therefore, by the Squeeze Theorem

$$\frac{\mathbb{P}\big(N(h) = j\big)}{h} = 0 \implies \mathbb{P}\big(N(h) = j\big) \text{ is of order } h \text{ for } j \geq 2.$$

Using this result, we obtain

$$\begin{aligned}
\phi_h(u) &= \mathbb{E}[e^{uN(h)}] \\
&= \sum_{j=0}^{\infty} e^{uj} \cdot \mathbb{P}\big(N(h) = j\big) \\
&= e^{u(0)}\,\mathbb{P}\big(N(h) = 0\big) + e^{u(1)}\,\mathbb{P}\big(N(h) = 1\big) + \sum_{j=2}^{\infty} e^{uj}\,\mathbb{P}\big(N(h) = j\big) \\
&= \big(1 - \lambda h + \mathrm{o}(h)\big) + e^u\big(\lambda h + \mathrm{o}(h)\big) + \underbrace{\sum_{j=2}^{\infty} c_j\,\mathbb{P}\big(N(h) = j\big)}_{=\mathrm{o}(h)} \text{ where } c_j = e^{uj} \\
&= 1 - \lambda h + e^u\lambda h + \mathrm{o}(h).
\end{aligned}$$

Returning to (4.8), we now have

$$\begin{aligned}
\phi_{t+h}(u) &= \phi_t(u)\big(1 - \lambda h + e^u\lambda h + \mathrm{o}(h)\big) \\
&= \phi_t(u) - \lambda h\phi_t(u) + e^u\lambda h\phi_t(u) + \mathrm{o}(h) \\
\phi_{t+h}(u) - \phi_t(u) &= \lambda h\phi_t(u)(e^u - 1) + \mathrm{o}(h) \\
\frac{\phi_{t+h}(u) - \phi_t(u)}{h} &= \frac{\lambda h\phi_t(u)(e^u - 1)}{h} + \frac{\mathrm{o}(h)}{h}
\end{aligned}$$

Letting $h \to 0$, we obtain:

$$\begin{aligned}
\lim_{h\to 0} \frac{\phi_{t+h}(u) - \phi_t(u)}{h} &= \lim_{h\to 0} \lambda(e^u - 1)\phi_t(u) + \lim_{h\to 0} \frac{\mathrm{o}(h)}{h} \\
\frac{\mathrm{d}}{\mathrm{d}t}\phi_t(u) &= \lambda(e^u - 1)\phi_t(u) \\
\frac{\mathrm{d}}{\mathrm{d}s}\phi_s(u) &= \lambda(e^u - 1)\phi_s(u) \text{ let } t = s \\
\frac{\frac{\mathrm{d}}{\mathrm{d}s}\phi_s(u)}{\phi_s(u)} &= \lambda(e^u - 1) \\
\frac{\mathrm{d}}{\mathrm{d}s}\ln\big(\phi_s(u)\big) &= \lambda(e^u - 1) \\
\mathrm{d}\Big(\ln\big(\phi_s(u)\big)\Big) &= \lambda(e^u - 1)\,\mathrm{d}s \\
\int_0^t \mathrm{d}\Big(\ln\big(\phi_s(u)\big)\Big) &= \int_0^t \lambda(e^u - 1)\,\mathrm{d}s \\
\Big[\ln\big(\phi_s(u)\big)\Big]_{s=0}^{s=t} &= \lambda(e^u - 1)t \\
\ln\big(\phi_t(u)\big) - \ln\big(\phi_0(u)\big) &= \lambda(e^u - 1)t.
\end{aligned}$$

Recall:

$$\phi_0(u) = \mathbb{E}[e^{u\overbrace{N(0)}^{=0}}] = 1.$$

It follows that

$$\begin{aligned}
\ln\big(\phi_t(u)\big) &= \lambda(e^u - 1)t \\
\phi_t(u) &= e^{\lambda(e^u-1)t} \\
&= e^{\lambda t(e^u-1)},\ u \in \mathbb{R}.
\end{aligned}$$

We recognize this function as the mgf of a POI(λt) rv. By the mgf uniqueness property, $N(t) \sim$ POI(λt).

Remark: As a direct consequence of Theorem 4.3, for all $s, t \geq 0$, we have

$$\mathbb{P}\big(N(s+t) - N(s) = k\big) = \mathbb{P}\big(N(t) = k\big) = \frac{e^{-\lambda t}(\lambda t)^k}{k!},\ k = 0, 1, 2, \ldots.$$

Interarrival Times

Interarrival Times: Define T_1 to be the elapsed time (from time 0) until the first event occurs. In general, for $i \geq 2$, let T_i be the elapsed time between the occurrences of the $(i-1)^{\text{th}}$ event and the i^{th} event. The sequence $\{T_i\}_{i=1}^{\infty}$ is called the *interarrival* or *interevent time sequence*. The diagram below depicts the relationship between $N(t)$ and $\{T_i\}_{i=1}^{\infty}$.

A very important result linking a Poisson process to its interarrival time sequence now follows.

Theorem 4.4. If $\big\{N(t), t \geq 0\big\}$ is a Poisson process at rate $\lambda > 0$, then $\{T_i\}_{i=1}^{\infty}$ is a sequence of iid EXP(λ) rvs.

Proof: We begin by considering the rv T_1. For $t \geq 0$, note that

$$\begin{aligned}\mathbb{P}(T_1 > t) &= \mathbb{P}(\text{no events occur before time } t)\\ &= \mathbb{P}\big(N(t) = 0\big)\\ &= \frac{e^{-\lambda t}(\lambda t)^0}{0!}\\ &= e^{-\lambda t},\end{aligned}$$

which is the tpf an EXP(λ) rv. Thus, $T_1 \sim$ EXP(λ). Next, for $s > 0$, and $t \geq 0$, consider

$$\begin{aligned}\mathbb{P}(T_2 > t \,|\, T_1 = s) &= \mathbb{P}\big(T_2 > t \,\big|\, N(w) = 0 \ \forall w \in [0, s) \text{ and } N(s) = 1\big)\\ &= \mathbb{P}\big(\underbrace{\text{no events occur in } (s, s+t]}_{N(s+t)-N(s)=0} \,\big|\, N(w) = 0 \ \forall w \in [0, s) \text{ and } N(s) = 1\big)\\ &= \mathbb{P}\big(N(s+t) - N(s) = 0\big) \text{ due to independent increments}\\ &= \mathbb{P}\big(N(t) = 0\big) \text{ due to stationary increments}\\ &= e^{-\lambda t},\end{aligned}$$

which is independent of s. Thus, T_1 and T_2 are independent rvs and

$$\begin{aligned}\mathbb{P}(T_2 > t) &= \mathbb{P}(T_2 > t \,|\, T_1 = s)\\ &= e^{-\lambda t},\end{aligned}$$

implying that $T_2 \sim$ EXP(λ). Carrying out this process inductively, this desired result is established.

Waiting Times

For $n \in \mathbb{Z}^+$, define S_n to be the total elapsed time until the n^{th} event occurs. In other words, S_n denotes the arrival time of the n^{th} event, or the *waiting time* until the n^{th} event occurs. Clearly, $S_n = \sum_{i=1}^{n} T_i$. If $\big\{N(t), t \geq 0\big\}$ is a Poisson process at rate λ, then $\{T_i\}_{i=1}^{\infty}$ is a sequence of iid EXP(λ) rvs by Theorem 4.4, implying that

$$S_n = \sum_{i=1}^{n} T_i \sim \text{Erlang}(n, \lambda).$$

From our earlier results on the Erlang distribution, we have $\mathbb{E}[S_n] = n/\lambda$, $\text{Var}(S_n) = n/\lambda^2$, and

$$\mathbb{P}(S_n > t) = e^{-\lambda t} \sum_{j=0}^{n-1} \frac{(\lambda t)^j}{j!}, \ t \geq 0.$$

Remarks:

(1) The above formula for the tpf of S_n could have been obtained without reference to the Erlang

distribution. In particular, note that

$$\begin{aligned}\mathbb{P}(S_n > t) &= \mathbb{P}(\text{arrival time of the n}^{\text{th}} \text{ event occurs after time } t) \\ &= \mathbb{P}(\text{at most } n-1 \text{ events occur by time } t) \\ &= \mathbb{P}\big(N(t) \le n-1\big) \\ &= \sum_{j=0}^{n-1} \frac{e^{-\lambda t}(\lambda t)^j}{j!} \qquad \text{since } N(t) \sim \text{POI}(\lambda t) \\ &= e^{-\lambda t} \sum_{j=0}^{n-1} \frac{(\lambda t)^j}{j!}.\end{aligned}$$

(2) If $\{X_i\}_{i=1}^{\infty}$ represents an iid sequence of EXP(λ) rvs and one constructs a counting process $\big\{N(t), t \ge 0\big\}$ defined by $N(t) = \max\{n \in \mathbb{N} : \sum_{i=1}^{n} X_i \le t\}$, then $\big\{N(t), t \ge 0\big\}$ is actually a Poisson process at rate λ. In other words, $\big\{N(t), t \ge 0\big\}$ has both independent and stationary increments (due to the memoryless property that the sequence of rvs $\{X_i\}_{i=1}^{\infty}$ processes), in addition to the fact that

$$\mathbb{P}\big(N(t) \le k\big) = \mathbb{P}\left(\sum_{i=1}^{k+1} X_i > t\right) = e^{-\lambda t} \sum_{j=0}^{k} \frac{(\lambda t)^j}{j!} \text{ since } \sum_{i=1}^{k+1} \sim \text{Erlang}(k+1, \lambda),$$

which subsequently leads to

$$\begin{aligned}\mathbb{P}\big(N(t) = k\big) &= \mathbb{P}\big(N(t) \le k\big) - \mathbb{P}\big(N(t) \le k-1\big) \\ &= e^{-\lambda t} \sum_{j=0}^{k} \frac{(\lambda t)^j}{j!} - e^{-\lambda t} \sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!} \\ &= \frac{e^{-\lambda t}(\lambda t)^k}{k!},\ k = 0, 1, 2, \ldots.\end{aligned}$$

Poisson Process

Example 4.4. At a local insurance company, suppose that fire damage claims come into the company according to a Poisson process at rate 3.8 expected claims per year.

