Stat 333

APPLIED PROBABILITY

University of Waterloo

Course notes by: TC Fraser
Instructor: Yi Shen

tcfraser@tcfraser.com

Version: 1.0

Table Of Contents

Disclaimer

These notes are intended to be a reference for my future self (TC Fraser). If you the reader find these notes useful in any capacity, please feel free to use these notes as you wish, free of charge. However, I do not guarantee their complete accuracy and mistakes are likely present. If you notice any errors please email me at **tcfraser@tcfraser.com**, or contribute directly at **https://github.com/tcfraser/course-notes**. If you are the professor of this course (Yi Shen) and you've managed to stumble upon these notes and would like to make large changes or additions, email me please.

Latest versions of all my course notes are available at **www.tcfraser.com/coursenotes**.

1 Review

If $X \perp\!\!\!\perp Y$ then $\mathsf{Cov}(X, Y) = 0$ and,

$$\mathsf{Var}(X + Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$$

We see that independence implies uncorrelated, but uncorrelation does not imply independence.

Remark 1. We have that,

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) \tag{1.1}$$

If $X \perp\!\!\!\perp Y$ then we also have that,

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) \tag{1.2}$$

and that,

$$\mathsf{Var}(XY) = \mathsf{Var}(X)\mathsf{Var}(Y) \tag{1.3}$$

It is important to remember that the first result (eq. (1.1)) and the other two results (eqs. (1.2) and (1.3)) have a very different natures. The first is a consequence of the linearity in the definition of expectation and holds unconditionally. However eqs. (1.2) and (1.3) require that $X \perp\!\!\!\perp Y$. As such it is more appropriate to consider eqs. (1.2) and (1.3) as properties of independence rather than the properties of expectation and variance.

1.1 Indicator

A r.v. $\mathbf{1}$ is called an **indicator** for an event A if,

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$$

The most important property of the indicator random variable is that the expectation of $\mathbf{1}_A$ is the same as the probability of the event A,

$$\mathbb{E}(\mathbf{1}_A) = P(A)$$

Proof. Since $\mathbf{1}_A$ is a Bernoulli random variable, the proof is easy. Consider,

$$\begin{aligned} P(\mathbf{1}_A = 1) &= P(\{\omega : \mathbf{1}_A(\omega) = 1\}) \\ &= P(\{\omega : \omega \in A\}) \\ &= P(A) \end{aligned}$$

Therefore the expectation of $\mathbf{1}_A$ must be,

$$\mathbb{E}(\mathbf{1}_A) = 1 \cdot P(\mathbf{1}_A = 1) + 0 \cdot P(\mathbf{1}_A = 0) = P(\mathbf{1}_A = 1) = P(A)$$

□

Example 1. We see $\mathbf{1}_A$ is just a Bernoulli random variable,

$$\mathbf{1}_A \sim \mathsf{Ber}(P(A))$$

Example 2. Let $X \sim \mathsf{Bin}(n, p)$; X is the number of successes in n Bernoulli trials, each with a probability p of success.

$$X = \mathbf{1}_1 + \cdots + \mathbf{1}_n \tag{1.4}$$

Where $\{\mathbf{1}_1, \ldots, \mathbf{1}_n\}$ are indicators for independent events. $\mathbf{1}_i = 1$ if the i-th trial is a success and $\mathbf{1}_i = 0$ if the i-th trial is a failure. Hence, I_i are **iid** (independent and identically distributed) r.v.s. It is known that the expectation of X is given by,

$$\mathbb{E}(X) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$$

However eq. (1.4) yields the following approach,

$$\begin{aligned}\mathbb{E}(X) &= \mathbb{E}(\mathbf{1}_1 + \cdots + \mathbf{1}_n)\\ &= \mathbb{E}(\mathbf{1}_1) + \cdots + \mathbb{E}(\mathbf{1}_n)\\ &= n\mathbb{E}(\mathbf{1}_1)\\ &= np\end{aligned}$$

Moreover,

$$\begin{aligned}\mathsf{Var}(X) &= \mathsf{Var}(\mathbf{1}_1 + \cdots + \mathbf{1}_n)\\ &= \mathsf{Var}(\mathbf{1}_1) + \cdots + \mathsf{Var}(\mathbf{1}_n) \quad \text{Independence}\\ &= n\mathsf{Var}(\mathbf{1}_1)\\ &= np(1-p)\end{aligned}$$

The variance $\mathsf{Var}(I_1)$ is given by,

$$\mathsf{Var}(I_1) = \mathbb{E}\left(I_1^2\right) - (\mathbb{E}(I_1))^2$$

But notice that $I_1^2 = I_1$ is idempotent. Therefore,

$$\mathsf{Var}(I_1) = p - p^2 = p(1-p)$$

Example 3. Let X be a r.v. taking values in non-negative integers $\{0, 1, 2\ldots\}$. Then we find that the expectation of X is given by,

$$\mathbb{E}(X) = \sum_{n=0}^{\infty} P(X > n)$$

Note that,

$$X = \sum_{n=0}^{\infty} \mathbf{1}_n$$

Where notationally $\mathbf{1}_n \equiv \mathbf{1}_{\{X>n\}}$. The intuition being that if $X = 3$, then $X = 1 + 1 + 1$ since $X = \underbrace{\mathbf{1}_0 + \mathbf{1}_1 + \mathbf{1}_2}_{3} + \underbrace{\mathbf{1}_3}_{0} + \ldots$.

$$\begin{aligned}\mathbb{E}(X) &= \mathbb{E}\left(\sum_{n=0}^{\infty} \mathbf{1}_n\right)\\ &= \sum_{n=0}^{\infty} \mathbb{E}(\mathbf{1}_n) \quad \text{Fubini's Theorem}\\ &= \sum_{n=0}^{\infty} P(X > n)\end{aligned}$$

Example 4. In particular let $X \sim \mathsf{Geo}(p)$ where $\mathbb{E}(X) = \sum_{k=0}^{\infty} k(1-p)^{k-1}p$. More easily we have seen that $P(X > n) = (1-p)^n$. Therefore by the geometric series,

$$\mathbb{E}(X) = \sum_{n=0}^{\infty} P(X > n) = \sum_{n=0}^{\infty} (1-p)^n = \frac{1}{1-(1-p)} = \frac{1}{p}$$

1.2 Moment Generating Function

Definition 1. Let X be a r.v. Then the function,

$$M(t) = \mathbb{E}\left(e^{tX}\right) \tag{1.5}$$

is called the **moment generating function (mgf)** if X if the expectation exists for all $t \in (-h, h)$ for some $h > 0$.

Remark 2. The moment generating function M is not always defined. It is important to check the existence of the expectation.
To compensate this, the latter condition in definition 1 is necessary because the expectation $\mathbb{E}\left(e^{tX}\right)$ might not always exist for some t. Also notice that $M(0) = 1$ always.

We will now discuss some important properties of the moment generating function.

Theorem 1. *The moment generating function generates moments. For $t = 0$,*

$$M(0) = 1$$

Also,

$$M^{(k)}(0) \equiv \frac{\mathrm{d}^k}{\mathrm{d}t^k} M(t) \mid_{t=0}$$

Has the nice property,

$$M^{(k)}(0) = \mathbb{E}\left(X^k\right)$$

Proof. Evidently,

$$M(0) = \mathbb{E}\left(e^{0\cdot X}\right) = \mathbb{E}(1) = 1$$

Moreover,

$$\begin{aligned}
M^{(k)}(0) &= \frac{\mathrm{d}^k}{\mathrm{d}t^k} M(t) \mid_{t=0} \\
&= \frac{\mathrm{d}^k}{\mathrm{d}t^k} \mathbb{E}\left(e^{tX}\right) \mid_{t=0} \\
&= \mathbb{E}\left(\frac{\mathrm{d}^k}{\mathrm{d}t^k} e^{tX} \mid_{t=0}\right) \quad \text{Dominant convergence theorem.} \\
&= \mathbb{E}\left(X \frac{\mathrm{d}^{k-1}}{\mathrm{d}t^{k-1}} e^{tX} \mid_{t=0}\right) \\
&= \cdots \\
&= \mathbb{E}\left(X^k e^{tX} \mid_{t=0}\right) \\
&= \mathbb{E}\left(X^k\right)
\end{aligned}$$

□

As a result Taylor series gives,

$$\begin{aligned}
M(t) &= \sum_{k=0}^{\infty} \frac{M^{(k)}(0)}{k!} t^k \\
&= \sum_{k=0}^{\infty} \frac{\mathbb{E}\left(X^k\right)}{k!} t^k
\end{aligned}$$

Which is a method that can be used to obtain the moment of a mgf.

Theorem 2. *Let $X \perp\!\!\!\perp Y$ with mgfs M_x and M_y be respective mgfs. Let M_{X+Y} be the mgf of $X + Y$. Then,*

$$M_{X+Y} = M_X M_Y$$

Proof.

$$\begin{aligned}
M_{X+Y}(t) &= \mathbb{E}\left(e^{t(X+Y)}\right) \\
&= \mathbb{E}\left(e^{tX} e^{tY}\right) \\
&= \mathbb{E}\left(e^{tX}\right)\mathbb{E}\left(e^{tY}\right) \quad \text{Independence} \\
&= M_X(t) M_Y(t)
\end{aligned}$$

□

Theorem 3. *The moment generating function completely determines the distribution of a r.v.*

$$M_X(t) = M_Y(t) \quad \forall t \in (-h, h)$$

For some $h > 0$, then

$$X \overset{d}{=} Y$$

Which denotes that the random variables have the same distribution.

How can the moment generating function help?

Example 5. Let $X \sim \mathsf{Poi}(\lambda_1)$ and $Y \sim \mathsf{Poi}(\lambda_2)$ where $X \perp\!\!\!\perp Y$. Find the distribution of $X + Y$.
To answer this, first derive the moment generating function of a Poisson distribution.

$$\begin{aligned}
M_X(t) &= \mathbb{E}\big(e^{tX}\big) \\
&= \sum_{n=0}^{\infty} e^{tn} P(X = n) \\
&= \sum_{n=0}^{\infty} e^{tn} \frac{\lambda_1^n}{n!} e^{-\lambda_1} \\
&= e^{-\lambda_1} \sum_{n=0}^{\infty} \frac{(e^t \lambda_1)^n}{n!} \\
&= e^{-\lambda_1} e^{\left(e^t \lambda_1\right)} \quad \text{Taylor series} \\
&= e^{\lambda_1\left(e^t - 1\right)}
\end{aligned}$$

Likewise, $M_Y(t) = e^{\lambda_2\left(e^t-1\right)}$. Therefore since $X \perp\!\!\!\perp Y$,

$$M_{X+Y}(t) = M_X(t) M_Y(t) = e^{(\lambda_1+\lambda_2)\left(e^t-1\right)}$$

Therefore by theorem 3, the distribution of $X + Y$ is the same distribution as $\mathsf{Poi}(\lambda_1 + \lambda_2)$.

In general, if $X_1, X_2, \ldots, X_n$ are independent and $X_i \sim \mathsf{Poi}(\lambda_i)$, then,

$$\sum_{i=1}^{n} X_i \sim \mathsf{Poi}\left(\sum_{i=1}^{n} \lambda_i\right)$$

Definition 2. Moreover, we define the **joint moment generating function (jmgf)** for X, Y random variables to be,

$$M(t_1, t_2) = \mathbb{E}\big(e^{t_1 X + t_2 Y}\big)$$

Provided that the expectation exists for $t_1 \in (-h_1, h_1)$ and $t_2 \in (-h_2, h_2)$ for $h_1, h_2 > 0$.

Evidently, the joint moment generating function can be defined for any number of random variables. More generally, we can define the joint moment generating function with parameters $t_1, \ldots, t_n$ to be,

$$M(t_1, \ldots, t_n) = \mathbb{E}\left(\exp\left(\sum_{i=1}^{n} t_i X_i\right)\right)$$

For r.v.s $X_1, \ldots, X_n$ provided that the expectation exists for $t_i \in (-h_i, h_i)$ for some $h_i > 0$ for $i = 1, \ldots, n$. There are some nice properties of the jmgf. First, it should be possible to obtain the mgf from a particular r.v. X_i from the jmgf which includes X_i.

