Notes on STAT-333:

Applied Probabilities / Stochastic Process I

Unversity of Waterloo

David Duan

Last Updated: September 16, 2021 (Draft)

Contents

Chapter 1. Preparation.

Probability Space and Random Variables

1.1. Definition: A **probability space** consists of a triple $(\Omega, \mathcal{E}, \Pr)$, where

- Ω: **sample space**.
  - The set of all possible outcomes of a random experiment.
- $\mathcal{E}$: **events**.
  - A σ-algebra that contains the collection of all **events**.
  - An **event** is a subset of Ω for which we can talk about probability.
- Pr: **probability measure**.
  - A set function $\mathcal{E} \to \mathbb{R}, E \mapsto \Pr(E)$ satisfying three axioms:
    1. $0 \leq \Pr(E) \leq 1$ for any $E \in \mathcal{E}$.
    2. $\Pr(\Omega) = 1$.
    3. $\Pr(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \Pr(E_i)$ for countable, disjoint events $\{E_i\}_{i=1}^{\infty}$.

1.2. Definition: A **random variable** (r.v.) X is a mapping from Ω to $\mathbb{R}$.

Stochastic Processes: Intuition

1.3. Intuitively, *stochastic* means *random* and *process* means *change/evolve over time.*

1.4. (Cont'd): There are two ways to approach a stochastic process. First, start with a number, we add some randomness to obtain a random number, then let this random variable to change over time. Since by definition a random variable does not change over time (i.e., it does not have a ”time” component), this construction gives us a group of random variables indexed by time stamps defined on the same probability space:

$$\text{Number} \xrightarrow{\text{randomness}} \text{Random Variable} \xrightarrow{\text{change over time}} \text{Stochastic Process.}$$

In other words, a stochastic process can be viewed it as *a sequence/family of random variables.* This is simple and we will take it as definition in this course.

1.5. (Cont'd): The second way is to first add the time component to obtain a function over time, and then add randomness. In other words, a stochastic process can be viewed as a *random function.* This is harder to formulate as we are working with function spaces.

$$\text{Number} \xrightarrow{\text{change over time}} \text{Function (of Time)} \xrightarrow{\text{Randomness}} \text{Stochastic Process.}$$

This will become useful in more advanced contexts.

Stochastic Processes: Definition

1.6. Definition: A **stochastic process** $\{X_t\}_{t\in T}$ is a collection of random variables defined on a common probability space.

1.7. Remark: It can be helpful to think of the index set T as "time" (as this is true in most cases), but the definition does not make such restriction (i.e., T can be other things as well). Note T can be **discrete** (e.g., $T = \{0, 1, 2, \dots\}$) or **continuous** (e.g., $T = [0, \infty)$). In the discrete case, we typically write $\{X_n\}_{n=0,1,2\dots}$, i.e., use n and N for index.

1.8. Definition: The possible values of X_t for all $t \in T$, are called the **states** of the processes. The collection of all states (of all X_t's) is called the **state space**, often denoted by S.

1.9. Remark: The state space can be **discrete** or **continuous**. In this course, we will focus on discrete state space. We may relabel the states in S to get the **standardized state space**, i.e., $\{0, 1, 2, \dots\}$ for a countable state space and $\{0, 1, 2, \dots, n\}$ for a finite state space.

1.10. Example: Let $X_0, X_1, \dots$ be independent normal random variables. Then $\{X_n\}_{n=0,1,\dots}$ is a stochastic process. This is sometimes called the **white noise**.

1.11. Example: Let $X_1, X_2, \dots$ be iid. Suppose for each i,

$$\Pr(X_i = 1) = p$$
$$\Pr(X_i = -1) = 1 - p.$$

Define $S_0 = 0$ and $S_n = \sum_{i=1}^n X_i$ for $n \geq 1$. Then $\{S_n\}_{n=0,1,\dots}$ is a stochastic process, with the state space $S = \mathbb{Z}$. This is known as a **simple random walk** as

$$S_n = S_{n-1} + X_n = \begin{cases} S_{n-1} + 1 & \text{with probability } p \\ S_{n-1} - 1 & \text{with probability } 1 - p \end{cases}$$

Stochastic Processes: Motivation

1.12. But why do we need the notion of stochastic process? Why can't we just look at the joint distribution of rvs? The answer is that the joint distribution of a large number of random variables is very complicated and it does not take advantage of the special structure of time T. For example, the full distribution of $S_0, S_1, \dots, S_n$ in the random walk example is very complicated for large n, and we could greatly simplify the structure by introducing time into the framework.