(a) What is the probability that exactly 5 claims occur in the time interval $(3.2, 5]$ (measured in years)?

Solution: Let $N(t)$ be the number of claims arriving to the company in the interval $[0, t]$. Since $\big\{N(t), t \ge 0\big\}$ is a Poisson process with $\lambda = 3.8$, we want to find

$$\mathbb{P}\big(N(5) - N(3.2) = 5\big) = \frac{e^{-3.8(1.8)}\big(3.8(1.8)\big)^5}{5!} \simeq 0.134.$$

(b) What is the probability that the time between the 2nd and 4th claims is between 2 and 5 months?

Solution: Let T be the time between the 2nd and 4th claims. Thus, $T = T_3 + T_4$ where $T_i \sim$ EXP(3.8) for $i = 3, 4$. Since T_3 and T_4 are independent rvs, we have $T \sim$ Erlang$(2, 3.8)$. Recall that

$$\begin{aligned}\mathbb{P}(T > t) &= e^{-3.8t} \sum_{j=0}^{2-1} \frac{(3.8t)^j}{j!} \\ &= e^{-3.8t}(1 + 3.8t),\ t \ge 0.\end{aligned}$$

We wish to calculate

$$\mathbb{P}\left(\frac{1}{6} < T < \frac{5}{12}\right) = \mathbb{P}\left(T > \frac{1}{6}\right) - \mathbb{P}\left(T > \frac{5}{12}\right) \simeq 0.337.$$

WEEK 11
1124 to 1st December

4.3 Further Properties of the Poisson Process

In this section, we shed further light on a number of interesting and important mathematical properties associated with the Poisson process. To begin with, the binomial distribution also arises in the context of Poisson processes, as the following theorem establishes.

Connection to Binomial Distribution

Theorem 4.5. If $\{N(t), t \geq 0\}$ is a Poisson process at rate λ, then

$$N(s) \mid \big(N(t) = n\big) \sim \text{BIN}(n, s/t),\ s < t.$$

Proof: We wish to determine the conditional distribution of $N(s) \mid \big(N(t) = n\big)$ for $s < t$. Clearly, $N(s) \mid \big(N(t) = n\big)$ takes on values in the set $\{0, 1, 2, \ldots, n\}$. Therefore, for $m = 0, 1, 2, \ldots, n$, note that

$$\begin{aligned}
\mathbb{P}\big(N(s) = m \bigm| N(t) = n\big) &= \frac{\mathbb{P}\big(N(s) = m, N(t) = n\big)}{\mathbb{P}\big(N(t) = n\big)} \\
&= \frac{\mathbb{P}\big(N(s) = m, N(t) - N(s) + N(s) = n\big)}{\mathbb{P}\big(N(t) = n\big)} \\
&= \frac{\mathbb{P}\big(N(s) = m, N(t) - N(s) = n - m\big)}{\mathbb{P}\big(N(t) = n\big)} \\
&= \frac{\mathbb{P}\big(N(s) = m\big)\,\mathbb{P}\big(N(t) - N(s) = n - m\big)}{\mathbb{P}\big(N(t) = n\big)} \text{ by independent increments} \\
&= \frac{\frac{e^{-\lambda s}(\lambda s)^m}{m!} \cdot \frac{e^{-\lambda(t-s)}(\lambda(t-s))^{n-m}}{(n-m)!}}{\frac{e^{-\lambda t}(\lambda t)^n}{n!}} \\
&= \binom{n}{m} \frac{(\lambda s)^m \big(\lambda(t-s)\big)^{n-m}}{(\lambda t)^m (\lambda t)^{n-m}} \\
&= \binom{n}{m} \left(\frac{s}{t}\right)^m \left(1 - \frac{s}{t}\right)^{n-m},
\end{aligned}$$

which we recognize as the pmf of a BIN$(n, s/t)$ rv.

Comparison of Event Occurrences

Suppose now that $\{N_1(t), t \geq 0\}$ and $\{N_2(t), t \geq 0\}$ are independent Poisson processes at rates λ_1 and λ_2, respectively. Let $S_m^{(1)}$ be the arrival time of the m^{th} event for $\{N_1(t), t \geq 0\}$. Likewise, let $S_n^{(2)}$ be the arrival time of the n^{th} event for $\{N_2(t), t \geq 0\}$.

Based on our knowledge of arrival times, we know that $S_m^{(1)} = \sum_{i=1}^m T_i^{(1)}$ where $\{T_i^{(1)}\}_{i=1}^\infty$ is a sequence of iid EXP(λ_1) rvs. Similarly, $S_n^{(2)} = \sum_{j=1}^n T_j^{(2)}$ where $\{T_j^{(2)}\}_{j=1}^\infty$ is a sequence of iid EXP(λ_2) rvs. Moreover, the sequences $\{T_i^{(1)}\}_{i=1}^\infty$ and $\{T_j^{(2)}\}_{j=1}^\infty$ are independent.

We are interested in the probability that the m^{th} event from the first process happens before the n^{th} event of the second process, or equivalently, $\mathbb{P}(S_m^{(1)} < S_n^{(2)})$.

Before looking at the general case, let us first examine a couple of special cases:

- Take $m = n = 1$:
$$\mathbb{P}(S_1^{(1)} < S_1^{(2)}) = \mathbb{P}(T_1^{(1)} < T_1^{(2)}) = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$
- Take $m = 2$, $n = 1$:
$$\begin{aligned}
\mathbb{P}(S_2^{(1)} < S_1^{(2)}) &= \mathbb{P}(T_1^{(1)} < T_1^{(2)})\, \mathbb{P}(S_2^{(1)} < S_1^{(2)} \mid T_1^{(1)} < T_1^{(2)}) \\
&\quad + \mathbb{P}(T_1^{(1)} > T_1^{(2)})\, \underbrace{\mathbb{P}(S_2^{(1)} < S_1^{(2)} \mid T_1^{(1)} > T_1^{(2)})}_{=0} \\
&= \mathbb{P}(T_1^{(1)} < T_1^{(2)})\, \mathbb{P}(T_1^{(1)} + T_2^{(1)} < T_1^{(2)} \mid T_1^{(1)} < T_1^{(2)}) \\
&= \mathbb{P}(T_1^{(1)} < T_1^{(2)})\, \mathbb{P}(T_1^{(2)} - T_1^{(1)} < T_2^{(1)} \mid T_1^{(1)} < T_1^{(2)}) \\
&= \mathbb{P}(T_1^{(1)} < T_1^{(2)})\, \mathbb{P}(T_1^{(2)} > T_2^{(1)}) \text{ due to the generalized memoryless property} \\
&= \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)\left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right) \\
&= \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^2.
\end{aligned}$$

In the general case, we realize, through the continued application of the memoryless property, that $\mathbb{P}(S_m^{(1)} < S_n^{(1)})$ is equivalent to the probability of observing m "successes" (where the success probability is $\lambda_1/(\lambda_1 + \lambda_2)$) occur before the n "failures" (where the failure probability is $\lambda_2/(\lambda_1 + \lambda_2)$) in a sequence of independent Bernoulli trials.

Specifically, in a sequence of $m + j$ Bernoulli trials (in which m are "successes" and j are "failures"), the $(m + j)^{\text{th}}$ trial must always be a "success" (i.e., the m^{th} one) and the number of "failures" j must be no larger than $n - 1$, which ultimately leads to

$$\mathbb{P}(S_m^{(1)} < S_n^{(1)}) = \sum_{j=0}^{n-1} \binom{m+j-1}{m-1} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^m \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^j. \tag{4.9}$$

Further Properties of the Poisson Process

Example 4.4. (*continued*) At a local insurance company, suppose that fire damage claims come into the company according to a Poisson process at rate 3.8 expected claims per year.

(c) If exactly 12 claims have occurred within the first 5 years, how many claims, on average, occurred within the first 3.5 years? How would this change if no claim history of the first 5 years was given?

Solution: We want to calculate $\mathbb{E}\big[N(3.5) \mid N(5) = 12\big]$. Using the binomial result of Theorem 4.5 with $s = 3.5$, $t = 5$, and $n = 12$, we obtain

$$\mathbb{E}\big[N(3.5) \mid N(5) = 12\big] = 12\left(\frac{3.5}{5}\right) = \frac{42}{5} = 8.4.$$

On the other hand,

$$\mathbb{E}\big[N(3.5)\big] = (3.8)(3.5) = 13.3 \neq 8.4,$$

implying that conditioning on knowledge of $N(5)$ does affect the mean of $N(3.5)$.

(d) At another competing insurance company, suppose that fire damage claims arrive to the company according to a Poisson process with rate 2.9 expected claims per year. What is the probability that 3 claims arrive to this company before 2 claims arrive to the other (first) company? Assume that the insurance companies operate independently of each other.

Solution: Let $N_1(t)$ denote the number of claims arriving to the first company by time t, whereas $N_2(t)$ denotes the number of claims arriving to this new (second) company by time t. We are assuming that $\{N_1(t), t \geq 0\}$ (i.e., a Poisson process with rate $\lambda_1 = 3.8$) and $\{N_2(t), t \geq 0\}$ (i.e., a Poisson process with rate $\lambda_2 = 2.9$) are independent processes. Using (4.9), we are able to calculate

$$\begin{aligned}
&\mathbb{P}(3 \text{ claims arrive to company } 2 \text{ before } 2 \text{ claims arrive to company } 1) \\
&= \mathbb{P}(S_3^{(2)} < S_2^{(1)}) \\
&= 1 - \mathbb{P}(S_2^{(1)} < S_3^{(2)}) \\
&= 1 - \sum_{j=0}^{3-1} \binom{2+j-1}{2-1} \left(\frac{3.8}{3.8+2.9}\right)^2 \left(\frac{2.9}{3.8+2.9}\right)^j \\
&\simeq 0.219.
\end{aligned}$$

Splitting and Merging Poisson Processes

The next property we examine concerns the classification (i.e., splitting) of events from a Poisson process into (potentially) several types.