Notice that,

$$M_X(t) = \mathbb{E}\big(e^{tX}\big) = \mathbb{E}\big(e^{t \cdot X + 0 \cdot Y}\big) = M_{XY}(t, 0)$$

Another property of the jmgf is,

$$\frac{\partial^{m+n}}{\partial t_1^m \partial t_2^n} M(t_1,t_2) \mid_{0,0} = \mathbb{E}(X^m Y^n)$$

The proof being very similar to the single r.v. case.
Thirdly, we have that if $X \perp\!\!\!\perp Y$, then,

$$M(t_1,t_2) = M_X(t_1) M_Y(t_2) \tag{1.6}$$

Proof.

$$\begin{aligned} M(t_1,t_2) &= \mathbb{E}\big(e^{t_1 X + t_2 Y}\big) \\ &= \mathbb{E}\big(e^{t_1 X} e^{t_2 Y}\big) \\ &= \mathbb{E}\big(e^{t_1 X}\big)\mathbb{E}\big(e^{t_2 Y}\big) \quad \text{Independence} \\ &= M_X(t_1) M_Y(t_2) \end{aligned}$$

□

Remark 3. It is important not to confuse this result with theorem 2. The difference being that theorem 2 is a single argument mgf while eq. (1.6) is a multi-parameter function,

$$M_{X+Y}(t) \neq M_{X,Y}(t_1,t_2)$$

Therefore knowing that $M_{X+Y}(t)$ is separable does not imply that $X \perp\!\!\!\perp Y$. eq. (1.6) is a stronger statement than theorem 2.

2 Conditional Distribution and Conditional Expectation

2.1 Conditional Distribution

We first begin with the discrete case.

Definition 3. Let X and Y ne discrete r.v.s. Then the **conditional distribution** of X given Y is given by,

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

We read this as the probability of the event $\{X = x\}$ given that $\{Y = y\}$ holds. We can also write this as a conditional pmf,

$$f_{X|Y=y}(x) \quad \text{or} \quad f_{X|Y}(x \mid y)$$

The conditional probability is a legitimate pmf. First note that as required,

$$f_{X|Y=y}(x) \geq 0 \quad \forall x$$

Also it should be clear that,

$$\sum_x f_{X|Y=y}(x) = 1$$

In fact, $P(Y = y)$ acts a *normalization constant* for the probabilities $\sum_x P(X = x, Y = y)$. Note that given $Y = y$, as X changes, the value of the function $f_{X|Y=y}(x)$ is proportional to the joint probability $P(X = x, Y = y)$.

$$f_{X|Y=y}(x) \propto P(X = x, Y = y) \tag{2.1}$$

Namely the proportionality constant is of course $(P(Y = y))^{-1}$. Although easy to understand, eq. (2.1) can be used to solve problems where the denominator $P(Y = y)$ is difficult to find.

Example 6. Let $X_1 \sim \mathsf{Poi}(\lambda_1)$ and $X_2 \sim \mathsf{Poi}(\lambda_2)$ such that $X_1 \perp\!\!\!\perp X_2$ and $Y = X_1 + X_2$. Then we can find $P(X_1 = k \mid Y = n) = f_{X|Y=y}(k)$ using the following process. Notice that $f_{X|Y=y}(k)$ can only be non-zero when $0 \le k \le n$ in order for $Y = X_1 + X_2$. In this case for fixed n,

$$P(X_1 = k \mid Y = n) = \frac{P(X_1 = k, Y = n)}{P(Y = n)} \propto P(X_1 = k, Y = n)$$

But since $Y = X_1 + X_2$ it must be that,

$$\begin{aligned} P(X_1 = k \mid Y = n) &\propto P(X_1 = k, X_2 = n - k) \\ &= e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n-k)!} \\ &\propto \frac{\lambda_1^k}{k!} \frac{\lambda_2^{-k}}{(n-k)!} \end{aligned} \tag{2.2}$$

If we want to find the exact proportionality constant for eq. (2.2), we simply need to normalize $P(X_1 = k \mid Y = n)$ by summing over all values of k in eq. (2.2),

$$P(X_1 = k \mid Y = n) = \frac{\lambda_1^k}{k!} \frac{\lambda_2^{-k}}{(n-k)!} \left\{ \frac{\lambda_1^k}{k!} \frac{\sum_{k=0}^{n} \lambda_2^{-k}}{(n-k)!} \right\}^{-1}$$

Proceeding using this technique is difficult because of the nasty summation. The easier way is to continue the proportionality analysis. Compare eq. (2.2) with the known result for common distributions. In particular, let's consider $X \sim \mathrm{Bin}(n, p)$,

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

Removing constants,

$$P(X = k) \propto \left(\frac{p}{1-p} \right)^k (k!(n-k)!)^{-1}$$

Choosing $p/(1-p) = \lambda_1/\lambda_2$, then,

$$P(X_1 = k \mid Y = n) \propto P(X = k)$$

Therefore we can conclude that $P(X_1 = k \mid Y = n)$ follows a binomial distribution with parameters n and p given by,

$$\frac{p}{1-p} = \frac{\lambda_1}{\lambda_2} \implies p = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

Therefore,

$$P(X_1 = k \mid Y = n) = \binom{n}{k} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^k \left(1 - \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right) \right)^{n-k}$$

We introduce the following notation. Denoted $X, Y \mid \{Z = k\} \overset{\text{iid}}{\sim} \cdots$, this means that X and Y are **conditionally independent** and follows a certain distribution. i.e. the conditional joint cdf/pmf/pdf equals to the product of conditional (marginal) cdfs/pmfs/pdfs.

2.2 Conditional Expectation

We have seen that conditional pmf/pdf are legitimate pmf/pdf. Correspondingly, a conditional distribution is nothing else but a probability distribution. It is simply a potentially different distribution from the original unconditional distribution, since it takes more information into account.

As a result, we can define everything which are previously defined for unconditional distributions for conditional distributions. In particular, it is natural to define the conditional expectation as follows.

Definition 4. The **conditional expectation** of $g(X)$ given $Y = y$ is defined as,

$$\mathbb{E}(g(X) \mid Y = y) = \begin{cases} \sum_{n=1}^{\infty} g(x_i) P(X = x_i \mid Y = y) & X \mid Y = y \text{ is discrete} \\ \int_{-\infty}^{+\infty} g(x) f_{X|Y}(x \mid y) \mathrm{d}x & X \mid Y = y \text{ is continuous} \end{cases}$$

The conditional expectation is nothing else but the expectation taken under the conditional distribution. There are of course different way t understand conditional expectations.

1. Fix a value y, $\mathbb{E}(g(X) \mid Y = y)$ is a number
2. As y changes $\mathbb{E}(g(X) \mid Y = y)$ is a *function* of y
3. Since Y is actually a random variable, we can define $\mathbb{E}(g(X) \mid Y) = h(Y)$ as a random variable itself.

$$\mathbb{E}(g(X) \mid Y)_{(\omega)} = \mathbb{E}\big(g(X) \mid Y = Y_{(\omega)}\big)$$

This random variable takes value $\mathbb{E}(g(X) \mid Y = y)$ when $Y = y$.

Theorem 4. *Properties of conditional expectation:*

1. *Linearity (inherited from expectation)*

$$\mathbb{E}(aX + b \mid Y = y) = a\mathbb{E}(X \mid Y = y) + b$$

$$\mathbb{E}(X + Z \mid Y = y) = \mathbb{E}(X \mid Y = y) + \mathbb{E}(Z \mid Y = y)$$

2. $\mathbb{E}(g(X, Y) \mid Y = y) = \mathbb{E}(g(X, y) \mid Y = y)$

Proof. (Discrete Case)

$$\mathbb{E}(g(X, Y) \mid Y = y) = \sum_{x_i} \sum_{y_j} g(x_i, y_j) P(X = x_i, Y = y_j \mid Y = y)$$

Where $P(X = x_i, Y = y_j \mid Y = y)$ is self-contradictory if $y_j \neq y$. Therefore,

$$P(X = x_i, Y = y_i \mid Y = y) = \begin{cases} 0 & y_j \neq y \\ P(X = x_i \mid Y = y) & y_j = y \end{cases}$$

Therefore,

$$\begin{aligned} \mathbb{E}(g(X, Y) \mid Y = y) &= \sum_{x_i} g(x_i, y_j) P(X = x_i \mid Y = y) \\ &= \mathbb{E}(g(X, y) \mid Y = y) \end{aligned}$$

Where $g(X, y)$ is simply regarded as a function of X. □

Remark 4. In particular if $g(X, Y)$ is separable, we have that,

$$\mathbb{E}(g(X)h(Y) \mid Y = y) = h(y)\mathbb{E}(g(X) \mid Y = y)$$

Which implies that,

$$\mathbb{E}(g(X)h(Y) \mid Y) = h(Y)\mathbb{E}(g(X) \mid Y)$$

Theorem 5. *If $X \perp\!\!\!\perp Y$ then,*

$$\mathbb{E}(g(X) \mid Y = y) = \mathbb{E}(g(X))$$

Proof. If $X \perp\!\!\!\perp Y$ then the conditional distribution of X given $Y = y$ is the same as the unconditional distribution of X. We can easily prove this for the pmf in the discrete case.

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

□

Theorem 6. ***Law of iterated expectation:***

$$\mathbb{E}(\mathbb{E}(X \mid Y)) = \mathbb{E}(X)$$

The law of iterated expectation is easy to digest when one understands that $\mathbb{E}(X \mid Y)$ is r.v. function of Y.

Proof. (Discrete Case) We know that when $Y = y_j$,

$$\mathbb{E}(X \mid Y = y_j) = \sum_{x_i} x_i P(X = x_i \mid Y = y_j)$$

This happens with probability that $Y = y_j$. Therefore,

$$\begin{aligned}
\mathbb{E}(\mathbb{E}(X \mid Y)) &= \sum_{y_j} (\mathbb{E}(X \mid Y = y_j)) P(Y = y_j) \\
&= \sum_{y_j} \left(\sum_{x_i} x_i P(X = x_i \mid Y = y_j) \right) P(Y = y_j) \\
&= \sum_{x_i} x_i \sum_{y_j} P(X = x_i \mid Y = y_j) P(Y = y_j) \\
&= \sum_{x_i} x_i \sum_{y_j} P(X = x_i, Y = y_j) \\
&= \sum_{x_i} x_i P(X = x_i) \\
&= \mathbb{E}(X)
\end{aligned}$$

$\square$

Example 7. Let's say that Y is the number of claims received by an insurance company and $X \sim \mathsf{Exp}(\lambda)$ is some random parameter. We know that $Y \mid X \sim \mathsf{Poi}(X)$ which denotes that,

$$\forall x : Y \mid X = x \sim \mathsf{Poi}(x)$$

What then is $\mathbb{E}(Y)$? Since $Y \mid X \sim \mathsf{Poi}(X)$ we have that,

$$\mathbb{E}(Y \mid X = x) = x \implies \mathbb{E}(Y \mid X) = X$$

Therefore by theorem 6,

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y \mid X)) = \mathbb{E}(X) = \frac{1}{\lambda}$$

What then is $P(Y = n)$?

$$\begin{aligned}
P(Y = n) &= \int_0^\infty P(Y = n \mid X = x) f_X(x) \mathrm{d}x \\
&= \int_0^\infty \frac{e^{-x}\lambda^n}{n!} \lambda e^{-\lambda x} \mathrm{d}x \\
&= \frac{\lambda}{n!} \int_0^\infty x^n e^{-(\lambda+1)x} \mathrm{d}x \\
&= \frac{\lambda}{(\lambda+1)^{n+1} n!} \int_0^\infty ((\lambda+1)x)^n e^{-(\lambda+1)x} \mathrm{d}(\lambda+1)x \\
&= \frac{\lambda}{(\lambda+1)^{n+1} n!} \Gamma(n+1)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\lambda}{(\lambda+1)^{n+1} n!} n! \\
&= \frac{\lambda}{(\lambda+1)^{n+1}} \\
&= \left(\frac{1}{1+\lambda}\right)^n \left(1 - \frac{1}{\lambda+1}\right)
\end{aligned}
$$

Therefore $Y + 1 \sim \mathsf{Geo}(\lambda/\lambda + 1)$.