1.13. (Cont'd): For the simple random walk, we find that if we know S_n, then the distribution of S_{n+1} will not depend on the history of S_i for $i = 0, \dots, n-1$. This is a very useful property and it motivates the notion of **markov chain**.

Conditional Probability

1.14. Definition: The **conditional probability** of an event B given an event A with $\Pr(A) > 0$ is given by

$$\Pr(B \mid A) = \frac{\Pr(B \cap A)}{\Pr(A)}.$$

Events A and B are **independent**, denoted $A \perp B$, iff $\Pr(A \cap B) = \Pr(A)\Pr(B)$.

1.15. Theorem: *Let* $\{A_1, A_2, \ldots\}$ *be a partition of the sample space* Ω, *i.e., they are disjoint events such that* $\bigcup_i A_i = \Omega$.

- *Law of Total Probability:*

$$\Pr(B) = \sum_i \Pr(B \mid A_i) \cdot \Pr(A_i).$$

- *Bayes' Theorem:*

$$\Pr(A_i \mid B) = \frac{\Pr(B \mid A_i) \cdot \Pr(A_i)}{\sum_j \Pr(B \mid A_j) \cdot \Pr(A_j)}.$$

Conditional Distribution

1.16. Definition: Let X, Y be discrete random variables and suppose $\Pr(Y = y) > 0$ for a specific y. Then the **conditional distribution** of X given $Y = y$ is given by

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

$\Pr(X = y \mid Y = y)$ is called the **conditional probability mass function**, denoted by $f_{X|Y=y}(x)$ or $f_{X|Y}(x \mid y)$.[1]

1.17. Remark: The continuous case is similar: just replace pmf with pdf when computing probabilities. Note we are primarily interested in the discrete case in this course because we will be dealing with discrete state spaces most of the times.

1.18. Proposition: *Conditional pmf is a valid pmf, i.e., for any* y *with* $\Pr(Y = y) > 0$,

- $f_{X|Y=y}(x) \geq 0$ *for all* x, *and*
- $\sum_x f_{X|Y=y}(x) = 1$.

1.19. Remark: Since the conditional pmf is a valid pmf, the conditional distribution is a valid probability distribution. Intuitively, it is a (potentially different) probability distribution after we acquired new knowledge $Y = y$. Having seen it is a valid distribution, we can define *expectation* for this distribution, known as *conditional expectation*.

[1] Be very careful here: y is a condition, not an argument.

Conditional Expectation

1.20. Definition: Let X, Y be discrete random variables and g be a function. Then the **conditional expectation** of $g(X)$ given $Y = y$ is given by

$$\mathbb{E}[g(X) \mid Y = y] = \sum_x g(x) \Pr(X = x \mid Y = y).$$

In other words, the conditional expectation is the expectation under the conditional distribution with an additional condition $Y = y$.

1.21. Note: We now present three different approaches to understand/interpret conditional expectation.

1. First, fix y, $\mathbb{E}[g(X) \mid Y = y]$ is just a number.
2. As y changes, $h(y) := \mathbb{E}[g(X) \mid Y = y]$ is a function of y.
3. Since Y is a random variable, we can define $\mathbb{E}[g(X) \mid Y] =: h(Y)$. Since $\mathbb{E}[g(X) \mid Y]$ is a function of a random variable Y, it is also a random variable.[2]

1.22. (Cont'd): Note $\mathbb{E}[g(X) \mid Y]$ is a random variable, so we want to determine its value for each event $\omega \in \Omega$, denoted $\mathbb{E}[g(X) \mid Y]_\omega$. We define

$$\mathbb{E}[g(X) \mid Y]_\omega = \mathbb{E}[g(X) \mid Y = Y(\omega)], \qquad \omega \in \Omega.$$

Again, since Y is a random variable and ω is an event, $Y(\omega)$ is a real number. In other words, you are substituting in $y = Y(\omega)$ to evaluate the conditional expectation here.