For a Poisson process $\{N(t), t \geq 0\}$ at rate λ, suppose that events can independently be classified as being one of the k possible types, with probability p_i of being type i, $i = 1, 2, \ldots, k$, with $\sum_{i=1}^k p_i = 1$.

Let $\{N_i(t), t \geq 0\}$ be the associated counting process for type-i events, $i = 1, 2, \ldots, k$. Clearly, by construction, $\sum_{i=1}^k N_i(t) = N(t)$.

First, for $s, t \geq 0$ and $i = 1, 2, \ldots, k$, note that

$$\begin{aligned}
&\mathbb{P}\big(N_i(s+t) - N_i(s) = m_i\big) \\
&= \sum_{n=m_i}^{\infty} \mathbb{P}\big(N_i(s+t) - N_i(s) = m_i \,\big|\, N(s+t) - N(s) = n\big)\, \mathbb{P}\big(N(s+t) - N(s) = n\big) \\
&= \sum_{n=m_i}^{\infty} \binom{n}{m_i} p_i^{m_i} (1-p_i)^{n-m_i}\, \mathbb{P}\big(N(t) = n\big) \\
&\qquad \text{by the stationary increments property of } \{N(t), t \geq 0\} \\
&= \sum_{n=m_i}^{\infty} \mathbb{P}\big(N_i(t) = m_i \,\big|\, N(t) = n\big)\, \mathbb{P}\big(N(t) = n\big) \\
&= \mathbb{P}\big(N_i(t) = m_i\big),
\end{aligned}$$

which proves that $\{N_i(t), t \geq 0\}$ also has the stationary increments property.

Next, suppose that $(s_1, t_1]$ and $(s_2, t_2]$ are non-overlapping time intervals.

For $i = 1, 2, \ldots, k$, note that by the independent increments property of $\{N(t), t \geq 0\}$, the number of events in each of these time intervals, $N(t_1) - N(s_1)$ and $N(t_2) - N(s_2)$, are independent.

Therefore, in combination with the fact that the classification of each event is an independent process, it must hold that the number of type-i events to occur in these intervals, $N_i(t_1) - N_i(s_1)$ and $N_i(t_2) - N_i(s_2)$, are also independent, implying that $\{N_i(t), t \geq 0\}$ possesses the independent increments property too.

Finally, consider

$$
\begin{aligned}
&\mathbb{P}\big(N_1(t) = m_1, N_2(t) = m_2, \ldots, N_k(t) = m_k\big) \\
&= \sum_{n=0}^{\infty} \mathbb{P}\big(N_1(t) = m_1, N_2(t) = m_2, \ldots, N_k(t) = m_k \,\big|\, N(t) = n\big)\, \mathbb{P}\big(N(t) = n\big) \\
&= \underbrace{\mathbb{P}\bigg(N_1(t) = m_1, N_2(t) = m_2, \ldots, N_k(t) = m_k \,\bigg|\, N(t) = \sum_{j=1}^{k} m_j\bigg)}_{\text{MN}\left(\sum_{j=1}^{k} m_j, p_1, p_2, \ldots, p_k\right) \text{ probability}} \mathbb{P}\bigg(N(t) = \sum_{j=1}^{k} m_j\bigg) \\
&= \frac{(m_1 + m_2 + \cdots + m_k)!}{m_1! m_2! \cdots m_k!} p_1^{m_1} p_2^{m_2} \cdots p_k^{m_k} \frac{e^{-\lambda t} (\lambda t)^{m_1 + m_2 + \cdots + m_k}}{(m_1 + m_2 + \cdots + m_k)!} \\
&= \left(\frac{e^{-\lambda p_1 t} (\lambda p_1 t)^{m_1}}{m_1!}\right) \left(\frac{e^{-\lambda p_2 t} (\lambda p_2 t)^{m_2}}{m_2!}\right) \cdots \left(\frac{e^{-\lambda p_k t} (\lambda p_k t)^{m_k}}{m_k!}\right) \\
&= \prod_{i=1}^{k} \mathbb{P}\big(N_i(t) = m_i\big).
\end{aligned}
$$

Thus, $N_1(t), N_2(t), \ldots, N_k(t)$ are independent Poisson random variables.

As a result, we have

$$
\begin{aligned}
\mathbb{P}\big(N_i(h) = 1\big) &= \frac{e^{-\lambda p_i h} (\lambda p_i h)^1}{1!} \\
&= \lambda p_i h \sum_{j=0}^{\infty} \frac{(-\lambda p_i h)^j}{j!} \\
&= \lambda p_i h (1 - \lambda p_i h + \mathrm{o}(h)) \\
&= \lambda p_i h + \mathrm{o}(h),
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{P}\big(N_i(h) \geq 2\big) &= 1 - \mathbb{P}\big(N_i(h) = 0\big) - \mathbb{P}(N_i(h) = 1) \\
&= 1 - \frac{e^{-\lambda p_i h} (\lambda p_i h)^0}{0!} - \lambda p_i h - \mathrm{o}(h) \\
&= 1 - \big(1 - \lambda p_i h + \mathrm{o}(h)\big) - \lambda p_i h - \mathrm{o}(h) \\
&= \mathrm{o}(h).
\end{aligned}
$$

By definition, we have shown that $\{N_1(t), t \geq 0\}, \{N_2(t), t \geq 0\}, \ldots, \{N_k(t), t \geq 0\}$ are independent Poisson processes in which $\{N_i(t), t \geq 0\}$ has rate λp_i, $i = 1, 2, \ldots, k$.

Example 4.4. (*continued*) At a local insurance company, suppose that fire damage claims come into the company according to a Poisson process at rate 3.8 expected claims per year.

(e) Suppose that fire damage claims can be categorized as being either commercial, business, or residential. At the first insurance company, past history suggests that 15% of the claims are commercial, 25% of them are business, and the remaining 60% are residential. Over the course of the next 4 years, what is the probability that the company experiences fewer than 5 claims in each of the 3 categories?

Solution: Let $N_c(t)$ be the number of commercial claims by time t. Likewise, let $N_b(t)$ and $N_r(t)$ represent the number of business and residential claims by time t, respectively. It follows that:

$$\begin{aligned}N_c(4) &\sim \text{POI}(3.8 \cdot 0.15 \cdot 4 = 2.28),\\ N_b(4) &\sim \text{POI}(3.8 \cdot 0.25 \cdot 4 = 3.8),\\ N_r(4) &\sim \text{POI}(3.8 \cdot 0.6 \cdot 4 = 9.12).\end{aligned}$$

We wish to calculate

$$\begin{aligned}&\mathbb{P}\big(N_c(4) < 5, N_b(4) < 5, N_r(4) < 5\big)\\ &= \mathbb{P}\big(N_c(4) < 5\big)\,\mathbb{P}\big(N_b(4) < 5\big)\,\mathbb{P}\big(N_r(4) < 5\big)\\ &\qquad \text{since } N_c(4),\ N_b(4), \text{ and } N_r(4) \text{ are independent}\\ &= \left(\sum_{i=0}^{4} \frac{e^{-2.28}(2.28)^i}{i!}\right)\left(\sum_{i=0}^{4} \frac{e^{-3.8}(3.8)^i}{i!}\right)\left(\sum_{i=0}^{4} \frac{e^{-9.12}(9.12)^i}{i!}\right)\\ &\simeq (0.91857)(0.66784)(0.05105)\\ &\simeq 0.0313.\end{aligned}$$

Remark: It is also possible to merge independent Poisson processes together. In particular, if $\{N_1(t), t \geq 0\}, \{N_2(t), t \geq 0\}, \ldots, \{N_m(t), t \geq 0\}$ are m independent Poisson processes at respective rates $\lambda_1, \lambda_2, \ldots, \lambda_m$, then it is straightforward to show that $\{N(t), t \geq 0\}$, where $N(t) = \sum_{i=1}^{m} N_i(t)$, is also a Poisson process at rate $\sum_{i=1}^{m} \lambda_i$.

Conditional Distribution of Arrival Times

Theorem 4.5 indicated that the conditional distribution of $N(s) \mid \big(N(t) = n\big)$, where $s < t$ is Binomial with n trials and success probability s/t. In other words, it is possible to view each event that occurred within $[0, t]$ as being independent of the others, and the probability of any one event landing within the interval $[0, s]$ as being governed by the cdf of a U$(0, t)$ rv evaluated at s. The idea that we can view s/t as a *uniform probability* is no coincidence. In fact, the following result confirms this notion.

Theorem 4.6. Suppose that $\{N(t), t \geq 0\}$ is a Poisson process at rate λ. Given $N(t) = 1$, the conditional distribution of the first arrival time is uniformly distributed on $(0, t)$ (i.e., $S_1 \mid \big(N(t) = 1\big) \sim$ U$(0, t)$).