Todo (TC Fraser): Finish this lecture.

Theorem 7. ***Decomposition of Variance (EVE's law):*** *Let X, Y be random variables. Then,*

$$\mathit{Var}(Y) = \mathbb{E}(\mathit{Var}(Y \mid X)) + \mathit{Var}(\mathbb{E}(Y \mid X))$$

3 Stochastic Process

To begin, the entomology of *stochastic* is that it comes an Asian word which means to *aim (at)*. As an example, in archery, aiming at a target is related to the idea of guessing where the target will be. In short, stochastic should be taken as a synonym for *random*. A *process* is something that changes over time. In conclusion, a **stochastic process** is a system that random changes over time.

As a simple system consider the system in question to be a number. There are two distinct ways to understand the definition of a stochastic process.

$$
\begin{tikzcd}
\text{Number} \arrow[r] \arrow[d] & \text{Random Variable} \arrow[d] \\
\text{Function of Time} \arrow[r] & \text{Stochastic Process}
\end{tikzcd}
$$

Therefore the two interpretations of a stochastic process are as follows:

1. A sequence of random variables
2. A random function

Attempts to formulate the second interpretation have faced many difficulties. As such, we define a stochastic process as:

4 DTMC

4.1 Review of Probability

A **random variable (r.v.)** X is a real valued function of the outcomes of a random experiment.

$$X : \Omega \to R$$

Where $\Omega = \{\omega_1, \omega_2, \ldots\}$ is the **sample space** corresponding to all possible outcomes ω_i. The outcomes can in principle be any objects (numbers, strings, etc.). We say that X maps each outcome ω to a real number $\omega \mapsto X(\omega) \in \mathbb{R}$.

A **stochastic process** is a family of random variables $\{X_t\}_{t\in T}$, defined on a common sample space Ω. T is referred to as the index set for the stochastic process which is often understood as time. The index set T can take a discrete spectrum,

$$T = \{0, 1, 2, \ldots\} \qquad \{X_n \mid n = 0, 1, 2, \ldots\}$$

Alternatively, T can take on a continuous spectrum,

$$T = \{t \mid t \geq 0\} = [0, \infty)$$

The **state space** S is the collection of all possible values of X_t's. It is important to understand the distinction of between sample space and state space. Additionally, the state space can either have discrete or continuous spectrum.

A question remains, *Why do we need the family of random variables to be defined on a common sample space?* The answer being that we would like to be able to discuss the joint behaviour of X_t's. If X_1 has domain Ω_1 and X_2 has domain Ω_2 (where $\Omega_1 \neq \Omega_2$), then one can *not* talk about common ideas of correlations and associations between X_1 and X_2. As such we assert that all members of a stochastic process share the same sample space domain Ω.

4.2 Discrete-time Markov Chain

A **discrete-time stochastic process** $\{X_n \mid n \in 0, 1, 2, \ldots\}$ is said to be a **Discrete-time Markov Chain (DTMC)** if the following conditions hold:

1. The state space is at most *countable*[1] (i.e. finite or countable).

$$S = \{0, 1, \ldots, k\} \quad \text{or} \quad S = \{0, 1, 2, \ldots\}$$

2. **Markov Property**: For any $n = 0, 1, 2, \ldots$,

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

We use capital letters X to denote the random variable and lower case letters x to denote a specific realization or valuation of X. The motivation of the Markov property is that future events $X_{n+1} = x_{n+1}$ are independent of past histories $\{X_i = x_i \mid i = 0, 1, \ldots, n-1\}$ given the immediate past state $X_n = x_n$. The intuition being that the future and the past are probabilistically independent.

Given the present, the future and the past are independent.

4.3 Transition Probability

The **transition probability** from a state $i \in S$ at time n to state $j \in S$ (at time $n+1$) is given by,

$$P_{n,i,j} \equiv P(X_{n+1} = j \mid X_n = i) \qquad n = 0, 1, 2, \ldots \tag{4.1}$$

In full generality, the transition probability could depend on time n but in this course we will restrict ourselves to transition probabilities that *do not* depend on time n ($P_{n,i,j} = P_{i,j}$). We say that the markov chain is **(time-)homogeneous** if this property holds. From now on, this will be our default setting.

The matrix of all transition probabilities $P = \{P_{i,j} \mid i, j \in S\}$ is called the **one-step transition (probability) matrix** for $\{X_n \mid n \in T\}$.

$$P = \begin{pmatrix} P_{00} & P_{01} & \cdots & P_{0j} & \cdots \\ P_{10} & P_{11} & \cdots & P_{1j} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \cdots \\ P_{i0} & P_{i1} & \cdots & P_{ij} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

[1]Countable meaning there is a one-to-one mapping from the state space to the natural numbers.

The one-step transition matrix P has the following properties:

1. The entries of P are non-negative:
$$P_{i,j} \geq 0 \tag{4.2}$$

2. The rows of P sum to unity:
$$\forall i : \sum_{j \in S} P_{ij} = 1 \tag{4.3}$$

The **n-step transition probability** is defined via the homogeneous property,

$$\forall i, j \in S : P_{ij}^{(n)} \equiv P(X_{n+m} = j \mid X_n = i) = P(X_n = j \mid X_m = i)$$

Analogously, the **n-step transition matrix** is the matrix,

$$P^{(n)} = \left\{ P_{ij}^{(n)} \mid i, j \in S \right\}$$

Theorem 8. *There is a simple relation between the n-step transition matrix $P^{(n)}$ and the one step transition matrix P.*

$$P^{(n)} = P^{(n-1)} \cdot P = \underbrace{P \cdot P \cdot \cdots \cdot P}_{n} = P^n$$

Proof. Proof by induction:

$$P^{(1)} = P \quad \text{By definition.}$$

We also have $P^{(0)} = P^0 = \mathbf{1}$ is the identity matrix. We now assume $P^{(n)} = P^n$. Then $\forall i, j \in S$,

$$\begin{aligned}
P_{ij}^{(n+1)} &= P(X_{n+1} = j \mid X_0 = i) \\
&= \sum_{k \in S} P(X_{n+1} = j, X_n = k \mid X_0 = i) \quad \text{Total probability} \\
&= \sum_{k \in S} \frac{P(X_{n+1} = j, X_n = k, X_0 = i)}{P(X_0 = i)} \\
&= \sum_{k \in S} \frac{P(X_{n+1} = j, X_n = k, X_0 = i)}{P(X_n = k, X_0 = i)} \frac{P(X_n = k, X_0 = i)}{P(X_0 = i)} \\
&= \sum_{k \in S} P(X_{n+1} = j \mid X_n = k, X_0 = i) \cdot P(X_n = k \mid X_0 = i) \quad \text{Conditional total probability} \\
&= \sum_{k \in S} P(X_{n+1} = j \mid X_n = k) \cdot P(X_n = k \mid X_0 = i) \quad \text{Use Markov Property} \\
&= \sum_{k \in S} P_{kj} \cdot P_{ik}^{(n)} \quad \text{Matrix terms} \\
&= \left(P \cdot P^{(n)} \right)_{ij} \quad \text{Matrix product} \\
&= \left(P^{n+1} \right)_{ij} \quad \text{Inductive Hypothesis}
\end{aligned}$$

There we have proved that $P^{(n+1)} = P^{n+1}$ and so we have a completed the proof that $P^{(n)} = P^n$.

□

This result is very fundamental. We now have a relationship between the n-step transition matrix and the 1-step transition matrix (namely $P^{(n)} = P^n$). It is important to not to be confused by notation ($P^{(n)} = P^n$ is not a tautology). $P^{(n)}$ is a single matrix with entries populated by n-step transition probabilities while P^n is a single matrix multiplied by itself $n - 1$ times.

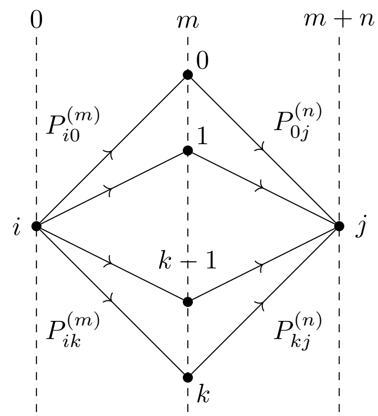
Corollary 9. *As a corollary, we have obtained that,*

$$P^{(n)} = P^{(m)} \cdot P^{(n-m)} \quad \forall 0 \leq m \leq n$$

Or equivalently the ***Chapman-Kolmogorov (C-K) Equation****,*

$$P_{ij}^{(n)} = \sum_{k \in S} P_{ik}^{(m)} P_{kj}^{(n-m)} \quad \forall i, j \in S, \forall 0 \leq m \leq n \tag{4.4}$$

It is very common to use the entry-wise form of the C-K equation (eq. (4.4)) instead of the more compact but less expressive matrix form ($P^{(n)} = P^{(m)} \cdot P^{(n-m)}$). Pictorially the C-K gives reveals the following picture that holds for all Markov chains,



So far, we have only been discussing transition probabilities. We will now divert our attention to actual distributions for a stochastic process.

Let $\alpha_n = (\alpha_{n,0}, \alpha_{n,1}, \ldots)$ be the **probability distribution vector** for X_n at time n.

$$\alpha_{n,k} = P(X_n = k) \qquad \forall k \in S$$

Note that $\alpha_{n,k} \geq 0$ and $\sum_{k \in S} \alpha_{n,k} = 1$ and $n = 0, 1, 2, \ldots$. We also define the initial distribution α_0,

$$\alpha_0 = (P(X_0 = 0), P(X_0 = 1), \ldots)$$

Theorem 10. *The transition probability matrix reveals the following relationship between the distribution α_n at time n and the distribution α_0 at time 0,*

$$\alpha_n = \alpha_0 \cdot P^n \tag{4.5}$$

Proof. The proof eq. (4.5) is quite trivial:

$$\begin{aligned}
\forall j \in S \quad \alpha_{n,j} &= P(X_n = j) \\
&= \sum_{i \in S} P(X_n = j \mid X_0 = i) \cdot P(X_0 = i) \\
&= \sum_{i \in S} \alpha_{0,i} \cdot P_{ij}^n \\
&= \alpha_{0,0} \cdot P_{0j}^n + \alpha_{0,1} \cdot P_{1j}^n + \ldots \\
&= (\alpha_0 \cdot P^n)_j
\end{aligned}$$

□

More generally, for any $n = 1, 2, \ldots$ the finite dimensional distribution can be obtained from the following process iterative process,

$$\begin{aligned} &P(X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = \\ &\qquad P(X_0 = x_0)\cdot \\ &\qquad P(X_1 = x_1 \mid X_0 = x_0)\cdot \\ &\qquad P(X_2 = x_2 \mid X_1 = x_1, X_0 = x_0)\cdots \\ &\qquad P(X_n = x_n \mid X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) \end{aligned}$$

But by the Markov condition, it must be that,

$$\begin{aligned} &P(X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = \\ &\qquad P(X_0 = x_0)\cdot \\ &\qquad P(X_1 = x_1 \mid X_0 = x_0)\cdot \\ &\qquad P(X_2 = x_2 \mid X_1 = x_1)\cdots \\ &\qquad P(X_n = x_n \mid X_{n-1} = x_{n-1}) \end{aligned}$$

First recognize the first term on the RHS ($P(X_0 = x_0) = \alpha_{0,x_0}$), and also the remaining terms are transition probabilities as per eq. (4.1). Therefore it must be that,

$$P(X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = \alpha_{0,x_0} P_{x_0 x_1} P_{x_1 x_2} \cdots P_{x_{n-1} x_n}$$

Even more generally, for $0 \le t_1 < t_2 < \cdots < t_n$,

$$P\big(X_{t_n} = x_{t_n}, X_{t_{n-1}} = x_{t_{n-1}}, \ldots, X_{t_1} = x_{t_1}\big) = P(X_{t_1} = x_{t_1})\big(P^{t_2 - t_1}\big)_{x_{t_1} x_{t_2}} \big(P^{t_3 - t_2}\big)_{x_{t_2} x_{t_3}} \cdots \big(P^{t_n - t_{n-1}}\big)_{x_{t_{n-1}} x_{t_n}}$$

Since $P(X_{t_1} = x_{t_1}) = \alpha_{t_1 x_{t_1}} = \sum_{k \in S} \alpha_{0,k} P^{t_1}_{k, x_{t_1}}$,

$$\alpha_{t_1} = \alpha_0 \cdot P^{t_1} \tag{4.6}$$

Remark 5. Equation (4.5) carries a very important interpretation. The probabilistic properties of a Discrete-Time Markov Chain (DTMC) are fully characterized by two things:

1. The initial distribution α_0
2. Transition matrix P

Knowing both the initial distribution and transition matrix fully determines the distribution α_n for all times n.