Properties of Conditional Expectation

1.23. Proposition: *Conditional expectation is linear:*

$$\mathbb{E}[aX + bY + c \mid Y = y] = a \cdot \mathbb{E}[X \mid Y = y] + b \cdot \mathbb{E}[X \mid Y = y] + c.$$

Proof. Inherited from expectation. □

1.24. Proposition: *Plug-in property:*

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

Proof. We prove the discrete case. Observe

$$\mathbb{E}[g(X, Y) \mid Y = y) = \sum_{x_i} \sum_{y_j} g(x_i, y_j) \cdot \Pr(X = x_i, Y = y_j \mid Y = j)$$

[2] Note in this case we did not fix the value of Y. $h(Y)$ here is a function of a random variable, not a function of a fixed argument (as in the second case). This introduces randomness to h.

If $y_j = y$, then

$$\Pr(X = x_i, Y = y_j \mid Y = y) = \frac{\Pr(X = x_i, Y = y)}{\Pr(Y = y)} = \Pr(X = X_i \mid Y = j)$$

Otherwise (if $y_j \neq y$), $\Pr(X = x_i, Y = y_j \mid Y = y) = 0$. Continuing from above,

$$\begin{aligned}\mathbb{E}[g(X,Y) \mid Y = y] &= \sum_{x_i}\sum_{y_j} g(x_i, y_j) \cdot \Pr(X = x_i, Y = y_j \mid Y = j) \\ &= \sum_{x_i} g(x_i, y) \cdot \Pr(X = x_i \mid Y = y) \\ &= \mathbb{E}(g(X, y) \mid Y = y).\end{aligned}$$

as desired. Note that $\mathbb{E}(g(X, y) \mid Y = y)$ is a function of X as y is fixed. □

1.25. Remark: Note that $\mathbb{E}[g(X,Y) \mid Y = y] \neq \mathbb{E}[g(X, y)]$ in general. Intuitively, the value of $Y = y$ might influence the value of X as well, so only plugging in $Y = y$ does not allow you to drop the condition.

1.26. Corollary: *Let $g(X), h(Y)$ be two functions. If we are given $Y = y$, then we can take $h(Y)$ out of the conditional expectation, i.e.,*

$$\begin{aligned}\mathbb{E}[g(X)h(Y) \mid Y = y] &= \mathbb{E}[g(X)h(y) \mid Y = y) \\ &= h(y)\mathbb{E}[g(X) \mid Y = y].\end{aligned}$$

Since this holds for all values of y, we can derive the following random variable:

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

1.27. Proposition: *If $X \mid Y$, then $\mathbb{E}[g(X) \mid Y] = \mathbb{E}[g(X)]$.*

Proof. Since $X \mid Y$, the conditional distribution is the same as the unconditional distribution, so the conditional expectation is the same as the unconditional expectation. □

Law of Iterated Expectation

1.28. Theorem:

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X].$$

Proof. Before we start, note that $\mathbb{E}[X \mid Y]$ is a random variable; it is a function of Y. As before, we prove the discrete case. Let's look at what values this random variable can take. When $Y = y_j$, then the random variable $\mathbb{E}[X \mid Y] = \mathbb{E}[X \mid Y = j] = \sum_{x_i} \Pr(X = x_i \mid Y = y_j)$, which is now just a number. This happens with probability $\Pr(Y = y_j)$. Thus,

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \sum_{y_j} \mathbb{E}[X \mid Y = y_j] \cdot \Pr(Y = y_j)$$

$$
\begin{aligned}
&= \sum_{y_j} \left(\sum_{x_i} \Pr(X = x_i \mid Y = y_j) \right) \Pr(Y = y_j) \\
&= \sum_{x_i} x_i \sum_{y_j} \Pr(X = x_i \mid Y = y_j) \cdot \Pr(Y = y_j) \\
&= \sum_{x_i} x_i \Pr(X = x_i) && \text{Law of total probability} \\
&= \mathbb{E}[X].
\end{aligned}
$$