Proof: In order to identify the conditional distribution of $S_1 \mid \big(N(t) = 1\big)$, we consider the cdf of $S_1 \mid \big(N(t) = 1\big)$, to be denoted by

$$G(s) = \mathbb{P}\big(S_1 \leq s \,\big|\, N(t) = 1\big),\ 0 < s < t.$$

Note that

$$\begin{aligned}
G(s) &= \mathbb{P}\big(S_1 \le s \,\big|\, N(t) = 1\big) \\
&= \frac{\mathbb{P}\big(S_1 \le s, N(t) = 1\big)}{\mathbb{P}\big(N(t) = 1\big)} \\
&= \frac{\mathbb{P}\big(\text{1 event in } [0, s] \cap \text{0 events in } (s, t]\big)}{\mathbb{P}\big(N(t) = 1\big)} \\
&= \frac{\mathbb{P}\big(N(s) = 1, N(t) - N(s) = 0\big)}{\mathbb{P}\big(N(t) = 1\big)} \\
&= \frac{\mathbb{P}\big(N(s) = 1\big)\,\mathbb{P}\big(N(t) - N(s) = 0\big)}{\mathbb{P}\big(N(t) = 1\big)} \\
&\qquad \text{due to the independent increments property} \\
&= \frac{\frac{e^{-\lambda s}(\lambda s)^1}{1!} \cdot \frac{e^{-\lambda(t-s)}(\lambda(t-s))^0}{0!}}{\frac{e^{-\lambda t}(\lambda t)}{1!}} \\
&= \frac{s}{t}.
\end{aligned}$$

However, this is the cdf of a U$(0, t)$ rv. Thus, $S_1 \mid \big(N(t) = 1\big) \sim$ U$(0, t)$.

Theorem 4.6 specifies how S_1 behaves distributionally when $N(t) = 1$. A natural question to ask is: *How are the n arrival times $S_1, S_2, \ldots, S_n$ distributed if it is known that exactly n arrival times have occurred by time t?*

Before we can address this more general question, we must first familiarize ourselves with some distributional results about *order statistics*.

In what follows, let $\{Y_i\}_{i=1}^n$ be an iid sequence of rvs having a common *continuous* distribution on $(0, \infty)$ with cdf $F(y) = \mathbb{P}(Y_i \le y)$ and pdf $f(y) = F'(y)$ for each $i = 1, 2, \ldots, n$.

Order Statistics

Definition: The sequence of random variables $\{Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}\}$ are called the *order statistics* of $\{Y_1, Y_2, \ldots, Y_n\}$, satisfying:

$$\begin{aligned}
Y_{(1)} &\equiv 1^{\text{st}} \text{ smallest among } \{Y_1, Y_2, \ldots, Y_n\}, \\
Y_{(2)} &\equiv 2^{\text{nd}} \text{ smallest among } \{Y_1, Y_2, \ldots, Y_n\}, \\
&\vdots \\
Y_{(n)} &\equiv n^{\text{th}} \text{ smallest among } \{Y_1, Y_2, \ldots, Y_n\}.
\end{aligned}$$

Remark: By its very definition, we observe that $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}$, and moreover, $Y_{(1)} = \min\{Y_1, Y_2, \ldots, Y_n\}$ and $Y_{(n)} = \max\{Y_1, Y_2, \ldots, Y_n\}$.

We wish to determine the joint distribution of the random vector $(Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)})$. Let the joint cdf of $(Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)})$ be denoted by

$$G(y_1, y_2, \ldots, y_n) = \mathbb{P}(Y_{(1)} \le y_1, Y_{(2)} \le y_2, \ldots, Y_{(n)} \le y_n).$$

Also, define

$$g(y_1, y_2, \ldots, y_n) = \frac{\partial^n G(y_1, \ldots, y_n)}{\partial y_1\, \partial y_2 \cdots \partial y_n}$$

to be the corresponding joint pdf of $\{Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}\}$.

To begin, consider the case when $n = 2$ and assume $0 < y_1 < y_2 < \infty$. Note that

$$\begin{aligned}\mathbb{P}(Y_{(1)} \le y_1, y_1 < Y_{(2)} \le y_2) &= \mathbb{P}(Y_{(1)} \le y_1, Y_{(2)} \le y_2) - \mathbb{P}(Y_{(1)} \le y_1, Y_{(2)} \le y_1)\\ &= G(y_1, y_2) - G(y_1, y_1),\end{aligned}$$

and so

$$\begin{aligned}G(y_1, y_2) &= \mathbb{P}(Y_{(1)} \le y_1, y_1 < Y_{(2)} \le y_2) + G(y_1, y_1)\\ \frac{\partial^2 G(y_1, y_2)}{\partial y_1\,\partial y_2} &= \frac{\partial^2\,\mathbb{P}(Y_{(1)} \le y_1, y_1 < Y_{(2)} \le y_2)}{\partial y_1\,\partial y_2} + \underbrace{\frac{\partial^2 G(y_1, y_1)}{\partial y_1\,\partial y_2}}_{=0}\\ g(y_1, y_2) &= \frac{\partial^2\,\mathbb{P}(Y_{(1)} \le y_1, y_1 < Y_{(2)} \le y_2)}{\partial y_1\,\partial y_2}.\end{aligned}$$

As a result of the above equality, the joint pdf of $Y_{(1)}$ and $Y_{(2)}$ can be obtained by taking the partial derivatives of the quantity $\mathbb{P}(Y_{(1)} \le y_1, y_1 < Y_{(2)} \le y_2)$ with respect to y_1 and y_2. This fact is true for general n (as can be readily verified), so that

$$g(y_1, y_2, \ldots, y_n) = \frac{\partial^n\,\mathbb{P}(Y_{(1)} \le y_1, y_1 < Y_{(2)} \le y_2, \ldots, y_{n-1} < Y_{(n)} \le y_n)}{\partial y_1\,\partial y_2\,\cdots\,\partial y_n},$$

where $0 < y_1 < y_2 < \cdots < y_n < \infty$.

If we now examine the case when $n = 2$ again, with $0 < y_1 < y_2 < \infty$, we see that

$$\begin{aligned}&\mathbb{P}(Y_{(1)} \le y_1, y_1 < Y_{(2)} \le y_2)\\ &= \mathbb{P}\big(Y_{(1)} \le y_1, y_1 < Y_{(2)} \le y_2, \{Y_1 < Y_2\} \cup \{Y_1 > Y_2\}\big)\\ &= \mathbb{P}(Y_{(1)} \le y_1, y_1 < Y_{(2)} \le y_2, Y_1 < Y_2) + \mathbb{P}(Y_{(1)} \le y_1, y_1 < Y_{(2)} \le y_2, Y_1 > Y_2)\\ &= \mathbb{P}(Y_1 \le y_1, y_1 < Y_2 \le y_2, Y_1 < Y_2) + \mathbb{P}(Y_2 \le y_1, y_1 < Y_1 \le y_2, Y_1 > Y_2)\\ &= \mathbb{P}(Y_1 \le y_1, y_1 < Y_2 \le y_2) + \mathbb{P}(Y_2 \le y_1, y_1 < Y_1 \le y_2)\\ &= 2\,\mathbb{P}(Y_1 \le y_1, y_1 < Y_2 \le y_2) \text{ since } Y_1 \text{ and } Y_2 \text{ are identically distributed}\\ &= 2\,\mathbb{P}(Y_1 \le y_1)\,\mathbb{P}(y_1 < Y_2 \le y_2) \text{ since } Y_1 \text{ and } Y_2 \text{ are independent rvs}\\ &= 2F(y_1)\big(F(y_2) - F(y_1)\big).\end{aligned}$$

As a result, we subsequently obtain

$$\begin{aligned}g(y_1, y_2) &= \frac{\partial^2\,\mathbb{P}(Y_{(1)} \le y_1, y_1 < Y_{(2)} \le y_2)}{\partial y_1\,\partial y_2}\\ &= \frac{\partial^2}{\partial y_1\,\partial y_2}\Big[2F(y_1)\big(F(y_2) - F(y_1)\big)\Big]\\ &= 2f(y_1)f(y_2),\ 0 < y_1 < y_2 < \infty.\end{aligned}$$

To verify that this is indeed a joint pdf, note that

$$\int_0^\infty \int_0^\infty g(y_1, y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2 = \int_0^\infty 2f(y_2) \int_0^{y_2} f(y_1)\,\mathrm{d}y_1\,\mathrm{d}y_2 = \int_0^\infty 2f(y_2)F(y_2)\,\mathrm{d}y_2.$$

Let

$$u = F(y_2) \implies \frac{\mathrm{d}u}{\mathrm{d}y_2} = f(y_2) \implies \mathrm{d}u = f(y_2)\,\mathrm{d}y_2,$$

so that

$$\int_0^\infty \int_0^\infty g(y_1, y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2 = 2\int_0^1 u\,\mathrm{d}u = 2\left[\frac{u^2}{2}\right]_{u=0}^{u=1} = 1.$$

Remarks:

(1) The above results can be extended beyond the $n = 2$ case. In fact, it can be shown that the joint pdf of $(Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)})$ is given by

$$g(y_1, y_2, \ldots, y_n) = n! \prod_{i=1}^{n} f(y_i),\ 0 < y_1 < y_2 < \cdots < y_n < \infty, \tag{4.10}$$

the marginal cdf of $Y_{(i)}$, $i = 1, 2, \ldots, n$ is given by

$$G_i(y_i) = \mathbb{P}(Y_{(i)} < y_i) = 1 - \sum_{j=0}^{i-1} \binom{n}{j} F(y_i)^j \big(1 - F(y_i)\big)^{n-j},\ 0 \le y_i < \infty,$$

and the marginal pdf of $Y_{(i)}$ is given by

$$g_i(y_i) = G_i'(y_i) = \frac{n!}{(n-i)!(i-1)!} F(y_i)^{i-1} f(y_i) \big(1 - F(y_i)\big)^{n-i},\ 0 < y_i < \infty. \tag{4.11}$$

(2) If $\{Y_i\}_{i=1}^n$ happens to be a sequence of iid $\mathrm{U}(0, t)$ rvs, then (4.10) and (4.11) simplify to become

$$g(y_1, y_2, \ldots, y_n) = \frac{n!}{t^n},\ 0 < y_1 < y_2 < \cdots < y_n < t, \tag{4.12}$$

and

$$g_i(y_i) = \frac{n! y_i^{i-1} (t - y_i)^{n-i}}{(n-i)!(i-1)! t^n},\ 0 < y_i < t.$$

Conditional Distribution of Arrival Times

With these results in place, we are now in position to state another important result concerning the Poisson process.