4.4 Stationary Distribution (Invariant Distribution)

In this section, we are interested in determining which distributions α_0 remain unchanged for all time $n \in T$.

Definition 5. A probability distribution $\pi = (\pi_0, \pi_1, \cdots)$ is called a **stationary (invariant) distribution** of the DTMC $\{X_n\}_{n=0,1,\cdots}$ with transition matrix P if the following conditions hold,

1. The transition matrix does not change π:
$$\pi = \pi \cdot P \tag{4.7}$$
2. The vector π is a valid probability distribution,
$$\sum_{i \in S} \pi_i = 1 \qquad \pi_i \ge 0 \tag{4.8}$$

Notice that if we posit that π is a probability distribution, then the second condition is already satisfied. Nonetheless, in practice we are able to find candidate π's using the the first condition and then we need to check these candidates against the second condition. In general, if $\sum_{i\in S}\pi_i$ is bounded (i.e. $\sum_{i\in S}\pi_i < \infty$) then it is possible to normalize the candidate π in order to satisfy eq. (4.8).

Why are such π's called stationary/invariant distributions? Notice that eq. (4.7) completely answers this question. Assume that the MC starts with initial distribution $\alpha_0 = \pi$ for X_0. In this case, the distribution of X_1 is determined by P,

$$\alpha_1 = \alpha_0 \cdot P$$

But since α_0 is π and π satisfies eq. (4.7),

$$\alpha_1 = \pi \cdot P = \pi$$

The distribution for X_1 is the *same* as the distribution for X_0. This process continues,

$$\alpha_2 = \alpha_1 \cdot P = \pi \cdot P = \pi$$

$$\alpha_n = \alpha_0 \cdot P^n = \pi \cdot P^n = \pi \cdot P^{n-1} = \cdots = \pi$$

Thus if the Markov chain starts with a stationary/invariant distribution then its marginal distribution will *never change*; hence why we refer π as stationary. Also not that this *does not* indicate that the value of X_i does not change over time (it almost certainly will), but its distribution does.

Example 8. Consider an electron with two states: ground (0) and excited (1). Let X_n be the state at time n. At each step, with probability α the MC chains state if it is in the ground state. With probability β the MC will transition to the ground state if it is in the excited state. Then $\{X_n\}_{n=0,1,\ldots}$ is a DTMC and its transition matrix is,

$$P = \begin{array}{c c} & \begin{array}{cc} \scriptstyle(0) & \scriptstyle(1) \end{array} \\ \begin{array}{c} \scriptstyle(0) \\ \scriptstyle(1) \end{array} & \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix} \end{array}$$

Now let us solve for the stationary distribution π.

$$\pi = \pi \cdot P \qquad \pi = (\pi_0, \pi_1) \qquad \pi_0 + \pi_1 = 1$$

Therefore,

$$\pi_0 = (1-\alpha)\pi_0 + \beta\pi_1 \tag{4.9}$$

$$\pi_1 = \alpha\pi_0 + (1-\beta)\pi_1 \tag{4.10}$$

However note that these two equations are not linearly independent. This is evident because summing eq. (4.9) with eq. (4.10) results in the trivial statement of $\pi_0 + \pi_1 = \pi_0 + \pi_1$. Nonetheless rearranging eq. (4.9) gives,

$$\alpha\pi_0 = \beta\pi_1 \implies \frac{\pi_0}{\pi_1} = \frac{\beta}{\alpha}$$

This is where we need $\pi_0 + \pi_1 = 1$.

$$\pi_0 = \frac{\beta}{\alpha+\beta} \qquad \pi_1 = \frac{\alpha}{\alpha+\beta}$$

Where $\alpha + \beta$ is considered the normalizing constant.

An important remark: sometimes the candidate distribution is not normalizable. In particular, there are configurations where eq. (4.7) is satisfiable but eq. (4.8) is not. In the above example, there exists a unique stationary distribution,

$$\pi = \left(\frac{\alpha}{\alpha+\beta}, \frac{\beta}{\alpha+\beta}\right)$$

If $\alpha_0 = \pi$ then we know immediately that,

$$P(X_n = 0) = \frac{\beta}{\alpha+\beta} \qquad P(X_n = 1) = \frac{\alpha}{\alpha+\beta} \qquad \forall n = 1, 2, \ldots$$

Remark 6. By the above procedure of solving for stationary distribution is typical.

1. Use eq. (4.7) to get proportions between different components of π.
2. Use eq. (4.8) to normalize π and get exact values.

Remark 7. Note that if $\beta = 2\alpha$ then π is always $(2/3, 1/3)$ regardless the actual value of α.

4.5 Classification of States

Let $\{X_n\}_{n=0,1,\dots}$ be a DTMC with state space S. State j is **accessible** from state i (denoted $i \to j$) if there exists $n = 0, 1, \dots$ such that $P_{ij}^{(n)} > 0$. Intuitively, one can transition from state i to state j in finite steps n with positive probability. If i is also accessible from j, then we say i and j **communicate**, denoted as $i \leftrightarrow j$.

$$i \leftrightarrow j \iff \exists m, n \geq 0, P_{ij}^{(m)} > 0, P_{ji}^{(n)} > 0$$

The relation "$\leftrightarrow$" either holds or does not hold for every state $i \in S$.

Theorem 11. *The binary communication relation "$\leftrightarrow$" is in fact a equivalence relation:*

- *Reflexivity* $i \leftrightarrow i$
- *Symmetry* $i \leftrightarrow j \implies j \leftrightarrow i$
- *Transitivity* $i \leftrightarrow j, j \leftrightarrow k \implies i \leftrightarrow k$

Proof. First, reflexivity is easy to prove by definition. Let $n = 0$ and recognize that $P_{ii}^{(0)}$ has a certain probability by definition,

$$P_{ii}^{(0)} = 1 \implies i \leftrightarrow i$$

Second, symmetry follows by definition,

$$P_{ij}^{(m)} > 0, P_{ji}^{(n)} > 0 \Leftrightarrow P_{ji}^{(n)} > 0, P_{ij}^{(m)} > 0$$

Third, transitivity can be proven by letting m and n be the unknown quantifiers:

$$\exists m \quad P_{ij}^{(m)} > 0, \exists n \quad P_{jk}^{(n)} > 0$$

Then by the CK equation eq. (4.4),

$$P_{ik}^{(m+n)} = \sum_{l \in S} P_{il}^{(m)} P_{lk}^{(n)}$$

Let $l = j$ be a single, fixed entry in the summation,

$$P_{ik}^{(m+n)} \geq P_{ij}^{(m)} P_{jk}^{(n)} > 0$$

Therefore we have that k is accessible from i $(i \to j)$. Analogously we have that $i \to j$ therefore $i \leftrightarrow k$. ∎

The communication equivalence relations then divides the state space S into different **equivalence classes**. That is, the states in one class comm with each other; the states in different classes do not comm. The equivalent classes form a *partition* of the state space S.

The family $\{S_1, S_2, \dots S_n\}$ is a **partition** of S if,

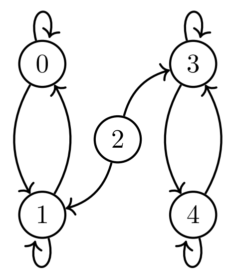
1. $S_i \subset S \mid \forall i \in 1, 2, \dots, n$
2. $S_i \cap S_j \neq \emptyset$ for all $i \neq j$
3. $\bigcup_i S_i = S$

We can find the equivalent classes by drawing a graph where the states in S are the nodes of the graph and a directed edge is placed going from i to j if j is accessible from i in one-step: $P_{ij} > 0$. Then identifying the the equivalent classes corresponds to identifying the loops of this graph within one step.

Example 9. As an example, consider the transition matrix P as follows.

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0.2 & 0.8 & 0 & 0 & 0 \\ 0.6 & 0.4 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.7 & 0.3 \\ 0 & 0 & 0 & 0.1 & 0.9 \end{pmatrix} \end{array}$$

The associated one-step accessibility graph is then,



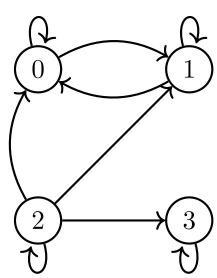
Where the loops of $S = \{0, 1, 2, 3, 4\}$ form the following partition,

$$S_1 = \{0, 1\} \quad S_2 = \{2\} \quad S_3 = \{3, 4\}$$

Example 10. As another example, consider the transition matrix P.

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

The associated one-step accessibility graph is then,



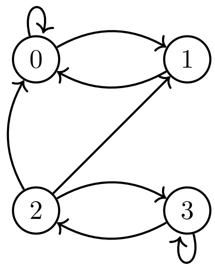
While the loops of $S = \{0, 1, 2, 3\}$ form the following partition,

$$S_1 = \{0, 1\} \quad S_2 = \{2\} \quad S_3 = \{3\}$$

Example 11. As another example, consider the transition matrix P.

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

The associated one-step accessibility graph is then,



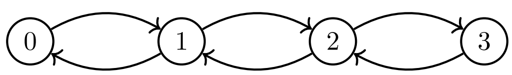
This contrasts with the previous example because $P_{01}, P_{12}, P_{20} > 0$ implies that $0, 1, 2$ are in the same class and $P_{23}, P_{32} > 0$ implies that $2, 3$ are in the same class. By transitivity $0, 1, 2, 3$ are all in the same class.

$$S_1 = \{0, 1, 2, 3\}$$

These equivalent classes are useful for Markov chains because it allows one to separate the behaviour of the equivalence classes and study them individually. A MC which has only one equivalent class is called **irreducible**.

Example 12.

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



Clearly if we start at state 0, we can only go back to 0 in $2, 4, 6, \ldots$ (i.e. an even number of) steps.

Furthermore, let us define the **period** of state i as,

$$d(i) = \gcd\left\{n \in \mathbb{Z}^+ \mid P_{ii}^n > 0\right\}$$

Additionally, if $P_{ii}^n = 0$ holds for all $n > 0$, we say that $d(i) = \infty$. If the period of i happens to be $d(i) = 1$ then the state i is said to be **aperiodic**. Alternatively, locus of steps that we can go back by are *co-prime*. A MC is called aperiodic if all its states S are aperiodic.

Note that $P_{ii} < 0$ then the greatest common divisor must be one. This implies that $d_i = 1$. In this case, the state is immediately aperiodic.

The period of a state is useful do to the following theorem,

Theorem 12. *The period of a state is a class property. If $i \leftrightarrow j$, then $d(i) = d(j)$.*

Proof. If $i = j$ we are already done. If $i \neq j$, since $i \leftrightarrow j$, then $\exists n, m$ such that,

$$P_{ij}^n > 0 \quad P_{ji}^m > 0$$

Then for any l such that $P_{jj}^l > 0$,

$$P_{ii}^{n+m+l} \geq P_{ij}^n P_{jj}^l P_{ji}^m \tag{4.11}$$

Because $P_{ij}^n P_{jj}^l P_{ji}^m$ happens to be a specific way for P_{ii}^{n+m+l} to occur. Since $i \leftrightarrow j$ and l was chosen carefully,

$$P_{ii}^{n+m+l} > 0$$

Moreover, we also have that,

$$P_{ii}^{n+m} \geq P_{ij}^{n} P_{ji}^{m} \tag{4.12}$$

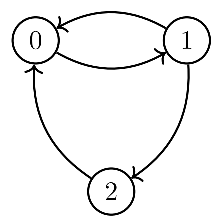
Since $d(i)$ divides both $n+m$ and $n+m+l$ by eqs. (4.12) and (4.11), then $d(i)$ also divides l. This holds for all l such that $P_{ii}^{l} > 0$. This implies that $d(i)$ is a common divisor of $\{l \mid P_{jj}^{l} > 0\}$ an thus $d(i)$ divides,

$$d(j) = \gcd\{l \mid P_{jj}^{l} > 0\}$$

By symmetry $d(j)$ divides $d(i)$. Therefore $d(i) = d(j)$.