□

1.29. Example: Let Y be the number of claims received by an insurance company and X be some random parameters that influences Y. In particular,

$$
\begin{aligned}
X &\sim \text{Exponential}(\lambda) \\
Y \mid X &\sim \text{Poisson}(X)
\end{aligned}
$$

Our goal is to find $\mathbb{E}[Y]$. Note that to find the expectation of a random variable, don't rush to determine its distribution. You can often derive the expectation of a random variable with nice properties of the expectation operator, as we illustrate here.

By the Law of Iterated Expectation, $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid X]]$. Since $Y \mid X$ follows a Poisson distribution, for a fixed $X = x$, $\mathbb{E}[Y \mid X = x] = x$. Since this holds for every x, we have $\mathbb{E}[Y \mid X] = X$. Thus,

$$
\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}[\mathbb{E}[Y \mid X]] \\
&= \mathbb{E}[X] && Y \mid X \sim \text{Poisson}(X) \\
&= 1/\lambda && X \sim \text{Exponential}(X).
\end{aligned}
$$

Part I

Discrete-Time Markov Chain.

Contents

Chapter 2. Basics of DTMCs

Section 1. Definition of DTMC

2.1. Definition: A discrete-time stochastic process $\{X_n\}_{n=0}^{\infty}$ is called a **discrete time Markov chain** (**DTMC**) with **transition matrix** $P = \{P_{i,j}\}_{i,j\in S}$ if for any index $n \in \mathbb{N}$ and any states $j, i, i_{n-1}, \ldots, i_0 \in S$, the following condition holds:

$$\boxed{\Pr(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P_{i,j}.} \tag{2.1}$$

This condition known as the **Markov property**.

2.2. Intuition: Suppose we are at time n.

$$\Pr(\underbrace{X_{n+1} = j}_{\text{future}} \mid \underbrace{X_n = i}_{\text{present}}, \underbrace{X_{n-1} = i_{n-1}, \ldots, X_0 = i_0}_{\text{history / past}}) = P_{i,j}.$$

Intuitively, the Markov property says that *given the current state, the history and the future are independent*. Some equivalent statements include:

- *The past influences the future only through the current state*.
- *If you know the current state, then knowing the past will not help you predict the future*. The other direction also holds, that is, *if you know the current state, then knowing the future will not help you trace back your past*.

2.3. Remark: A more general form of Markov property is as follows:

$$\Pr(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \Pr(X_{n+1} = j \mid X_n = i). \tag{2.2}$$

Note this version states that the transition probability depends on three variables:

- current state i,
- future state j,
- current time n,

while the Markov property in Definition 2.1 has only two variables, current state i and future state j. Thus, in addition to (2.2), Definition 2.1 requires that the transition probability

$$\Pr(X_{n+1} = j \mid X_n = i)$$

does not depend on n, i.e.,

$$\forall n \geq 1 : \Pr(X_{n+1} = j \mid X_n = i) = P_{i,j}.$$

This property is called **time-homogeneity**. In this course, we only consider time-homogeneous discrete-time Markov chains.

Section 2. Transition Matrix

2.4. Note: Let's take a closer look at the transition matrix P of a DTMC.

$$P = \{P_{ij}\}_{i,j\in S} = \begin{bmatrix} P_{00} & P_{01} & \cdots & P_{0j} & \cdots \\ P_{10} & P_{11} & \cdots & P_{1j} & \cdots \\ \vdots & \vdots & & \vdots & \\ P_{i0} & P_{i1} & \cdots & P_{ij} & \cdots \\ \vdots & \vdots & & \vdots & \end{bmatrix}$$

It is important to note that the *rows correspond to the initial/starting/current state* and the *columns correspond to the ending/target/next state*. For example, P_{ij} in the ith row, jth column corresponds to the (one-step) transition probability from state i to state j.