Theorem 4.7. Let $\{N(t), t \ge 0\}$ be a Poisson process at rate λ. Given $N(t) = n$, the conditional joint distribution of the n arrival times is identical to the joint distribution of the n order statistics from the $\mathrm{U}(0, t)$ distribution. In other words,

$$(S_1, S_2, \ldots, S_n) \mid \big(N(t) = n\big) \sim (Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}),$$

where $\{Y_i\}_{i=1}^n$ is an iid sequence of $\mathrm{U}(0, t)$ rvs.

Proof: For $0 < s_1 < s_2 < \cdots < s_n < t$, consider

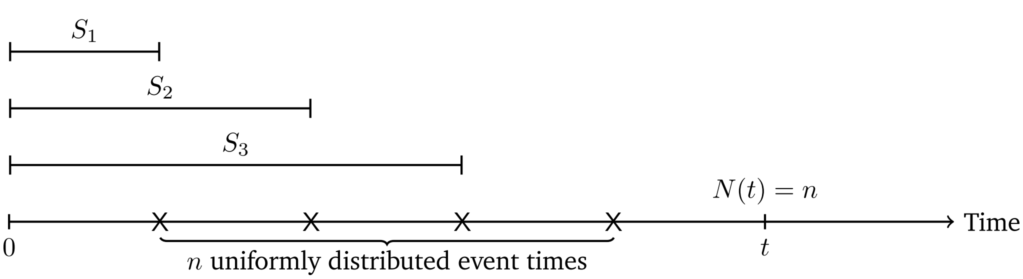
$$\begin{aligned}
&\mathbb{P}\big(S_1 \le s_1, s_1 < S_2 \le s_2, \ldots, s_{n-1} < S_n \le s_n \,\big|\, N(t) = n\big) \\
&= \frac{\mathbb{P}\big(S_1 \le s_1, s_1 < S_2 \le s_2, \ldots, s_{n-1} < S_n \le s_n \,\big|\, N(t) = n\big)}{\mathbb{P}\big(N(t) = n\big)} \\
&= \frac{\mathbb{P}\big(N(s_1) = 1, N(s_2) - N(s_1) = 1, \ldots, N(s_n) - N(s_{n-1}) = 1, N(t) - N(s) = 0\big)}{\mathbb{P}\big(N(t) = n\big)} \\
&= \frac{\mathbb{P}\big(N(s_1) = 1\big)\,\mathbb{P}\big(N(s_2) - N(s_1) = 1\big)\cdots\mathbb{P}\big(N(s_n) - N(s_{n-1}) = 1\big)\,\mathbb{P}\big(N(t) - N(s) = 0\big)}{\mathbb{P}\big(N(t) = n\big)} \\
&\qquad \text{due to the independent increments property of the Poisson process } \{N(t), t \ge 0\} \\
&= \frac{(e^{-\lambda s_1}\lambda s_1)\big(e^{-\lambda(s_2 - s_1)}\lambda(s_2 - s_1)\big)\cdots\big(e^{-\lambda(s_n - s_{n-1})}\lambda(s_n - s_{n-1})\big)\big(e^{-\lambda(t-s)}(\lambda s_1)\big)}{e^{-\lambda t}(\lambda t)^n/n!} \\
&= \frac{n! s_1 (s_2 - s_1)\cdots(s_n - s_{n-1})}{t^n}.
\end{aligned}$$

Thus, the joint pdf of $(S_1, S_2, \ldots, S_n)$ given that $N(t) = n$, can be obtained via differentiation, thereby yielding

$$\frac{\partial^n\, \mathbb{P}\big(S_1 \le s_1, s_1 < S_2 \le s_2, \ldots, s_{n-1} < S_n \le s_n \,\big|\, N(t) = n\big)}{\partial s_1\, \partial s_2 \,\cdots\, \partial s_n} = \frac{n!}{t^n}, \quad 0 < s_1 < s_2 < \cdots < s_n < t.$$

Note that the form above agrees with that of (4.12), and hence the result is proven.

Remark: What this result essentially implies is that under the condition that n events have occurred by time t in a Poisson process, the n times at which those events occur are distributed *independently* and *uniformly* over the interval $[0, t]$.

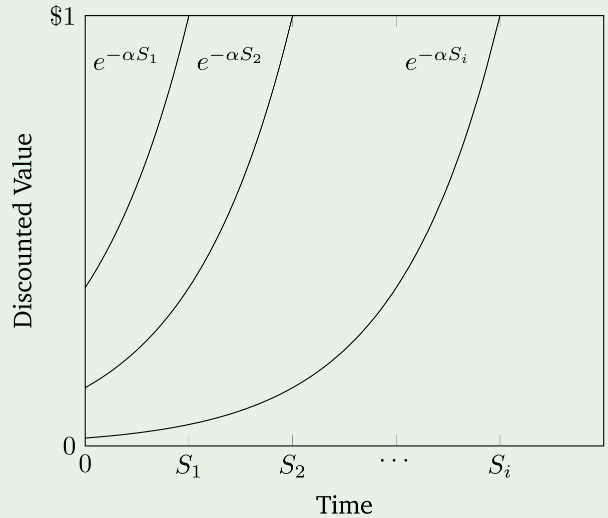


Example 4.5. Cars arrive to a toll bridge according to a Poisson process at rate λ, where each car pays a toll of \$1 upon arrival. Calculate the mean and variance of the total amount collected by time t, *discounted* back to time 0 where $\alpha > 0$ is the discount rate per unit time.

Solution: Let $N(t)$ count the number of cars arriving to the toll bridge by time t, where $N(t) \sim \text{POI}(\lambda t)$. If S_i denotes the arrival time of the i^{th} car of the toll bridge, $i \in \mathbb{Z}^+$, then the discounted value (i.e., back to time 0) of \$1 paid by the i^{th} arrival time is given by

$$1 \cdot e^{-\alpha S_i} = e^{-\alpha S_i},$$

as demonstrated by the following diagram:



If T represents the (discounted) total amount collected by time t, then $T = \sum_{i=1}^{N(t)} e^{-\alpha S_i}$. We wish to find $\mathbb{E}[T]$ and $\mathrm{Var}(T)$. To find $\mathbb{E}[T]$, note that

$$\begin{aligned}\mathbb{E}[T] &= \mathbb{E}\left[\sum_{i=1}^{N(t)} e^{-\alpha S_i}\right] \\ &= \sum_{n=1}^{\infty} \mathbb{E}\left[\sum_{i=1}^{N(t)} e^{-\alpha S_i} \,\middle|\, N(t)=n\right] \mathbb{P}\big(N(t)=n\big) \text{ since } t=0 \text{ if } N(t)=0 \\ &= \sum_{n=1}^{\infty} \mathbb{E}\left[\sum_{i=1}^{n} e^{-\alpha S_i} \,\middle|\, N(t)=n\right] \mathbb{P}\big(N(t)=n\big).\end{aligned}$$

By Theorem 4.7,

$$(S_1, S_2, \ldots, S_n) \mid \big(N(t)=n\big) \sim (Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}),$$

where $(Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)})$ are the n order statistics from $\mathrm{U}(0,t)$ distribution. As a result,

$$\begin{aligned}\mathbb{E}[T] &= \sum_{n=1}^{\infty} \mathbb{E}\left[\sum_{i=1}^{n} e^{-\alpha Y_{(i)}}\right] \mathbb{P}\big(N(t)=n\big) \\ &= \sum_{n=1}^{\infty} \mathbb{E}\left[\sum_{i=1}^{n} e^{-\alpha Y_i}\right] \mathbb{P}\big(N(t)=n\big) \text{ since } \sum_{i=1}^{n} e^{-\alpha Y_{(i)}} = \sum_{i=1}^{n} e^{-\alpha Y_i},\ Y_i \sim \mathrm{U}(0,t)\ \forall i \\ &= \sum_{n=1}^{\infty} n\,\mathbb{E}[e^{-\alpha Y_1}]\,\mathbb{P}\big(N(t)=n\big).\end{aligned}$$

Clearly,

$$\begin{aligned}\mathbb{E}[e^{-\alpha Y_1}] &= \int_0^t e^{-\alpha y} \frac{1}{t}\,\mathrm{d}y \\ &= \frac{1}{t}\left[\frac{-e^{-\alpha y}}{\alpha}\right]_{y=0}^{y=t} \\ &= \frac{1-e^{-\alpha t}}{\alpha t}.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}[T] &= \frac{1-e^{-\alpha t}}{\alpha t} \underbrace{\sum_{n=1}^{\infty} n\,\mathbb{P}\big(N(t)=n\big)}_{\mathbb{E}[N(t)]} \\ &= \frac{1-e^{-\alpha t}}{\alpha t}\lambda t \\ &= \frac{\lambda}{\alpha}(1-e^{-\alpha t}).\end{aligned}$$

To determine $\mathrm{Var}(T)$, we once again apply Theorem 4.7 to first obtain:

$$\begin{aligned}\mathrm{Var}\big(T \mid N(t)=n\big) &= \mathrm{Var}\left(\sum_{i=1}^{n} e^{-\alpha Y_i}\right) \\ &= \mathrm{Var}\left(\sum_{i=1}^{n} e^{-\alpha Y_i}\right) \text{ since } \sum_{i=1}^{n} e^{-\alpha Y_{(i)}} = \sum_{i=1}^{n} e^{-\alpha Y_i} \\ &= \sum_{i=1}^{n} \mathrm{Var}(e^{-\alpha Y_i}) \\ &= n\,\mathrm{Var}(e^{-\alpha Y_1}),\end{aligned}$$

due to the independence and iid nature of $\{Y_i\}_{i=1}^{n}$. Note that

$$\begin{aligned}\mathrm{Var}(e^{-\alpha Y_1}) &= \mathbb{E}\big[(e^{-\alpha Y_1})^2\big] - \big(\mathbb{E}[e^{-\alpha Y_1}]\big)^2 \\ &= \mathbb{E}\big[(e^{-2\alpha Y_1})\big] - \frac{(1-e^{-\alpha t})^2}{\alpha^2 t^2} \\ &= \frac{1-e^{-2\alpha t}}{2\alpha t} - \frac{(1-e^{-\alpha t})^2}{\alpha^2 t^2},\end{aligned}$$

and so

$$\begin{aligned}\mathrm{Var}\big(T \mid N(t)\big) &= \mathrm{Var}\big(T \mid N(t)=n\big)\Big|_{n=N(t)} \\ &= N(t)\left(\frac{1-e^{-2\alpha t}}{2\alpha t} - \frac{(1-e^{-\alpha t})^2}{\alpha^2 t^2}\right).\end{aligned}$$