□

Remark 8. It is important to note that $d(i) = k \not\Rightarrow P_{ii}^{(k)} > 0$. As a counterexample consider the following one step accessibility graph,



Evidently,

$$P_{00}^{(2)} > 0 \quad P_{00}^{(3)} > 0$$

So we have $d(0) = 1$ because $d(0) = \gcd\{2, 3, \dots\}$. However this doesn't imply that $P_{00}^{(1)} > 0$ because we do have that $P_{00}^{(1)} = 0$

Remark 9. If the MC is irreducible (having only one class) then all the states have the same period. In this case we ascribe the entire MC the period $d(i)$ for some representative $i \in S$.

Remark 10. If the period of i is $d_i = 1$ then there exists some N such that $P_{ii}^{(n)} > 0$ for any $n \geq N$. Intuitively, if state i is aperiodic then after a long time, the probability of going back to i is always positive.

4.6 Recurrence and Transience

In order to define transience and recurrence, let T_i be the waiting time of a MC to visit/revisit state i for the first time since time 0.

$$T_i = \min\{n > 0 : X_n = i\}$$

Of course we have that $\{n > 0 : X_n = i\}$ is a random collection of numbers and also $\min\{n > 0 : X_n = i\}$ is a random number. Therefore T_i is a random variable. Notice that if $T_i = \infty$ if the MC never returns to state i.

Definition 6. A state i is called **transient** if,

$$P(T_i < \infty \mid X_0 = i) < 1$$

Or equivalently,

$$P(T_i = \infty \mid X_0 = i) > 0$$

We say that the MC never goes back to state i with positive probability.

Moreover,

Definition 7. A state i is **recurrent** if,

$$P(T_i < \infty \mid X_0 = i) = 1$$

Or equivalently,

$$P(T_i = \infty \mid X_0 = i) = 0$$

We say that the MC always goes back to i.

Remark 11. If we have that $P(T < \infty) = 1$ then it does not imply that $\mathbb{E}(T) < \infty$. As a counter example, let $s_p = 2^p$ for $p = 1, \ldots, \infty$,

$$P(T = s_p) = \frac{1}{s_p}$$

Where the expectation of T is unbounded,

$$\mathbb{E}(T) = \sum_{p=1}^{\infty} s_p \frac{1}{s_p} = \sum_{p=1}^{\infty} 1 = \infty$$

To make the distinction,

Definition 8. A recurrent state i is said to be **positive recurrent** if,

$$\mathbb{E}(T_i \mid X_0 = i) < \infty$$

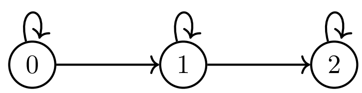
A recurrent state i is said to be **null recurrent** if,

$$\mathbb{E}(T_i \mid X_0 = i) = \infty$$

Example 13. Consider the following example with transition matrix,

$$P \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \!\!\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix} \end{array}$$

And MC graphically expressed as,



Given that $X_0 = 0$ there are only two possibilities for transitions,

$$P\left(\underbrace{X_1 = 0 \mid X_0 = 0}_{T_0 = 1}\right) = P\left(\underbrace{X_1 = 1 \mid X_0 = 0}_{T_0 = \infty}\right) = \frac{1}{2}$$

The second term is recognized as $T_0 = \infty$ because after transitioning to state 1, it is impossible to return to state 0. This tells us that,

$$P(T_0 < \infty \mid X_0 = 0) = \frac{1}{2} < 1$$

Therefore state 0 is said to be transient. Analogously we have that state 1 is also transient. Given $X_0 = 2$, $P\left(\underbrace{X_1 = 2 \mid X_0 = 2}_{T_2 = 1}\right) = 1$ we have that,

$$P(T_2 < \infty \mid X_0 = 2) - 1$$

Which tells is that state 2 is recurrent.

In general, the distribution of T_i is very hard to derive. In particular it will be hard to know whether T_i takes value ∞ and with what probability. This suggests that we will need handier criteria for recurrence/transience.

To facilitate this define $f_{ii} = P(T_i < \infty \mid X_0 = i)$ and,

$$f_{ij} = P(T_j < \infty \mid X_0 = i)$$

We also defined V_i to be the number of times that the MC visits state i,

$$V_i = \sum_{n=1}^{\infty} \mathbb{1}_{\{X_n = i\}}$$

First consider the case that i is transient. If i is transient, then it must be that $f_{ii} < 1$. This can be seen from the definition of T_i. The PMF is,

$$P(V_i = k \mid X_0 = i) = f_{ii}^k (1 - f_{ii})$$

This can be derived by considering that if the MC is to visit state i exactly k times but not more than k times.

$$P(V_i = k \mid X_0 = i) = [P(T_i < \infty \mid X_0 = i)]^k [P(T_i = \infty \mid X_0 = i)]$$

This PMF tells us that $V_i + 1$ follows a geometric distribution with parameter $1 - f_{ii}$,

$$V_i + 1 \sim \mathsf{Geo}(1 - f_{ii})$$

In particular, $P(V_i < \infty \mid X_0 = i) = 1$. Therefore if i is transient, it is visited only finitely many times with probability 1. Afterwards, the MC will leave state i forever sooner or later.

Second consider the case that i is recurrent. If state i is recurrent, then $f_{ii} = 1$ by definition. The we have that,

$$P(V_i = k \mid X_0 = i) = 0 \quad \forall k = 0, 1, \ldots$$

Since V_i can not take on any finite values, it must be that

$$P(V_i = \infty \mid X_0 = i) = 1$$

If the MC starts from recurrent state i is will visit that state infinitely many times. Before identifying our more versatile criteria, we generalize some of these notions.

4.7 Recurrence and Transience Again

For $n \in \mathbb{Z}^+$ define,

$$f_{ij}^{(n)} = P(X_n = j, X_{n-1} \neq j, \cdots, X_1 \neq j \mid X_0 = i) \quad \forall i, j \in S$$

Intuitively, $f_{ij}^{(n)}$ is the probability that X visits state j at time n for the first time since $X_0 = i$. A looming question: What is the relation between $f_{ij}^{(n)}$ and $P_{ij}^{(n)}$? First notice that,

$$P_{ij}^{(n)} \geq f_{ij}^{(n)}$$

These reads: the probability that X visits j at time n is more larger that the probability that X visits j at time n provided it did not visit j prior. A more detailed equality is the following,

$$P_{ij}^{(n)} = \sum_{k=1}^{n} f_{ij}^{(k)} P_{jj}^{(n-k)} \tag{4.13}$$

Expanded out gives,

$$P_{ij}^{(n)} = f_{ij}^{(n)} + \sum_{k=1}^{n-1} f_{ij}^{(k)} P_{jj}^{(n-k)}$$

Proof.

$$\begin{aligned} P_{ij}^{(n)} &= P(X_n = j \mid X_0 = i) \\ &= \sum_{k=1}^{n} P(X_n = j, X \text{ first visits } j \text{ at time } k \mid X_0 = i) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{n} P(X_n = j, | \ X \text{ first visits } j \text{ at time } k, X_0 = i) \cdot P(X \text{ first visits } j \text{ at time } k \mid X_0 = i) \\
&= \sum_{k=1}^{n} P(X_n = j, | \ X_k = j, X_{k-1} \neq j, \ldots, X_1 \neq j, X_0 = i) \cdot P(X_k = j, X_{k-1} \neq j, \ldots, X_1 \neq j \mid X_0 = i) \\
&= \sum_{k=1}^{n} P(X_n = j, | \ X_k = j, X_{k-1} \neq j, \ldots, X_1 \neq j, X_0 = i) \cdot f_{ij}^{(k)} \\
&= \sum_{k=1}^{n} P(X_n = j, | \ X_k = j) \cdot f_{ij}^{(k)} \quad \text{Markov Condition} \\
&= \sum_{k=1}^{n} P_{jj}^{(n-k)} \cdot f_{ij}^{(k)}
\end{aligned}
$$

□

In fact eq. (4.13) defines a recurrence relation to compute $f_{ij}^{(n)}$ from $f_{ij}^{(k)}$ where $k < n$,

$$f_{ij}^{(n)} = P_{ij}^{(n)} - \sum_{k=1}^{n-1} f_{ij}^{(k)} P_{jj}^{(n+k)}$$

We now define f_{ij} *without* the superscript to be,

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)}$$

The probability that X will *ever* reach state $j \in S$ provided it started at i ($f_{ij} \leq 1$). Whether or not f_{ij} is certain or not defines the following two properties.

A state i is called **transient** if $f_{ii} < 1$; and **recurrent** if $f_{ii} = 1$. Intuitively, f_{ii} is the probability the MC returns to state i given it started in state i. If i is transient, then there is a non-negative probability that the MC does not return to i and if $f_{ii} = 1$ then the MC always returns to state i.

Another way to characterize recurrence and transience: Define V_i to be the total number of times the MC (re)visits i after time 0. In more mathematical terms,

$$V_i = \sum_{n=1}^{\infty} \mathbf{1}_{[X_n = i]}$$

Where $\mathbf{1}_{[X_n=i]}$ is the indicator defined by,

$$\mathbf{1}_{[X_n=i]} = \begin{cases} 1 & X_n = i \\ 0 & X_n \neq i \end{cases}$$

If $f_{ii} < 1$ we have that the probability of visiting state i k times is given by,

$$P(V_i = k \mid X_0 = i) = \underbrace{f_{ii} \cdot f_{ii} \cdots f_{ii}}_{k} \underbrace{(1 - f_{ii})}_{\text{never return}}$$

Where $(1 - f_{ii})$ is necessary because it guarantees that we never return to state i more that k times. Given $X_0 = i$, V_i follows a geometric distribution with parameter $(1 - f_{ii})$. Thus,

$$\mathbb{E}(V_i \mid X_0 = i) = \frac{f_{ii}}{1 - f_{ii}} < \infty$$

Therefore if i is transient, there a finite number revisits are expected. In contrast if $f_{ii} = 1$ we have that,

$$\mathbb{E}(V_i \mid X_0 = i) = \lim_{f_{ii} \to 1} \frac{f_{ii}}{1 - f_{ii}} \to \infty$$

Recalling our construction of the random variable V_i we can alternatively write $\mathbb{E}(V_i \mid X_0 = i)$ as,

$$\begin{aligned}
\mathbb{E}(V_i \mid X_0 = i) &= \mathbb{E}\left(\sum_{n=1}^{\infty} \mathbb{1}_{\{X_n = i\}} \mid X_0 = i\right) \\
&= \sum_{n=1}^{\infty} \mathbb{E}\big(\mathbb{1}_{\{X_n = i\}} \mid X_0 = i\big) \\
&= \sum_{n=1}^{\infty} P(X_n = i \mid X_0 = i) \\
&= \sum_{n=1}^{\infty} P_{ii}^{(n)} \\
&= \sum_{n=1}^{\infty} P_{ii}^{n}
\end{aligned}$$

Therefore we have equivalent criteria for recurrent and transience for a state i.

recurrent	transient
$P(T_i < \infty \mid X_0 = i) = 1$	$P(T_i < \infty \mid X_0 = i) < 1$
$P(V_i = \infty \mid X_0 = i) = 1$	$P(V_i < \infty \mid X_0 = i) = 1$
$\mathbb{E}(V_i \mid X_0 = i) = \infty$	$\mathbb{E}(V_i \mid X_0 = i) < \infty$
$\sum_{n=1}^{\infty} P_{ii}^{n} = \infty$	$\sum_{n=1}^{\infty} P_{ii}^{n} < \infty$

The final criteria being the easiest to use.