2.5. Note: There are two key properties of a transition matrix. First, *all entries in P has to be non-negative* as they encode probabilities:

$$\forall i, j \in S : P_{i,j} \geq 0. \tag{2.3}$$

Next, *the row sums of P are always 1*:

$$\forall i \in S : \sum_{j\in S} P_{i,j} = 1. \tag{2.4}$$

To see this, fix $i \in S$ as the starting state. The entries in the i-th row in P correspond to the transition probability from state i to each state $j \in S$. Since we must arrive at some state in S, the probability sum to 1:

$$\sum_{j\in S} P_{i,j} = \sum_{j\in S} \Pr(X_{n+1} = j \mid X_n = i) = \Pr(X_{n+1} \in S \mid X_{n=i}) = 1.$$

Any (square) matrix that satisfies these two conditions qualify to be the transition matrix for some DTMC.

2.6. Example (Simple Random Walk)**:** Recall the simple random walk example. Let S_n denote the state of the walk at time n. We have $S_{n+1} = S_n + X_{n+1}$ where $\Pr(X_{n+1} = 1) = p$ and $\Pr(X_{n+1} = -1) = 1 - p$. Let us find $P_{i,j} = \Pr(S_{n+1} = j \mid S_n = i)$ for all $i, j \in \mathbb{Z}$, which gives the transition matrix of SRW. Observe that

$$\begin{aligned} P_{ij} &= \Pr(S_n + X_{n+1} = j \mid S_n = i) \\ &= \Pr(i + X_{n+1} = j \mid S_n = i) && S_n = i \text{ is given} \\ &= \Pr(X_{n+1} = j - i \mid S_n = i) \\ &= \Pr(X_{n+1} = j - i) && S_n \perp\!\!\!\perp X_{n+1} \\ &= \begin{cases} p & j = i + 1 \\ 1 - p & j = i - 1 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

2.7. Remark: As an easy exercise, show SRW is a DTMC. [1]

2.8. Example (The Ehrenfest's Urn)**:** Suppose there are two urns A and B containing M balls in total. At each time stamp, we pick one ball uniformly randomly and put it into the opposite urn. Define X_n to be the number of balls in urn A after n steps, i.e., $X_n = i$ means there are i balls in A and $M - i$ balls in B. Then $\{X_n\}_{n\in\mathbb{N}}$ is a (discrete) stochastic process with state space $S = \{0, 1, \dots, M\}$. Show that the transition matrix is given by

$$P_{i,j} = \Pr(X_{n+1} = j \mid X_n = i) = \begin{cases} \dfrac{i}{M} & j = i - 1 \\ \dfrac{M-i}{M} & j = i + 1 \\ 0 & \text{otherwise} \end{cases}$$

Solution. Suppose $X_n = i$, i.e., after n steps we have i balls in urn A. Then there are two possible outcomes after the next move. If we move a ball from A to B, then we end up with $j = i - 1$ balls in A, i.e., $X_{n+1} = i - 1$. Since there are i balls in A, the probability that we picked a ball from urn A is i/M. The other case is similar, where we move a ball from urn B to urn A, so urn A ends up with $j = i + 1$ balls. This case has probability $(M - i)/M$. It is impossible for urn A to contain any other number of balls, so the probability is 0 otherwise. □

[1] Hint: Show that the Markov property holds for SRW.

Section 3. Multi-Step Transition Matrix

2.9. We are now interested in the behavior of the DTMC in n steps (rather than 1 step). Define the **k-step transition probability** from state i to state j as

$$P_{ij}^{(n)} := \Pr(X_n = j \mid X_0 = i) = \Pr(X_{m+n} = j \mid X_m = i), \quad m = 1, 2, \ldots$$

Again, we assume **time-homogeneity**, i.e., the starting time m is irrelevant.