Finally, applying the conditional variance formula, we get

$$\begin{aligned}\mathrm{Var}(T) &= \mathrm{Var}\Big(\mathbb{E}\big[T \mid N(t)\big]\Big) + \mathbb{E}\Big[\mathrm{Var}\big(T \mid N(t)\big)\Big] \\ &= \mathrm{Var}\left(N(t)\cdot\frac{1-e^{-\alpha t}}{\alpha t}\right) + \mathbb{E}\left[N(t)\left(\frac{1-e^{-2\alpha t}}{2\alpha t} - \frac{(1-e^{-\alpha t})^2}{\alpha^2 t^2}\right)\right] \\ &= \frac{(1-e^{-\alpha t})^2}{\alpha^2 t^2}\underbrace{\mathrm{Var}\big(N(t)\big)}_{=\lambda t} + \left(\frac{1-e^{-2\alpha t}}{2\alpha t} - \frac{(1-e^{-\alpha t})^2}{\alpha^2 t^2}\right)\underbrace{\mathbb{E}\big[N(t)\big]}_{=\lambda t} \\ &= \frac{\lambda}{2\alpha}(1-e^{-2\alpha t}).\end{aligned}$$

Example 4.6. Satellites are launched at times according to a Poisson process at rate 3 per year. During the past year, it was observed that only two satellites were launched. What is the joint probability that the first of these two satellites was launched in the first 5 months of the year and the second satellite was launched prior to the last 2 months of the year?

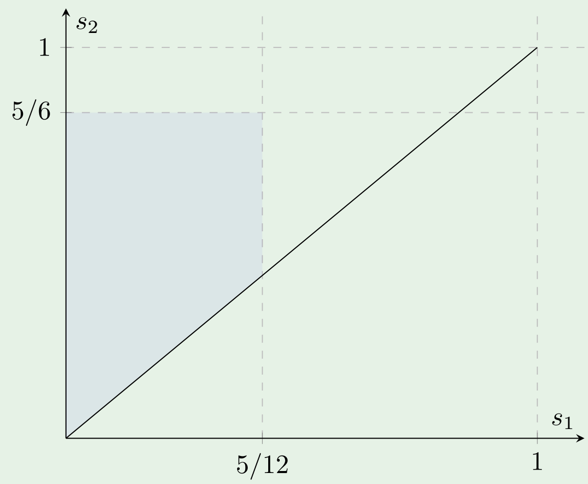
Solution: Let $\{N(t), t \geq 0\}$ be the Poisson process at rate $\lambda = 3$ governing satellite launches. We are interested in calculating

$$\mathbb{P}\left(S_1 \leq \frac{5}{12}, S_2 \leq \frac{5}{6} \,\middle|\, N(1) = 2\right).$$

To do so, we use Theorem 4.7 (specifically, (4.12)), to obtain the joint conditional pdf of $(S_1, S_2) \mid \big(N(1) = 2\big)$ as

$$g(s_1, s_2) = \frac{2!}{1^2} = 2,\ 0 < s_1 < s_2 < 1,$$

which we integrate over the shaded region:



Thus,

$$\begin{aligned}
\mathbb{P}\left(S_1 \leq \frac{5}{12}, S_2 \leq \frac{5}{6} \,\middle|\, N(1) = 2\right) &= \int_0^{5/12} \int_{s_1}^{5/6} g(s_1, s_2)\, \mathrm{d}s_2\, \mathrm{d}s_1 \\
&= \int_0^{5/12} \int_{s_1}^{5/6} 2\, \mathrm{d}s_2\, \mathrm{d}s_1 \\
&= 2 \int_0^{5/12} \left(\frac{5}{6} - s_1\right) \mathrm{d}s_1 \\
&= 2 \left[\frac{5}{6} s_1 - \frac{s_1^2}{2}\right]_{s_1=0}^{s_1=5/12} \\
&= 2 \left(\frac{25}{72} - \frac{25}{288}\right) \\
&= 2 \times \frac{25}{96} \\
&= \frac{25}{48} \\
&\simeq 0.521.
\end{aligned}$$

4.4 Two Important Generalizations

The Non-homogeneous Poisson Process

Oftentimes, we find the Poisson process difficult to apply in applications of real-life phenomena, largely due to the fact that the Poisson process assumes a *constant* arrival rate of λ for all time. In what follows, we consider a more general type of process in which the arrival rate is allowed to vary as a function of time.

Definition: The counting process $\{N(t), t \geq 0\}$ is a *non-homogeneous* (or non-stationary) Poisson process with *rate function* $\lambda(t)$ if the following three conditions hold true:

(1) $\{N(t), t \geq 0\}$ has independent increments.

(2) $\mathbb{P}\big(N(t+h) - N(t) = 1\big) = h\lambda(t) + \mathrm{o}(h)$.

(3) $\mathbb{P}\big(N(t+h) - N(t) \geq 2\big) = \mathrm{o}(h)$.

Many applications that generate random points in time are modelled more realistically with non-homogeneous processes. For instance:

- The rate of customers entering a supermarket is not the same during the entire day.
- The average arrival rate of vehicles on a highway fluctuates between its maximum during rush hours and its minimum during low traffic times.

The mathematical cost of this generalization, however, is that we lose the stationary increments property.

For $s_1, s_2 \geq 0$, the following theorem specifies the distribution of $N(s_1+s_2) - N(s_1)$.

Theorem 4.8. If $\{N(t), t \geq 0\}$ is a non-homogeneous Poisson process with rate function $\lambda(t)$, then $N(s_1+s_2) - N(s_1) \sim \text{POI}\big(m(s_1+s_2) - m(s_1)\big)$ where the *mean value function* $m(t)$ is given by

$$m(t) = \int_0^t \lambda(\tau)\,\mathrm{d}\tau,\ t \geq 0.$$

Proof: Let $\phi_u(s_1, s_2) = \mathbb{E}\Big[e^{u\left(N(s_1+s_2)-N(s_1)\right)}\Big]$, which is the mgf of $N(s_1+s_2) - N(s_1)$, where u serves as the argument of the mgf. For $h \geq 0$, we first note that

$$\begin{aligned}
\phi_u(s_1, s_2+h) &= \mathbb{E}\Big[e^{u\left(N(s_1+s_2+h)-N(s_1)\right)}\Big] \\
&= \mathbb{E}\Big[e^{u\left(N(s_1+s_2+h)-N(s_1+s_2)+N(s_1+s_2)-N(s_1)\right)}\Big] \\
&= \mathbb{E}\Big[e^{u\left(N(s_1+s_2+h)-N(s_1+s_2)\right)}e^{u\left(N(s_1+s_2)-N(s_1)\right)}\Big] \\
&= \mathbb{E}\Big[e^{u\left(N(s_1+s_2+h)-N(s_1+s_2)\right)}\Big]\,\mathbb{E}\big[e^{u\left(N(s_1+s_2)-N(s_1)\right)}\big] \text{ due to independent increments} \\
&= \phi_u(s_1+s_2, h)\phi_u(s_1, s_2). \qquad (4.13)
\end{aligned}$$

Applying a similar approach to that used in the proof of Theorem 4.3, it can be shown that (4.13) ultimately give rise to the first-order differential equation (see Exercise 4.4.1).

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}s_2}\phi_u(s_1,s_2) &= \lambda(s_1+s_2)(e^u-1)\phi_u(s_1,s_2)\\
\frac{\mathrm{d}}{\mathrm{d}t}\phi_u(s_1,t) &= \lambda(s_1+t)(e^u-1)\phi_u(s_1,t) \text{ let } s_2=t\\
\frac{\frac{\mathrm{d}}{\mathrm{d}t}\phi_u(s_1,t)}{\phi_u(s_1,t)} &= \lambda(s_1+t)(e^u-1)\\
\frac{\mathrm{d}}{\mathrm{d}t}\ln\big(\phi_u(s_1,t)\big) &= \lambda(s_1+t)(e^u-1)\\
\mathrm{d}\ln\big(\phi_u(s_1,t)\big) &= \lambda(s_1+t)(e^u-1)\,\mathrm{d}t\\
\int_0^{s_2}\mathrm{d}\ln\big(\phi_u(s_1,t)\big) &= \int_0^{s_2}\lambda(s_1+t)(e^u-1)\,\mathrm{d}t\\
\Big[\ln\big(\phi_u(s_1,t)\big)\Big]_{t=0}^{t=s_2} &= (e^u-1)\int_0^{s_2}\lambda(s_1+t)(e^u-1)\,\mathrm{d}t\\
\ln\big(\phi_u(s_1,s_2)\big)-\ln\big(\phi_u(s_1,0)\big) &= (e^u-1)\int_{s_1}^{s_1+s_2}\lambda(\tau)\,\mathrm{d}\tau.
\end{aligned}$$

