Stat 433

STOCHASTIC PROCESSES

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 $\mathbb{E}(e^{tX})$ 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,

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

Proof. Evidently,

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

Moreover,

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

□

As a result Taylor series gives,

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

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.

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

□

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.

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

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,

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

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

2 DTMC

2.1 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.

2.2 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.

2.3 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,

$$P_{n,i,j} \equiv P(X_{n+1} = j \mid X_n = i) \qquad n = 0, 1, 2, \ldots \tag{2.1}$$

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 MC 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,

$$P_{i,j} \geq 0 \tag{2.2}$$

$$\forall i : \sum_{j \in S} P_{ij} = 1 \tag{2.3}$$

The entries are non-negative because they represent probabilities and the row sums for P are always unitary.

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

$$\forall i, j \in S : P_{ij}^{(n)} \equiv P(X_{n+m} = j \mid X_n = i) = P(X_n = j \mid X_0 = i)$$

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

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

Theorem 4. *There is a simple relation between the n-step transition matrix $P^{(n)}$ and the one step transition matrix P.*

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

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 complete proof that $P^{(n)} = P^n$.

□

Corollary 5. *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 Equation** or simply C-K equation,*

$$P_{ij}^{(n)} = \sum_{k \in S} P_{ik}^{(m)} P_{kj}^{(n-m)} \quad \forall i, j \in S, \forall 0 \leq m \leq n \tag{2.4}$$

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

Let $\alpha_n = (\alpha_{n,0}, \alpha_{n,1}, \dots)$ 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 $n = 0, 1, 2, \dots$. We also define the initial distribution α_0,

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

Theorem 6. *The transition probability matrix reveals the following relationship between the distribution α_n at time n and the distribution α_0 at time 0,*

$$\alpha_n = \alpha_0 \cdot P^n \tag{2.5}$$

Proof. The proof eq. (2.5) is quite trivial:

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

□

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

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

But by the Markov condition, it must be that,

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

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. (2.1). Therefore it must be that,

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

Even more generally, for $0 \leq t_1 < t_2 < \dots < t_n$,

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

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

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

This means the probabilistic properties of a DTMC are fully characterized by two things:

1. The initial distribution α_0
2. Transition matrix P

2.4 Classification of States

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 $i \leftrightarrow j$.

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

Theorem 7. *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 proving 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. (2.4),

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

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,

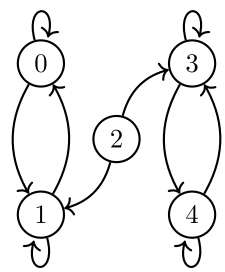
1. $S_i \subset S \mid \forall i \in 1, 2, \dots, n$
2. $S_i \cap S_j \neq \emptyset$ for all $i \neq j$
3. $\bigcup_i S_i = S$

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 6. 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,

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

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**.

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

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

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.

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

Theorem 8. *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 $P_{jj}^l > 0$,

$$P_{ii}^{n+m+l} \geq P_{ij}^n P_{jj}^l P_{ji}^m \tag{2.6}$$

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,

$$P_{ii}^{n+m} \geq P_{ij}^n P_{ji}^m \tag{2.7}$$

Since $d(i)$ divides both $n + m$ and $n + m + l$ by eqs. (2.7) and (2.6), 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 $\left\{l \mid P_{jj}^l > 0\right\}$ an thus $d(i)$ divides,

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

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

$\square$

Remark 3. 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,

$$\begin{tikzcd} 0 \arrow[r, bend right] \arrow[dr, bend right] & 1 \arrow[l, bend right] \arrow[d] \\ & 2 \end{tikzcd}$$

Evidently $P_{00} = 0$ but we have $d(0) = 1$ because $d(0) = \gcd\{2, 3, \ldots\}$.

Remark 4. 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 $i \in S$.

2.5 Recurrence and Transience

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

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

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,

$$P_{ij}^{(n)} = \sum_{k=1}^{n} f_{ij}^{(k)} P_{jj}^{(n-k)} \tag{2.8}$$

Expanded out gives,

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

Proof.

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

□

In fact eq. (2.8) 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 M_i to be the total number of times the MC (re)visits i after time 0. In more mathematical terms,

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

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,

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

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

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

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

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

Alternatively, we can look at $\mathbb{E}(M_i \mid X_0 = i)$ as,

$$\mathbb{E}(M_i \mid X_0 = i) = \sum_{k=1}^{\infty} P(M_i \geq k \mid X_0 = i) \tag{2.9}$$

The proof of eq. (2.9) is left as an exercise to the reader. Clearly if $f_{ii} = 1$,

$$P(M_i \geq k \mid X_0 = i) = {f_{ii}}^k = 1 \quad \forall k \tag{2.10}$$

Therefore,

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

Theorem 9. *Therefore i is recurrent if and only if $P(M_i \geq k \mid X_0 = i) = \infty$ and i is transient if and only if only if $P(M_i \geq k \mid X_0 = i) < \infty$.*

Remark 5. We actually also have that i is recurrent if and only if $M_i = \infty$. This can be seen from eq. (2.10). Since $P(M_i \geq k \mid X_0 = i)$ is strictly positive for all k, then $M_i = \infty$. Analogously, we have that i is transient if and only if $M_i < \infty$.

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

Theorem 10. *The expectation of the indicator is given by $\mathbb{E}(\mathbf{1}_A) = P(A)$ for any event A.*

Therefore,

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

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 11. *Recurrence/transience are class properties. If $i \leftrightarrow j$ and i is recurrent, then j is recurrent.*

Proof. Since $i \leftrightarrow j$, $\exists m, n \geq 0$ such that,

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

We now what to show that $\sum_{s=1}^{\infty} P_{jj}^{(s)}$ is infinite,

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

Now exchange of variables $l = s - n - m$,

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

Then by the eq. (2.4),

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

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 12. *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. Also, it is *impossible* for all states to be transient if the state space S is finite. 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 $i, j \in S$.

Theorem 13. *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; $P_{ij} = 0$. □