2.10. Example: Observe that the 2-step transition matrix is given by $P^{(2)} = P^2$:

$$\begin{aligned} P_{ij}^{(2)} &= \Pr(X_j = j \mid X_0 = i) \\ &= \sum_{k \in S} \Pr(X_2 = j \mid X_0 = i, X_1 = k) \cdot \Pr(X = k \mid X_0 = i) && \text{see below} \\ &= \sum_{k \in S} \Pr(X_2 = j \mid X_1 = k) \cdot \Pr(X_1 = k \mid X_0 = i) && \text{see below} \\ &= \sum_{k \in S} P_{kj} P_{ik} = \sum_{k \in S} P_{ik} P_{kj} && \text{definition of } P_{ij} \\ &= (PP)_{ij} = (P^2)_{ij} && \text{matrix multiplication} \end{aligned}$$

- Line 2: Law of Total Probability, conditioned on $X_0 = i$.
- Line 3: Markov property.

We are now ready to look at the general case.

2.11. Theorem (Chapman-Kolmogorov)**:** *The $m + n$ step transition matrix is given by*

$$P^{(m+n)} = P^{(m)} P^{(n)}. \tag{2.5}$$

In other words,

$$\Pr(X_{m+n} = j \mid X_0 = i) = (P^{(m)} P^{(n)})_{ij}. \tag{2.6}$$

Proof. For $n, m \in \mathbb{Z}^+$,

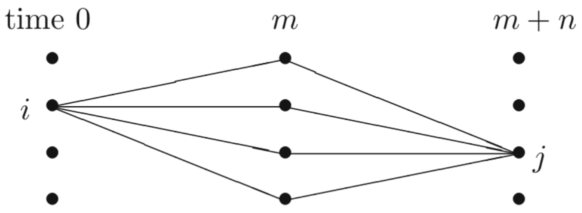
$$\begin{aligned} P_{ij}^{m+n} &= \Pr(X_{m+n} = j \mid X_0 = i) \\ &= \sum_{k \in S} \Pr(X_{m+n} = j \mid X_m = k, X_0 = i) \Pr(X_m = k \mid X_0 = i) \\ &= \sum_{k \in S} \Pr(X_{m+n} = j \mid X_m = k) \Pr(X_m = k \mid X_0 = i) \\ &= \sum_{k \in S} P_{kj}^{(n)} P_{ik}^{(m)} = \sum_{k \in S} P_{ik}^{(m)} P_{kj}^{(n)} = (P^{(m)} P^{(n)})_{ij}. \end{aligned}$$

We introduced an intermediate state k on line 2, then used Markov property to get line 3: knowing $X_m = k$ renders $X_0 = i$ irrelevant in finding the probability of $X_{m+n} = j$. The rest are just definition and elementary algebra. □

2.12. Corollary: *The n-step transition matrix is given by $P^{(n)} = P^n$.*

Proof. Trivial. □

2.13. Intuition: Here's more intuition on the (proof of the) CK equation above.



Suppose we are at state i at time 0. To go from state i to state j in $m+n$ steps, we have to go from i to some state k in m steps and then from k to j in n steps. The Markov property implies that the two parts of our journey are independent. This justifies the second line in the proof above.

2.14. Remark: DTMCs can be represented by weighted DAGs:

- States $\mapsto$ Nodes.
- (One-Step) Transitions $\mapsto$ Edges.
- Transition Probabilities $\mapsto$ Edge Weights.

Section 4. Distribution of X_n

2.15. Motivation: So far, we have examined transition probabilities

$$P_{ij}^{(n)} = \Pr(X_n = j \mid X_0 = i) = P_{ij}^n.$$

If the DTMC *always* starts from state i, then P_{ij}^n is the probability that $X_n = j$, and thus $\{P_{ij}^n\}_{j \in S}$ is the distribution of X_n. In other words, the ith row of P^n is the distribution of X_n if the chain starts from state i with probability 1. But what if the DTMC has a random starting state? Let us first introduce the notion of initial distribution.

2.16. Definition: Define $\mu_n(i) = \Pr(X_n = i)$ and

$$\mu_n = \begin{bmatrix} \mu_n(0) & \mu_n(1) & \cdots & \mu_n(i) & \cdots \end{bmatrix}.$$

This row vector μ_n is called the **distribution** of X_n. The case where $n = 0$ is known as the **initial distribution** of a DTMC and is often denoted by $\mu = \mu_0$.