A final criteria we can consider is,

$$\mathbb{E}(V_i \mid X_0 = i) = \sum_{k=1}^{\infty} P(V_i \geq k \mid X_0 = i) \tag{4.14}$$

The proof of eq. (4.14) is left as an exercise to the reader. Clearly if $f_{ii} = 1$,

$$P(V_i \geq k \mid X_0 = i) = {f_{ii}}^{k} = 1 \quad \forall k \tag{4.15}$$

Therefore,

$$\mathbb{E}(V_i \mid X_0 = i) = \sum_{k=1}^{\infty} 1 = \infty$$

Theorem 13. *Therefore i is recurrent if and only if $P(V_i \geq k \mid X_0 = i) = \infty$ and i is transient if and only if only if $P(V_i \geq k \mid X_0 = i) < \infty$.*

Remark 12. We actually also have that i is recurrent if and only if $V_i = \infty$. This can be seen from eq. (4.15). Since $P(V_i \geq k \mid X_0 = i)$ is strictly positive for all k, then $V_i = \infty$. Analogously, we have that i is transient if and only if $V_i < \infty$.

Yet *another* way to characterize recurrence and transience is much more tractable. First,

Theorem 14. *The expectation of the indicator is given by $\mathbb{E}(\mathbf{1}_A) = P(A)$ for any event A.*

Therefore,

$$\mathbb{E}(V_i \mid X_0 = i) = \mathbb{E}\left(\sum_{n=1}^{\infty} \mathbf{1}_{[X_n = i]} \mid X_0 = i\right)$$

$$
\begin{aligned}
&= \sum_{n=1}^{\infty} \mathbb{E}\big(\mathbf{1}_{[X_n=i]} \mid X_0 = i\big) \quad \text{Fubini's Theorem} \\
&= \sum_{n=1}^{\infty} P(X_n = i \mid X_0 = i) \\
&= \sum_{n=1}^{\infty} P_{ii}^{(n)}
\end{aligned}
$$

Thus i is recurrent if and only if $\sum_{n=1}^{\infty} P_{ii}^{(n)} = \infty$ and i is transient if and only if $\sum_{n=1}^{\infty} P_{ii}^{(n)} < \infty$.

Theorem 15. *Recurrence/transience are class properties. If $i \leftrightarrow j$ and i is recurrent (transient), then j is recurrent (transient).*

Proof. Since $i \leftrightarrow j$, there must be times $\exists m, n \geq 0$ that permit i to access j and j to access i,

$$P_{ij}^{(m)} > 0 \qquad P_{ji}^{(n)} > 0$$

Suppose that i is recurrent, then $\sum_{l=1}^{\infty} P_{ii}^{(l)} = \infty$. We now what to show that $\sum_{s=1}^{\infty} P_{jj}^{(s)}$ is infinite as well. Since $m, n \geq 0$ we can drop positive terms to get,

$$\sum_{s=1}^{\infty} P_{jj}^{(s)} \geq \sum_{s=n+m+1}^{\infty} P_{jj}^{(s)}$$

Now exchange the variables of summation by letting $l = s - n - m$,

$$\sum_{s=1}^{\infty} P_{jj}^{(s)} \geq \sum_{l=1}^{\infty} P_{jj}^{(n+l+m)}$$

Then by the C-K equation (eq. (4.4)) a example path from j to j is to transition $j \to i \to i \to j$ in steps m, l, n respectively,

$$\sum_{l=1}^{\infty} P_{jj}^{(n+l+m)} \geq \sum_{l=1}^{\infty} P_{ji}^{(n)} P_{ii}^{(l)} P_{ij}^{(m)} = P_{ji}^{(n)} P_{ij}^{(m)} \left\{ \sum_{l=1}^{\infty} P_{ii}^{(l)} \right\}$$

But since i is recurrent, $\sum_{l=1}^{\infty} P_{ii}^{(l)} = \infty$. Also, $P_{ji}^{(n)} P_{ij}^{(m)} > 0$ by the choice of m, n. Therefore $\sum_{l=1}^{\infty} P_{jj}^{(n+l+m)} = \infty$ and thus $\sum_{s=1}^{\infty} P_{jj}^{(s)} = \infty$. Therefore j is also recurrent. ☐

Corollary 16. *If $i \leftrightarrow j$ and i is transient, then j is transient.*

As a result, if we know that if a MC is irreducible (admitting only one class), then either all states are transient or they are all recurrent. In such cases we can say that the MC itself is recurrent/transient.

Theorem 17. *If an irreducible MC has a finite state space, then it is recurrent.*

To see why this is true, if all states are transient then each state $i \in S$ has a time k that is the *last* visit time for all states, this is impossible because $P_{ij} \neq 0$ for at least some choice $i, j \in S$.

Remark 13. We can actually show that the MC must be positive recurrent if the state space is finite and the MC is irreducible.

Theorem 18. *If i is recurrent, and i does not communicate with j, then $P_{ij} = 0$.*

Proof. Proof by contradiction. Assume that $P_{ij} > 0$. Since i and j do not communicate, then either j is not accessible from i or vice versa. But if $P_{ij} > 0$ then j is accessible from i. It must be that i is not accessible from j. Recall that f_{ii} is the probability that the MC ever revisits the state i given the starting state was i. Therefore $1 - f_{ii}$ is the probability that the MC never revisits state i.

$$f_{ii} \leq 1 - P_{ij} < 1$$

This inequality holds because if $X_1 = j$ then the MC never revisits i (i is not accessible from j). But there are other ways it never revisits i. Therefore,

$$P(X_1 = j \mid X_0 = i) = P_{ij} \leq P(\text{MC never revisits } i \mid X_0 = i)$$

But if $f_{ii} < 1$, then i is not recurrent; it is transient. Therefore the assumption that $P_{ij} > 0$ is wrong; $P_{ij} = 0$. □

4.8 Limiting Distribution

In this section we are interested in the **limiting transition probability** and the **limiting distribution** of a MC.

$$\lim_{n\to\infty} P_{ij}^{(n)} \qquad \lim_{n\to\infty} P(X_n = j)$$

To keep things simple, we assume that the process $\{X_n\}_{n=0,1,2,\dots}$ is irreducible.

Theorem 19. ***Basic Limit Theorem:*** *Let $\{X_n\}_{n=0,1,\dots}$ be an irreducible, aperiodic and positive recurrent DTMC, then a stationary distribution $\pi = (\pi_0, \pi_1, \dots)$ exists. Moreover we have,*

$$\lim_{n\to\infty} P_{ij}^{(n)} = \lim_{n\to\infty} \frac{\sum_{k=1}^{n} \mathbb{1}_{\{X_k = j\}}}{n} = \frac{1}{\mathbb{E}(T_j \mid X_0 = j)} = \pi_j \qquad \forall i, j \in S$$

Where $T_j = \min\{n > 0, X_n = j\}$. In words, the limiting distribution is equal to the long-run fraction of time spent in j is equal to the stationary distribution. Note that this does not depend on the starting state i.

Todo (TC Fraser): Missed a lecture By examining foil theorems, we have seen that irreducibility is related to the uniqueness of the stationary distribution, aperiodicity is related to the existence of $\lim_{n\to\infty} P_{ij}^{(n)}$ and positive recurrence is related to the existence of stationary distribution.

Example 14. Let us have an electron with two states and P given by,

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

Where $\lim_{n\to\infty} P^n$ becomes,

$$\lim_{n\to\infty} P^n = \begin{pmatrix} \frac{\alpha}{\alpha+\beta} & \frac{\beta}{\alpha+\beta} \\ \frac{\alpha}{\alpha+\beta} & \frac{\beta}{\alpha+\beta} \end{pmatrix}$$

It is important to notice that $\lim_{n\to\infty} P^n$ does not depend on the row considered (i.e. the initial state i considered). We have the stationary distribution is,

$$\pi = \begin{pmatrix} \frac{\alpha}{\alpha+\beta} & \frac{\beta}{\alpha+\beta} \end{pmatrix}$$

So we verify that $\pi_j = \lim_{n\to\infty} P_{ij}^{(n)}$. Also given that we start from state $X_0 = 0$, then clearly,

$$\begin{aligned}
P(T_0 = 1 \mid X_0 = 0) &= 1 - \alpha \\
P(T_0 = 2 \mid X_0 = 0) &= \alpha\beta \\
P(T_0 = 3 \mid X_0 = 0) &= \alpha(1-\beta)\beta \\
P(T_0 = k \mid X_0 = 0) &= \alpha(1-\beta)^{k-2}\beta \quad \forall k = 2, 3, \dots
\end{aligned}$$

Where α represents leaving state 0 and β represents returning to state 0 at step k. $(1-\beta)^{k-2}$ represents not returning to 0 until step k. We then obtain the expectation,

$$\mathbb{E}(T_0 \mid X_0 = 0) = 1 - \alpha + \sum_{k=2}^{\infty} \alpha(1-\beta)^{k-2}\beta k$$

We then split this sum into two components,

$$\mathbb{E}(T_0 \mid X_0 = 0) = 1 - \alpha + \sum_{k=2}^{\infty} \alpha(1-\beta)^{k-2}\beta(k-1) + \sum_{k=2}^{\infty} \alpha(1-\beta)^{k-2}\beta$$

We then notice that $(1-\beta)^{k-2}\beta$ is the pmf of $\mathsf{Geo}(\beta)$ while the first sum $(\sum_{k=2}^{\infty}(1-\beta)^{k-2}\beta(k-1))$ is simply the expectation of $\mathsf{Geo}(\beta)$. Therefore,

$$\mathbb{E}(T_0 \mid X_0 = 0) = 1 - \alpha + \alpha\frac{1}{\beta} + \alpha = 1 + \frac{\alpha}{\beta} = \frac{\alpha+\beta}{\beta}$$

We verify that $\mathbb{E}(T_0 \mid X_0 = 0) = \frac{1}{\pi_0}$. Of course, this is one of the few cases where we can explicitly verify the predictions of the Basic Limit Theorem. In the future, we will accept the BLT and move forward.

4.9 Generating Function

Before defining the generating function we recall the moment generating function $M_k(t) = \mathbb{E}(e^{tX})$. The moment generating function is defined for any distribution X. However the generating function is defined for distributions over the non-negative integers.

Definition 9. Let $p = (p_0, p_1, \ldots)$ be a distribution over the non-negative integers $\{0, 1, 2, \ldots\}$ such that $p_i \geq 0$ and $\sum_i p_i = 1$. Let ξ be a random variable following distribution $p : \xi \sim p$. Then the generating function of ξ (or for p) is defined as,

$$\varphi(s) = \mathbb{E}(s^{\xi}) = \sum_{k=0}^{\infty} s^k p_k \quad \forall 0 \leq s \leq 1$$

The generating function is useful for the following properties.