However,

$$\phi_u(s_1,0)=\mathbb{E}\Big[e^{u\big(N(s_1)-N(s_1)\big)}\Big]=\mathbb{E}[e^{u\cdot 0}]=\mathbb{E}[e^0]=1.$$

$$\begin{aligned}
\ln\big(\phi_u(s_1,s_2)\big) &= (e^u-1)\int_{s_1}^{s_1+s_2}\lambda(\tau)\,\mathrm{d}\tau\\
\phi_u(s_1,s_2) &= e^{(e^u-1)\int_{s_1}^{s_1+s_2}\lambda(\tau)\,\mathrm{d}\tau},\ u\in\mathbb{R}.
\end{aligned}$$

Since

$$\begin{aligned}
\int_{s_1}^{s_1+s_2}\lambda(\tau)\,\mathrm{d}\tau &= \int_0^{s_1+s_2}\lambda(\tau)\,\mathrm{d}\tau-\int_0^{s_1}\lambda(\tau)\,\mathrm{d}\tau\\
&= m(s_1+s_2)-m(s_1).
\end{aligned}$$

Thus, we have

$$\phi_u(s_1,s_2)=e^{(e^u-1)\big(m(s_1+s_2)-m(s_1)\big)},\ u\in\mathbb{R},$$

which is the mgf of a POI$\big(m(s_1+s_2)-m(s_1)\big)$ rv. By the mgf uniqueness property,

$$N(s_1+s_2)-N(s_1)\sim \text{POI}\big(m(s_1+s_2)-m(s_1)\big).$$

Remarks:

(1) As a direct consequence of Theorem 4.8, for all $s,t\geq 0$, we have

$$\mathbb{P}\big(N(s+t)-N(s)=n\big)=e^{-\big(m(s+t)-m(s)\big)}\frac{\big(m(s+t)-m(s)\big)^n}{n!},\ n=0,1,2,\ldots.$$

(2) If the rate function $\lambda(\tau)=\lambda\ \forall\tau\geq 0$, then note that

$$\int_s^{s+t}\lambda(\tau)\,\mathrm{d}\tau=\int_s^{s+t}\lambda\,\mathrm{d}\tau=\lambda(s+t-s)=\lambda t$$

and

$$\mathbb{P}\big(N(s+t)-N(s)=n\big)=\frac{e^{-\lambda t}(\lambda t)^n}{n!},\ n=0,1,2,\ldots.$$

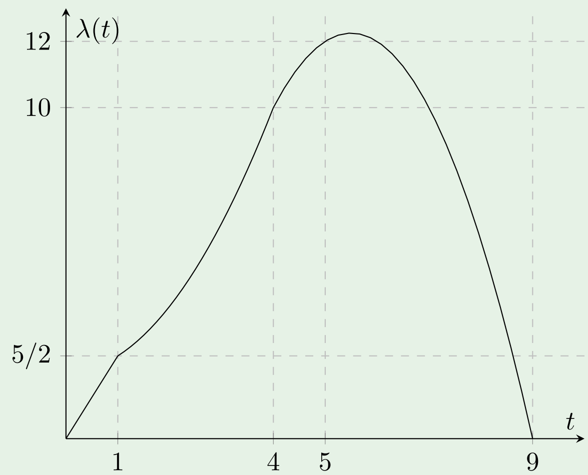
This is expected, since $\{N(t), t \geq 0\}$ simplifies to become the standard (i.e., stationary) Poisson process.

Example 4.7. Requests for technical support within the statistics department occur according to a non-homogeneous Poisson process $\{N(t), t \geq 0\}$ having rate function

$$\lambda(t) = \begin{cases} 5t/2, & \text{if } 0 \leq t < 1, \\ t^2/2 + 2, & \text{if } 1 \leq t < 4, \\ (9-t)(t-2), & \text{if } 4 \leq t \leq 9, \end{cases}$$

where t is measured in hours from the start of the workday. What is the probability that four requests occur in the first two hours of the workday and ten more occur in the final two hours of the workday?

Solution: The plot below provides a visual depiction of the rate function in use:



We wish to calculate

$$\mathbb{P}\big(N(2) = 4, N(9) - N(7) = 10\big).$$

Using the independent increments property, we first have

$$\mathbb{P}\big(N(2) = 4, N(9) - N(7) = 10\big) = \mathbb{P}\big(N(2) = 4\big)\, \mathbb{P}(N(9) - N(7) = 10).$$

Now, we need to calculate

$$\begin{aligned} m(2) - m(0) &= \int_0^2 \lambda(t)\,\mathrm{d}t \\ &= \int_0^1 \frac{5t}{2}\,\mathrm{d}t + \int_1^2 \frac{t^2}{2} + 2\,\mathrm{d}t \\ &= \left[\frac{5t^2}{4}\right]_{t=0}^{t=1} + \left[\frac{t^3}{6} + 2t\right]_{t=1}^{t=2} \\ &= \frac{53}{12}. \end{aligned}$$

Also,

$$\begin{aligned} m(9)-m(7) &= \int_7^9 \lambda(t)\,\mathrm{d}t \\ &= \int_7^9 (9-t)(t-2)\,\mathrm{d}t \\ &= \int_7^9 (-18+11t-t^2)\,\mathrm{d}t \\ &= \left[-18t+\frac{11t^2}{t}-\frac{t^3}{3}\right]_{t=7}^{t=9} \\ &= \frac{34}{3}. \end{aligned}$$

As a result, it follows that

$$\begin{aligned} \mathbb{P}\big(N(2)=4\big) &= e^{-\left(m(2)-m(0)\right)}\frac{\big(m(2)-m(0)\big)^4}{4!} \\ &= \frac{e^{-53/12}(53/12)^4}{4!}, \end{aligned}$$

and

$$\begin{aligned} \mathbb{P}\big(N(9)-N(7)=10\big) &= e^{-\left(m(9)-m(7)\right)}\frac{\big(m(9)-m(7)\big)^{10}}{10!} \\ &= \frac{e^{-34/3}(34/3)^{10}}{10!}. \end{aligned}$$

Finally,

$$\begin{aligned} \mathbb{P}\big(N(2)=4, N(9)-N(7)=10\big) &= \mathbb{P}\big(N(2)=4\big)\,\mathbb{P}(N(9)-N(7)=10) \\ &= \frac{e^{-53/12}(53/12)^4}{4!}\cdot\frac{e^{-34/3}(34/3)^{10}}{10!} \\ &\simeq 0.0221. \end{aligned}$$

The Compound Poisson Process

Another restriction we might wish to relax concerns the assumption that arrivals in a Poisson process occur strictly one after the other.

For instance, vehicles crossing the Canada-USA border would usually have more than just one passenger on board. Individuals arriving to a sporting event often arrive in groups of two or more.

There are also applications where each arrival might generate a random monetary amount. For example, in an insurance company where claims occur according to a Poisson process, the claim sizes themselves are (random) amounts of money, and one would typically be interested in the total amount of money paid out by the insurance company by time t.

In any of the above scenarios, each arrival in a Poisson process comes with an associated real-valued rv that represents the *value* of the arrival in a sense.

This gives rise to the following definition.

Definition: Let $\{Y_i\}_{i=1}^{\infty}$ be an iid sequence of rvs. Let $\big\{N(t), t\geq 0\big\}$ be a Poisson process at rate λ, independent of each Y_i, $i=1,2,3,\ldots$. If $X(t)=\sum_{i=1}^{N(t)} Y_i$, then the process $\big\{X(t), t\geq 0\big\}$ is a *compound*

Poisson process.

Remarks:

(1) The above definition is another way of generalizing the Poisson process, since if we choose the distribution of each Y_i to be degenerate at 1 (i.e., $Y_i = 1$ with probability 1 $\forall i \in \mathbb{Z}^+$), then $\{X(t), t \geq 0\}$ and $\{N(t), t \geq 0\}$ are identical processes.

(2) The compound Poisson process $\{X(t), t \geq 0\}$ inherits the independent and stationary increments properties from the Poisson processes $\{N(t), t \geq 0\}$. To see this formally, let $(s_1, t_1]$ and $(s_2, t_2]$ be time intervals such that $t_1 \leq s_2$ (resulting in $(s_1, t_1] \cap (s_2, t_2] = \emptyset$). Making use of the assumptions concerning $\{Y_i\}_{i=1}^{\infty}$ and $\{N(t), t \geq 0\}$, note that

$$
\begin{aligned}
&\mathbb{P}\big(X(t_1) - X(s_1) \leq a_1, X(t_2) - X(s_2) \leq a_2\big) \\
&= \mathbb{P}\Bigg(\sum_{i=1}^{N(t_1)} Y_i - \sum_{i=1}^{N(s_1)} Y_i \leq a_1, \sum_{i=1}^{N(t_2)} Y_i - \sum_{i=1}^{N(s_2)} Y_i \leq a_2\Bigg) \\
&= \mathbb{P}\Bigg(\sum_{i=N(s_1)+1}^{N(t_1)} Y_i \leq a_1, \sum_{i=N(s_2)+1}^{N(t_2)} Y_i \leq a_2\Bigg) \\
&= \sum_{m_1=0}^{\infty}\sum_{n_1=0}^{\infty}\sum_{m_2=0}^{\infty}\sum_{n_2=0}^{\infty} \\
&\quad \mathbb{P}\Bigg(\sum_{i=N(s_1)+1}^{N(t_1)} Y_i \leq a_1, \sum_{i=N(s_2)+1}^{N(t_2)} Y_i \leq a_2 \;\Bigg|\; \begin{matrix} N(s_1) = m_1, N(t_1) - N(s_1) = n_1, \\ N(s_2) - N(t_1) = m_2, N(t_2) - N(s_2) = n_2 \end{matrix}\Bigg) \\
&\quad \times \mathbb{P}\big(N(s_1) = m_1, N(t_1) - N(s_1) = n_1, N(s_2) - N(t_1) = m_2, N(t_2) - N(s_2) = n_2\big) \\
&= \sum_{m_1=0}^{\infty}\sum_{n_1=0}^{\infty}\sum_{m_2=0}^{\infty}\sum_{n_2=0}^{\infty} \mathbb{P}\Bigg(\sum_{i=m_1+1}^{m_1+n_1} Y_i \leq a_1, \sum_{i=m_1+n_1+m_2+1}^{m_1+n_1+m_2+n_2} Y_i \leq a_2\Bigg) \\
&\quad \times \mathbb{P}\big(N(s_1) = m_1\big)\,\mathbb{P}\big(N(t_1) - N(s_1) = n_1\big)\,\mathbb{P}\big(N(s_2) - N(t_1) = m_2\big)\,\mathbb{P}\big(N(t_2) - N(s_2) = n_2\big) \\
&= \Bigg\{\sum_{n_1=0}^{\infty} \mathbb{P}\Big(\sum_{i=1}^{n_1} Y_i \leq a_1\Big)\,\mathbb{P}\big(N(t_1 - s_1) = n_1\big)\Bigg\}\Bigg\{\sum_{n_2=0}^{\infty} \mathbb{P}\Big(\sum_{i=1}^{n_2} Y_i \leq a_2\Big)\,\mathbb{P}\big(N(t_2 - s_2) = n_2\big)\Bigg\} \\
&= \mathbb{P}\Bigg(\sum_{i=1}^{N(t_1-s_1)} Y_i \leq a_1\Bigg)\,\mathbb{P}\Bigg(\sum_{i=1}^{N(t_2-s_2)} Y_i \leq a_2\Bigg) \\
&= \mathbb{P}\big(X(t_1 - s_1) \leq a_1\big)\,\mathbb{P}\big(X(t_2 - s_2) \leq a_2\big).
\end{aligned}
$$