2.17. Remark: The row vector μ_n represents a distribution, hence we have

- $\mu_n(i) \geq 0$ for all $i \in S$.
- $\sum_{i \in S} \mu_n(i) = 1$.

We sometimes also write $\mu_n(X_n = i)$. In this case, we may view μ_n as a probability function.

2.18. Proposition: *The distribution of X_n is can be obtained by right-multiplying the n-step transition matrix to the initial distribution, i.e.,*

$$\mu_n = \mu \cdot P^n. \tag{2.7}$$

Proof. For any $j \in S$, the jth component in μ_n is given by

$$\begin{aligned} \mu_n(j) &= \Pr(X_n = j) \\ &= \sum_{i \in S} \Pr(X_n = j \mid X_0 = i) \cdot \Pr(X_0 = i) \\ &= \sum_{i \in S} \mu(i) \cdot P_{ij}^{(n)} \\ &= (\mu \cdot P^{(n)})_j = (\mu \cdot P^n)_j \end{aligned}$$

- Line 2: Law of Total Probability.
- Line 3: Definition of $\mu(i)$ and P_{ij}^n.

□

2.19. Remark: By this proposition, the distribution of a DTMC (that is, the distribution of all random variables $\{X_i\}_{i=0}^{\infty}$ of the DTMC) is completely characterized by its initial distribution μ and its transition matrix P.

Section 5. Expectation of $f(X_n)$

2.20. Motivation: Suppose we are interested in the expected reward/penalty we receive at time stamp n based on the state X_n. Define the **reward function** f as

$$\begin{aligned} f : S &\to \mathbb{R} \\ i &\mapsto f(i) \end{aligned}$$

Alternatively, f can be thought of as a row vector, i.e., $[f(0), f(1), \ldots]^T$, since its domain is the set of natural numbers. Our goal is then to evaluate $\mathbb{E}[f(X_n)]$. We present two approaches to derive its expression; both yield the same answer.

2.21. Theorem: *Let μ be the initial distribution, P be the transition matrix, and $f : S \to \mathbb{R}$ be the reward function (row vector). The expected reward at time n is given by*

$$\mathbb{E}[f(X_n)] = \mu P^n f^T. \tag{2.8}$$

Proof 1. The first approach is straightforward. Recall the distribution of X_n is given by the row vector μ_n. More specifically, $\mu_n(i) = \Pr(X_n = i)$ for $i \in S$. Thus,

$$\begin{aligned} \mathbb{E}[f(X_n)] &= \sum_{i \in S} f(i) \cdot \Pr(X_n = i) \\ &= \sum_{i \in S} f(i) \cdot \mu_n(i) && \mu_n(i) = \Pr(X_n = i) \\ &= \mu_n f^T \\ &= \mu P^n f^T && \mu_n = \mu P^n \end{aligned}$$

Since μ is a row vector, P^n is a square matrix, and f^T is a column vector, the result is a scalar as desired. □

Proof 2. Alternatively, we can calculate $P^n f^T$ first. By the Law of Iterated Expectation,

$$\begin{aligned} \mathbb{E}[f(X_n)] &= \mathbb{E}[\mathbb{E}[f(X_n) \mid X_0 = i]] \\ &= \sum_{i \in S} \mathbb{E}[f(X_n) \mid X_0 = i] \cdot \Pr(X_0 = i) \\ &= \sum_{i \in S} \mathbb{E}[f(X_n) \mid X_0 = i] \cdot \mu(i) \\ &=: \sum_{i \in S} f^{(n)}(i) \cdot \mu(i) \\ &= \mu \cdot (f^{(n)})^T \end{aligned}$$

where the row vector

$$f^{(n)} = \begin{bmatrix} \mathbb{E}[f(X_n)|X_0 = 0] \ , & [\mathbb{E}[f(X_n)|X_0 = 1] \ , & \cdots \ , & \mathbb{E}[f(X_n)|X_0 = i] \ , & \cdots \end{bmatrix}^T$$

can be interpreted as a function of expected reward at time n given your starting point and can be found by

$$\begin{aligned}(f^{(n)})^T(i) &= \mathbb{E}[f(X_n) \mid X_0 = i] \\ &= \sum_{j \in S} f(j) \cdot \Pr(X_n = j \mid X_0 = i) \\ &= \sum_{j \in S} P_{ij}^n f(j) \\ &= (P^n \cdot f^T)_i.\end{aligned}$$