1. The generating function always intersects two points: $\varphi(0) = p_0$, $\varphi(1) = 1$
2. The generating function determines the distribution:

$$p_k = \frac{1}{k!}\frac{\mathrm{d}^k\varphi(s)}{\mathrm{d}s^k}\bigg|_{s=0}$$

Proof. In order to proof the second property, simply take $\varphi(s)$ as a sum,

$$\varphi(s) = p_0 + p_1 s^1 + \cdots + p_{k-1}s^{k-1} + p_k s^k + p_{k+1}s^{k+1} + \cdots$$

Then the k-th derivative of $\varphi(s)$ is,

$$\frac{\mathrm{d}^k\varphi(s)}{\mathrm{d}s^k} = k! p_k s^0 + \frac{(k+1)!}{1!}p_{k+1}s^1 + \frac{(k+2)!}{2!}p_{k+2}s^2 + \cdots \tag{4.16}$$

But when evaluated at $s = 0$,

$$\frac{\mathrm{d}^k\varphi(s)}{\mathrm{d}s^k}\bigg|_{s=0} = k! p_k$$

□

Remark 14. In particular, if we restrict ourselves to $s \geq 0$ eq. (4.16) determines,

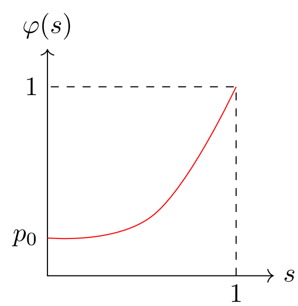
$$\varphi'(s) \geq 0 \implies \varphi(s) \text{ is increasing}$$
$$\varphi''(s) \geq 0 \implies \varphi(s) \text{ is convex}$$

3. Let $\xi_1, \xi_2, \ldots, \xi_n$ be independent random variables with generating functions $\varphi_1, \ldots, \varphi_2$. Then if $X = \xi_1 + \cdots + \xi_n$ then,
$$\varphi_X(s) = \varphi_1(s) \cdots \varphi_n(s)$$

   Proof. The proof of which is similar to the analogous property for moment generating functions.
$$\begin{aligned}\varphi_X(s) &= \mathbb{E}\left(s^X\right)\\ &= \mathbb{E}\left(s^{\xi_1+\cdots+\xi_n}\right)\\ &= \mathbb{E}\left(s^{\xi_1}\cdots s^{\xi_n}\right)\\ &= \mathbb{E}\left(s^{\xi_1}\right)\cdots\mathbb{E}\left(s^{\xi_n}\right) \quad \text{Independence}\\ &= \varphi_1(s)\cdots\varphi_n(s)\end{aligned}$$
   □

4. If we take the derivative evaluated at the point $s = 1$ then,
$$\left.\frac{\mathrm{d}^k\varphi(s)}{\mathrm{d}s^k}\right|_{s=1} = \left.\frac{\mathrm{d}^k\mathbb{E}\left(s^\xi\right)}{\mathrm{d}s^k}\right|_{s=1} = \left.\mathbb{E}\left(\frac{\mathrm{d}^k s^\xi}{\mathrm{d}s^k}\right)\right|_{s=1} = \left.\mathbb{E}\left(\xi(\xi-1)\cdots(\xi-k+1)s^{\xi-k}\right)\right|_{s=1} = \mathbb{E}\left(\frac{\xi!}{k!}\right)$$
   Clearly this is not the moment generating function but it is intimately related. In particular $\mathbb{E}(\xi) = \varphi(1)$ and,
$$\begin{aligned}\mathsf{Var}(\xi) &= \mathbb{E}\left(\xi^2\right) - \left(\mathbb{E}(\xi)\right)^2\\ &= \mathbb{E}(\xi(\xi-1)) + \mathbb{E}(\xi) - \left(\mathbb{E}(\xi)\right)^2\\ &= \varphi'(1) + \varphi'(1) - \left(\varphi'(1)\right)^2\end{aligned}$$

A characteristic plot is depicted below.



4.10 Branching Processes

Consider a population of organism such that each organism, at the end of its lifetime, produces a random number of ξ offsprings with probability $P(\xi = k) = p_k$ for $k = 0, 1, 2, \ldots$ such that $p_k \geq 0$ and $\sum_k p_k = 1$. The number of offsprings between different individuals are independent. We consider the first set of organisms the **ancestor generation** and each subsequent generation, **generation** $1, 2, 3, \ldots$.

We let X_n be the number of individuals or the population in the n-th generation. We restrict or focus to $X_0 = 1$. Then X_{n+1} is,
$$X_{n+1} = \xi_1^{(n)} + \cdots + \xi_{X_n}^{(n)}$$
Where $\xi_1^{(n)}, \ldots, \xi_{X_n}^{(n)}$ are the independent copies of ξ and $\xi_i^{(n)}$ is the number of offsprings of the i-th individual in generation n.

What is the mean and variance of X_n? First assume that $\mathbb{E}(\xi) = \mu$ and $\mathsf{Var}(\xi) = \sigma^2$. Remember that ξ is the number of offspring produced by an individual of the population. To find $\mathbb{E}(E_n)$ first make use of the $\mathbb{E}(X_{n+1})$.

$$\begin{aligned} \mathbb{E}(X_{n+1}) &= \mathbb{E}\left(\xi_1^{(n)} + \cdots + \xi_{X_n}^{(n)}\right) \\ &= \mathbb{E}\left(\mathbb{E}\left(\xi_1^{(n)} + \cdots + \xi_{X_n}^{(n)} \mid X_n\right)\right) \\ &= \mathbb{E}(\mu X_n) \\ &= \mu\mathbb{E}(X_n) \end{aligned}$$

Recursively if $\mathbb{E}(X_{n+1}) = \mu\mathbb{E}(X_n)$, then we have that,

$$\mathbb{E}(X_n) = \mu^n \mathbb{E}(X_0) = \mu^n \quad \forall n = 0, 1, 2, \ldots$$

Next we find $\mathsf{Var}(X_{n+1})$ using,

$$\mathsf{Var}(X_{n+1}) = \mathbb{E}(\mathsf{Var}(X_{n+1} \mid X_n)) + \mathsf{Var}(\mathbb{E}(X_{n+1} \mid X_n))$$

Examine the first term and make use of Wald's identity,

$$\begin{aligned} \mathbb{E}(\mathsf{Var}(X_{n+1} \mid X_n)) &= \mathbb{E}\left(\mathsf{Var}\left(\xi_1^{(n)} + \cdots + \xi_{X_n}^{(n)} \mid X_n\right)\right) \\ &= \mathbb{E}\left(\sigma^2 X_n\right) \\ &= \sigma^2 \mu \end{aligned}$$

$$\mathsf{Var}(\mathbb{E}(X_{n+1} \mid X_n)) = \mathsf{Var}(\mu X_n) = \mu \mathsf{Var}(X_n)$$

Combining these results and we arrive at the recursive relation,

$$\mathsf{Var}(X_{n+1}) = \sigma^2 \mu^n + \mu^2 \mathsf{Var}(X_n)$$

With initial variance $\mathsf{Var}(X_1) = \sigma^2$. Therefore,

$$\mathsf{Var}(X_2) = \sigma^2 \mu^1 + \mu^2 \sigma^2 = \sigma^2\left(\mu^1 + \mu^2\right)$$

Similarly for $n = 3$,

$$\mathsf{Var}(X_3) = \mu^2 \sigma^2 + \mu^2 \mathsf{Var}(X_2) = \sigma^2\left(\mu^2 + \mu^3 + \mu^4\right)$$

Altogether,

$$\mathsf{Var}(X_n) = \sigma^2\left(\mu^{n-1} + \cdots + \mu^{2n-2}\right)$$

Which becomes,

$$\mathsf{Var}(X_n) = \begin{cases} \sigma^2 \mu^{n-1} \frac{1-\mu^n}{1-\mu} & \mu \neq 1 \\ \sigma^2 n & \mu = 1 \end{cases}$$

Todo (TC Fraser): Missed a lecture

Theorem 20. *An irreducible DTMC with finite state space must be positive recurrent.*

For a branching process, extinction happens with probability $u < 1$ if $\mathbb{E}(\xi) > 1$, where u is the smallest solution between 0 and 1 of $\varphi(s) = s$ (where φ is the generating function for ξ). Extinction happens with probability 1 is $\mathbb{E}(\xi) \leq 1$.

5 Continuous Time Process

Recall that a continuous process is one where $\{X(t)\}_{t\geq 0}$. A counting process is one where,

$$0 \leq s_1 \leq s_2 \leq \cdots$$

Is the occurrence time of the events.

$$N(t) = \#\{n : s_n \leq t\} = \sum_{n=1}^{\infty} \mathbb{1}_{\{s_n \leq t\}}$$

Is the counting process of the events $\{s_n\}_{n=1,2,\ldots}$. Properties:

1. $N(t) \leq 0$ where $t \geq 0$
2. $N(t) \in \mathbb{Z}$
3. $N(t)$ is increasing

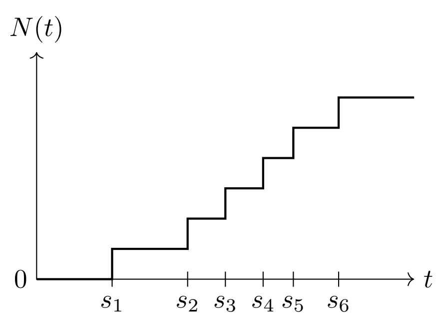
For continuous time processes we will assume some natural properties hold. First we require that,

$$N(0) = 0$$

Which means that we do not expect any events to be happening at time 0 (or earlier). The second assumption is that $N(t)$ only changes over time with size 1; i.e. $N(t)$ *jumps* one at a time. This assumption amounts to the assumption that no two events occur at *exactly* the same time. Thirdly we claim that $N(t)$ is **right-continuous**.

$$N(t) = \lim_{x\to t^+} N(s)$$

As it turns out that this is not an assumption; since $N(t) = \#\{n : s_n \leq t\}$, $N(t)$ *must* be right-continuous.



5.1 Definition of a Poisson Process

Definition 10. The **interarrival times** $W_1, W_2, \ldots$ are the times between subsequent event occurrence times.

$$W_1 = s_1 \qquad W_n = s_n - s_{n-1}$$

Definition 11. A **renewal process** is a counting process where the interarrival times $W_1, W_2, \ldots$ are i.i.d. (independent and identically distributed).

Intuitively after each event s_i, the amount of time we need to wait for the next event is identical to as if we started from the beginning again. In this sense, each time we get a new event s_i, our process is *renewed*.

All of the three previous examples of counting processes Todo (TC Fraser): These are in the last lecture that I missed can be reasonably modeled as renewal processes.

Definition 12. A **Poisson process** $\{N(t)\}_{t\geq 0}$ is a renewal process where the interarrival times $W_1, W_2, \dots$ are not only i.i.d. but are exponentially distributed,

$$W_n \overset{\text{i.i.d}}{\sim} \mathsf{Exp}(\lambda)$$

A Poisson process can be denoted as $\{N(t)\} \sim \mathsf{Poi}(\lambda t)$. Note that the only parameter here λ (not λt) and is called the *intensity*. For each fixed time $N(t)$ is a random variable that has Poisson distribution. We are going to be looking at the properties of a Poisson process. These properties rely heavily on properties of the exponential distribution. Recall that an exponential distribution $X \sim \mathsf{Exp}(\lambda)$ has the following properties:

- pdf: $f(x) = \lambda e^{-\lambda x}$ for all $x \geq 0$
- cdf: $F(x) = 1 - e^{-\lambda x}$ for all $x \geq 0$
- $\mathbb{E}(X) = 1/\lambda$, $\mathsf{Var}(X) = 1/\lambda^2$
- Memoryless property:
$$P(X > s + t \mid X > s) = P(X > t)$$
- Minimum of exponentials: If we have a collection of exponentials,
$$X_1, \dots, X_n \quad X_i \perp\!\!\!\perp X_j \quad X_i \sim \mathbb{E}(\lambda_i)$$
Then,
$$\min\{X_1, \dots, X_n\} = \mathsf{Exp}(\lambda_1 + \cdots + \lambda_n)$$
$$P(X_i = \min(X_1, \dots, X_n)) = \frac{\lambda_i}{\lambda_1 + \cdots + \lambda_n}$$

5.2 Properties of Poisson Processes

To begin, we will define the continuous-time Markov property,

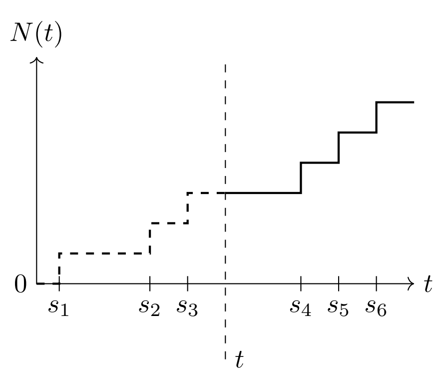
Definition 13. Continuous-Time Markov Property:

$$P\left(\underbrace{N(t_m) = j}_{\text{future}} \mid \underbrace{N(t_{m-1}) = i}_{\text{present}}, \underbrace{N(t_{m-2}) = i_{m-2}, N(t_1) = 1}_{\text{past}}\right) = P(N(t_m) = j \mid N(t_{m-1}) = i)$$

For any m, $t_1 < \cdots < t_{m-1} < t_m$ and $i_1, i_2, \dots, i_{m-2}, i, j \in S$.

Theorem 21. *The Poisson process is the only renewal process having the Markov property.*

The reason being that since the exponential distribution has the memoryless property, the future arrival time future arrival times will not depend on how long we have waited.