(3) For $t > 0$, determining the probability distribution of $X(t)$ is, generally speaking, a challenging mathematical problem. On the other hand, the mean and variance of $X(t)$ are readily accessible. In particular, by making use of the results for the mean and variance of a random sum (see Example 2.9), we immediately obtain

$$
\mathbb{E}\big[X(t)\big] = \mathbb{E}\big[N(t)\big]\,\mathbb{E}[Y_1] = \lambda t\,\mathbb{E}[Y_1]
$$

and

$$
\begin{aligned}
\mathrm{Var}\big(X(t)\big) &= \mathrm{Var}(Y_1)\,\mathbb{E}\big[N(t)\big] + \mathbb{E}[Y_1]^2\,\mathrm{Var}\big(N(t)\big) \\
&= \lambda t\big(\mathrm{Var}(Y_1) - \mathbb{E}[Y_1]^2\big) \\
&= \lambda t\,\mathbb{E}[Y_1^2].
\end{aligned}
$$

Example 4.8. Claims received by an insurance company occur according to a Poisson process at a rate of 30 claims per year. Individual claim amounts, which are assumed to be independent, are known to be either \$1000, \$2000, or \$3000. Company records indicate that one year ago, the average total amount paid out was \$56000 and that the standard deviation of the total amount paid out was \$11000. Based on this information, how likely was it that an individual claim of \$3000 took place last year?

Solution: Let $N(t)$ represent the number of claims received by the company by time t (measured in years), which is known to have a POI($30t$) distribution in the past year. In addition, let Y_i, $i \in \mathbb{Z}^+$, be the size of the i^{th} individual claim amount (measured in thousands of dollars), having pmf of the form

$$\mathbb{P}(Y_i = y) = \begin{cases} \alpha, & \text{if } y = 1, \\ \beta, & \text{if } y = 2, \\ 1 - \alpha - \beta, & \text{if } y = 3. \end{cases}$$

Therefore, $X(t) = \sum_{i=1}^{N(t)} Y_i$ represents the total amount paid out (measured in thousands of dollars) by time t. We have:

$$\mathbb{E}\big[X(1)\big] = 30\,\mathbb{E}[Y_1] = 56 \implies \mathbb{E}[Y_1] = \frac{56}{30} = \frac{28}{15}.$$

That is,

$$\alpha + 2\beta + 3(1 - \alpha - \beta) = \frac{28}{15} \implies 30\alpha + 15\beta = 17. \tag{4.14}$$

Furthermore,

$$\mathrm{Var}\big(X(1)\big) = (30)\,\mathbb{E}[Y_1^2] = 11^2 \implies \mathbb{E}[Y_1^2] = \frac{121}{30}.$$

That is,

$$\alpha + 4\beta + 9(1 - \alpha - \beta) = \frac{121}{30} \implies 240\alpha + 150\beta = 149. \tag{4.15}$$

The above equations (4.14) and (4.15) are linear, and can be solved to find α and β as follows:

$$10 \times (4.14) - (4.15) \implies 60\alpha = 21 \implies \alpha = \frac{7}{20}. \tag{4.16}$$

Plugging (4.16) into (4.14) leads to

$$30 \times \frac{7}{20} + 15\beta = 7 \implies \beta = \frac{13}{30}.$$

From above, the desired probability is simply

$$1 - \alpha - \beta = 1 - \frac{7}{20} - \frac{13}{30} = \frac{13}{60} \simeq 0.217.$$

Appendix A

Useful Documents

A.1 List of Acronyms

iff	if and only if
gcd	greatest common divisor
rv	random variable
pmf	probability mass function
pdf	probability density function
cdf	cumulative distribution function
tpf	tail probability function
mgf	moment generating function
iid	independent and identically distributed
SLLN	Strong Law of Large Numbers
a.s.	almost surely
DTMC	discrete-time Markov chain
TPM	transition probability matrix
BLT	Basic Limit Theorem (for DTMCs)

A.2 Special Symbols

$\rightarrow$	approaches
$\Longrightarrow$	implies
Ω	sample space for a probability model
$\mathbb{P}(\,\cdot\,)$	probability function
$\varnothing$	null event (or empty set)
$\cup$	union operator
$\cap$	intersection operator
A^c	complement of A
$\subseteq$	is a subset of
$\mathbb{R}$	set of all real numbers
$\mathbb{Z}$	set of all integers $\{0, \pm 1, \pm 2, \ldots\}$
$\mathbb{Z}^+$	set of positive integers $\{1, 2, \ldots\}$
$\mathbb{N}$	set of non-negative integers $\{0, 1, 2, \ldots\}$
$\mathcal{S}$	state space of a rv (or a DTMC)
$\approx$	approximately equal to
$\sim$	has the probability distribution of
$\mathbb{E}[\,\cdot\,]$	expected value operator
$\underline{a}$	row vector notation
$\underline{a}^\top$	column vector notation
$[A_{i,j}]$	matrix A with the elements of the form $A_{i,j}$
I	identity matrix (of appropriate dimension)
$\mathbf{0}$	zero matrix (of appropriate dimension)
$\underline{e}^\top$	vector of ones (of appropriate dimension)
$\mathcal{T}$	index set of a stochastic process
$n!$	n factorial
$\binom{n}{x}$	n choose x
$(n)_x$	n taken to x terms
$\delta_{i,j}$	1 if $i = j$, 0 if $i \neq j$ (Kronecker delta)
$\lvert x \rvert$	absolute value of x
$\lfloor x \rfloor$	greatest integer less than or equal to x
$\exp\{x\}$	exponential function e^x

A.3 Results for Some Fundamental Probability Distributions

Discrete Distribution	Probability Mass Function of X	Mean $\mathbb{E}[X]$	Variance $\text{Var}(X)$
DU(a,b)	$p(x) = \frac{1}{b-a+1}$, $x = a, a+1, \ldots, b$	$\frac{a+b}{2}$	$\frac{(b-a)(b-a+2)}{12}$
BIN(n,p)	$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, \ldots, n$	np	$np(1-p)$
BERN(p)	$p(x) = p^x(1-p)^{1-x}$, $x = 0, 1$	p	$p(1-p)$
HG(N,r,n)	$p(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$, $x = \max\{0, n-N+r\}, \ldots, \min\{n, r\}$	$\frac{nr}{N}$	$\frac{nr(N-r)(N-n)}{N^2(N-1)}$
POI(λ)	$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, $x = 0, 1, 2, \ldots$	λ	λ
NB$_t(k,p)$	$p(x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}$, $x = k, k+1, k+2, \ldots$	$\frac{k}{p}$	$\frac{k(1-p)}{p^2}$
GEO$_t(p)$	$p(x) = (1-p)^{x-1}p$, $x = 1, 2, 3, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
NB$_f(k,p)$	$p(x) = \binom{x+k-1}{k-1} p^k (1-p)^x$, $x = 0, 1, 2, \ldots$	$\frac{k(1-p)}{p}$	$\frac{k(1-p)}{p^2}$
GEO$_f(p)$	$p(x) = (1-p)^x p$, $x = 0, 1, 2, \ldots$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$
Continuous Distribution	**Probability Mass Function of X**	**Mean $\mathbb{E}[X]$**	**Variance $\text{Var}(X)$**
U(a,b)	$f(x) = \frac{1}{b-a}$, $a < x < b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Beta(m,n)	$f(x) = \frac{(m+n-1)!}{(m-1)!(n-1)!} x^{m-1}(1-x)^{n-1}$, $0 < x < 1$	$\frac{m}{m+n}$	$\frac{mn}{(m+n)^2(m+n+1)}$
Erlang(n,λ)	$f(x) = \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}$, $x > 0$	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$
EXP(λ)	$f(x) = \lambda e^{-\lambda x}$, $x > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

"DU"	stands for	*Discrete Uniform*	"BIN"	stands for	*Binomial*
"BERN"	stands for	*Bernoulli*	"HG"	stands for	*Hypergeometric*
"POI"	stands for	*Poisson*	"NB$_t$"	stands for	*Negative Binomial (for trials)*
"GEO$_t$"	stands for	*Geometric (for trials)*	"NB$_f$"	stands for	*Negative Binomial (for failures)*
"GEO$_f$"	stands for	*Geometric (for failures)*	"U"	stands for	*(Continuous) Uniform*
"EXP"	stands for	*Exponential*			