At time t, the next arrival time will be exponentially distributed and be independent of the previous arrival times. If the interarrival times have similar distributions (and possibly not exponential), we can attempt to predict the next arrival time based on when the previous arrival time. In these cases, we do not retain the Markov property. Therefore, the fact that the exponential distribution is the only distribution with the memoryless property translates to implying that the Poisson process is the only process with the Markov property. In summayr, the future of the counting process only depends on its current value. Otherwise, we do not need to retain or keep track of any of the previous times of arrival.

In fact we have that for $0 < s < t$,

$$P(N(t) = j \mid N(s) = i) = P(N(t) = j - 1 \mid N(s) = 0)$$

Moreover, the Poisson process has **independent increments**. Let there be time intervals $t_1 < t_2 \leq t_3 < t_4$, then,

$$N(t_2) - N(t_1) \perp\!\!\!\perp N(t_4) - N(t_3)$$

This follows directly (again) from the memoryless property of the exponential distribution. It is important to note that in order for this to hold, the intervals need to have no overlap.

Theorem 22. ***Poisson Increments***: *For Poisson process* $\{N(t)\}_{t \geq 0}$ *and interval* (t_1, t_2) *where* $t_2 < t_1$,

$$N(t_2) - N(t_1) \sim \mathsf{Poi}(\lambda(t_2 - t_1))$$

Proof. The reasoning is as follows. Let the arrival times between t_1 and t_2 be $s_1, \ldots, s_N$ where $N = N(t_2) - N(t_1)$. We let $W_1 = s_1 - t_1$ and $W_i = s_i - s_{i-1}$ are i.i.d. random variables following $\mathsf{Exp}(\lambda)$. For $N = n$ the sum of all interarrival times must be less than $t_2 - t_1$

$$W_1 + \cdots W_n \leq t_2 - t_1$$

While the next event occurs after t_2,

$$W_1 + \cdots W_n + W_{n+1} > t_2 - t_1$$

Now we use the following fact: If $W_1, \ldots, W_n$ are i.i.d r.v.s following $\mathsf{Exp}(\lambda)$ then the sum follows an Erlang distribution,

$$W_1 + \ldots W_n \sim \mathsf{Erlang(n}, \lambda)$$

Where Erlang distributions are just a special type of Gamma distributions. The cdf of an Erlang distribution is,

$$F(x) = 1 - \sum_{k=0}^{n-1} \frac{1}{k!} e^{-\lambda x} (\lambda x)^k$$

Thus,

$$P(W_1 + \cdots + W_n \leq t_2 - t_1) = 1 - \sum_{k=0}^{n-1} \frac{1}{k!} e^{-\lambda(t_2 - t_1)} (\alpha(t_2 - t_1))^k$$

While the addition of W_{n+1},

$$P(W_1 + \cdots + W_{n+1} \leq t_2 - t_1) = 1 - \sum_{k=0}^{n} \frac{1}{k!} e^{-\lambda(t_2 - t_1)} (\alpha(t_2 - t_1))^k$$

Combining these two results,

$$P(N = n) = \frac{1}{n!} e^{-\lambda(t_2 - t_1)} (\alpha(t_2 - t_1))^n$$

Which is nothing more than the pmf for $\mathsf{Poi}(\lambda(t_2 - t_1))$. $\square$

In particular we often look at the interval between 0 and t such that,

$$N(t) = N(t) - N(0) \sim \mathsf{Poi}(\lambda t)$$

Consequently, we have that the expectation of $N(t)$ is the expectation of $\mathsf{Poi}(\lambda t)$,

$$\mathsf{Exp}(N(t)) = \lambda$$

On average, we have λ arrivals in a unit of time t. This interpretation motivates us to call λ the **intensity** of $N(t)$.

5.3 Combining and Thinning Poisson Processes

Suppose we have two Poisson processes $\{N_1(t)\} \sim \mathsf{Poi}(\lambda_1 t)$ and $\{N_2(t)\} \sim \mathsf{Poi}(\lambda_2 t)$. We can combine $N_1(t)$ and $N_2(t)$ in the following way. For any event $s_{1,i}$ in N_1 or $s_{2,i}$ in N_2, we attribute all events to $N(t)$. Mathematically,

$$N(t) = N_1(t) + N_2(t)$$

Definition 14. **Combining:** Let $\{N_1(t)\} \sim \mathsf{Poi}(\lambda_1 t)$ and $\{N_2(t)\} \sim \mathsf{Poi}(\lambda_2 t)$ be two Poisson processes with intensities λ_1 and λ_2 where $N_1(t) \perp\!\!\!\perp N_2(t)$. Then,

$$\{N(t)\}_{t\geq 0} \sim \mathsf{Poi}((\lambda_1 + \lambda_2)t)$$

Intuitively if $N(t)$ is defined in this way, the sum of two Poisson processes is also a Poisson process. The reason being that this is a combination of two properties of exponential random variables. The first property being the memoryless property of exponential distributions and the second being that if $W_1 \sim \mathsf{Exp}(\lambda_1)$ and $W_2 \sim \mathsf{Exp}(\lambda_2)$ are two independent exponential random variables $W_1 \perp\!\!\!\perp W_2$,

$$\min(W_1, W_2) \sim \mathsf{Exp}(\lambda_1 + \lambda_2)$$

The combined process is the counting process of events interarrival times following i.i.d $\mathsf{Exp}(\lambda_1 + \lambda_2)$ which means that $\{N(t)\} \sim \mathsf{Poi}((\lambda_1 + \lambda_2)t)$ where the combined intensity is $\lambda_1 + \lambda_2$. From this analysis we can interpret the intensity as the expected number of events in a unit time.

Thinning a Poisson process is dual to combining a Poisson process. Instead of combining together two or many Poisson processes, thinning a Poisson process reduces the intensity of the process. To do this we classify the events within a Poisson process.

Definition 15. **Thinning:** Let $\{N(t)\} \sim \mathsf{Poi}(\lambda t)$. Each arrival is then labeled as type 1 or type 2 with probability p and $1 - p$ independently from others. Then we can count each type separately. Let $N_1(t)$ and $N_2(t)$ be the number of customers of type 1 and type 2 who arrived before time t. The result of this separation is predictable.

$$\{N_1(t)\} \sim \mathsf{Poi}(p\lambda t)$$
$$\{N_2(t)\} \sim \mathsf{Poi}((1-p)\lambda t)$$

Less predictability we have that $\{N_1(t)\} \perp\!\!\!\perp \{N_2(t)\}$. The counter-intuition being that if an event is not classified as type 1 then it is classified as type 2 with certainty.

The reason for their independence is that thinning is the inverse procedure of combining two independent Poisson processes into one Poisson process.

5.4 Order Statistic Property

Let $X_1, \dots, X_n$ be i.i.d r.v.s. The **order statistics** of $X_1, \dots, X_n$ are random variables defined as follows. The first order statistic:

$$X_{(1)} = \min\{X_1, \dots, X_n\}$$

While $X_{(2)}$ is defined as the second smallest value,

$$X_{(2)} = \min\{\{X_1, \dots, X_n\} \setminus \min\{X_1, \dots, X_n\}\}$$

Continuing we define $X_{(n)}$ to be the maximum value,

$$X_{(n)} = \max\{X_1, \dots, X_n\}$$

Theorem 23. *Let $\{N(t)\} \sim \mathsf{Poi}(\lambda t)$. Condition on $N(t) = n$, the points/arrivals of N in $[0, t]$ are distributed as the order statistics of n i.i.d uniform r.v.s on $[0, t]$. That is,*

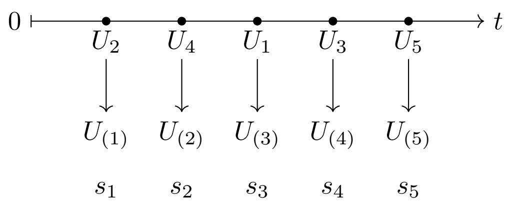
$$\{s_1, \dots, s_n \mid N(t) = n\} \overset{d}{=} \left(U_{(1)}, \dots, U_{(n)}\right)$$

Where '$\overset{d}{=}$' denotes that the two quantities share the same distribution. Also,

$$U_1, \dots, U_n \overset{i.i.d}{\sim} \mathsf{Unif}[0, t]$$

and $U_{(1)}, \dots, U_{(n)}$ are their order statistics.

Example 15. Given $N(t) = 5$,



Let us see the reason for this with $N(t) = 1$. Suppose we have that $s_1 = s$ with $W_2 > t - s$ since the second event has not occurs in $[0, t]$. Then we have the pdf,

$$f_{s_1 \mid N(t)=1}(s) = \frac{f_{s_1}(s) P(W_2 > t - s)}{P(N(t) = 1)} \mathbb{1}_{[0,t]}(s)$$

But we know that W_2 is distributed exponentially,

$$\begin{aligned} f_{s_1 \mid N(t)=1}(s) &= \frac{\lambda e^{-\lambda s} e^{-\lambda(t-s)}}{\lambda t e^{-\lambda t}} \mathbb{1}_{[0,t]}(s) \\ &= \frac{1}{t} \mathbb{1}_{[0,t]}(s) \end{aligned}$$

This is nothing more than the pdf for $\mathsf{Unif}[0, t]$.

$$s_1 \mid N(t) = 1 \sim \mathsf{Unif}[0, t]$$

As a result of the order statistics property, we have the following.

Theorem 24.

$$N(s) \mid N(t) = n \sim \mathsf{Bin}\left(n, \frac{s}{t}\right) \quad \textit{for } s \leq t$$

The reason being that given $N(t) = n$, the order statistics property gives,

$$\begin{aligned} N(s) &= \#\{S_i : S_i \leq s, i = 1, \dots, n\} \\ &= \#\left\{U_{(i)} : U_{(i)} \leq s, i = 1, \dots, n\right\} \\ &= \#\{U_i : U_i \leq s, i = 1, \dots, n\} \end{aligned}$$

The last step validated by the fact that $\left\{U_{(i)}\right\}$ is just a particular permutation of $\{U_i\}$. Therefore $U_i \overset{\text{i.i.d}}{\sim} \mathsf{Unif}[0, t]$ because,

$$P(U_i \leq s) = \frac{s}{t} \quad i = 1, \dots, n$$

6 Continuous-time Markov chain

The Poisson process is the simplest example of a continuous time process; it also happens to be a counting process.

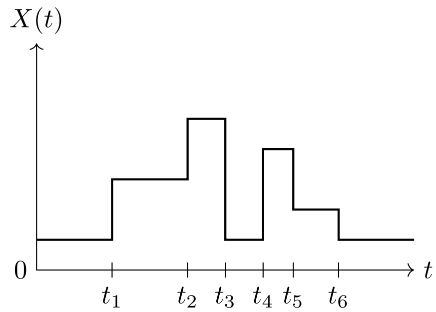
Definition 16. A **continuous time stochastic process** $\{X(t)\}_{t\geq 0}$ is called a **continuous-time Markov chain** is its state space is at most countable and it satisfies the **continuous-time Markov property**.

Recall the continuous-time Markov property (definition 13),

$$P(X(t_m) = j \mid X(t_{m-1}) = i, X(t_{m-2}) = i_{m-2}, \ldots, X(t_1) = 1) = P(X(t_m) = i \mid X(t_{m-1}) = i)$$

For any $m, t_1 < \cdots < t_m, i_1, \ldots, i_{m-2}, i, j \in S$.

For a DTMC, S can be considered as finite $\{0, \ldots, M\}$ or $\{1, \ldots, M\}$ or countably infinite $\{0, 1, \ldots\}$, $\{0, \pm 1, \pm 2, \ldots\} = \mathbb{Z}$. For a CTMC, time is continuous, but the state space is only countable and thus is discrete. Since time is continuous and the state space is discrete, the CTMC must linger/stay at a given state for a finite amount of time and then jump to another state instantaneously.



Therefore we need to specify two things:

1. When do the jumps happen and how long does the process stays in a state?
2. When it jumps, where does it jump to?

To answer the first question, let the process be in state i, it will stay in this state for an exponential random time, with parameter denoted λ_i. This distribution is fixed by the Markov property. When the process jumps in the future only depends on its current state, not on how long it has been there. Therefore the waiting time must be exponentially distributed. Moreover the Markov property enforces that the exponential parameter can only depend on state i.