Thus, $(f^{(n)})^T = P^n f^T$ and $\mathbb{E}[f(X_n)] = \mu P^n f^T$, which matches our previous approach. □

2.22. Note: In summary, there are two ways to interpret this result.

- If we calculate μP^n first, then

$$\mathbb{E}[f(X_n)] = \mu_n f^T,$$

i.e., standing at n, find the distribution of X_n, then multiply it by the reward vector.

- If we calculate $P^n f^T$ first, then

$$\mathbb{E}[f(X_n)] = \mu \cdot \begin{bmatrix} \mathbb{E}[f(X_n) \mid X_0 = 0] \\ \mathbb{E}[f(X_n) \mid X_0 = 1] \\ \vdots \end{bmatrix},$$

i.e., compute the expected reward at X_n for each starting state, then weight the result by the initial distribution μ.

2.23. Remark: Conventions:

- Vectors are row vectors by default.
- Row vectors represent distributions, e.g., $\mu, \mu^{(n)}$, etc.
- Column vectors represent functions, e.g., $f^T, (f^{(n)})^T$, etc.

Section 6. Stationary Distribution

2.24. Definition: A probability distribution $\pi = (\pi_0, \pi_1, \dots)^T$ is called a **stationary** or **invariant distribution** of a DTMC $\{X_n\}_{n=0}^{\infty}$ with transition matrix P if

1. $\pi = \pi P$ ("*stationarity condition*", as the distribution will not change over time).
2. $\sum_{i \in S} \pi_i = \pi^T \mathbf{1} = 1$ ("*normalization condition*, so π is a probability distribution).

Intuitively, if a DTMC starts from a stationary distribution π, then its distribution never changes: $\mu = \mu^{(1)} = \cdots = \mu^{(n)} = \cdots = \pi$.

2.25. Example: An electron has two states: *ground state* (0) and *excited state* (1). Let $X_n \in \{0, 1\}$ be its state at time n. At each step, the electron changes state with probability α if it is in state 0, with probability β if it is in state 1. Then $\{X_n\}$ is a DTMC with transition matrix

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

To find a stationary distribution, we just need to solve $\pi = \pi P$:

$$\begin{bmatrix} \pi_0 & \pi_1 \end{bmatrix} \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\alpha \end{bmatrix} = \begin{bmatrix} \pi_0 & \pi_1 \end{bmatrix}$$
$$\pi_0(1-\alpha) + \pi_1\beta = \pi_0$$
$$\pi_0\alpha + \pi_1(1-\beta) = \pi_1$$

We have two equations and two unknowns. However, they are linearly dependent, meaning that you could derive one equation from the other. In particular, subtracting the first equation from the identity $\pi_0 + \pi_1 = \pi_0 + \pi_1$ gives the second equation. Thus, the second one is redundant. Now from equation 1, we have

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

Combine this with the normalization condition, we have

$$\pi_0 = \frac{\beta}{\alpha+\beta}, \; \pi_1 = \frac{\alpha}{\alpha+\beta},$$

which is the unique stationary distribution.

2.26. Note: Here's the typical procedure of computing the stationary distribution:

1. Use $\pi = \pi P$ to get proportions among different components of π.
2. Use $\sum_{i \in S} \pi_i = 1$ to solve for exact values (to normalize the values).

You are never able to solve the system of equations just using the stationarity condition because the system is homogeneous, that is, if π is a solution, then $a\pi = (a\pi_0, a\pi_1, \dots)^T$ is also a solution. Also, note that π is the transpose of an eigenvector of P with eigenvalue 1.

Chapter 3. Recurrence and Transience

Section 1. Definition of Recurrence and Transience

3.1. Definition: Let $y \in S$ be a state.

- Define $T_y = \min\{n \geq 1 : X_n = y\}$ as the time of the first (re)visit to y.
- Define $\rho_{yy} = \Pr_y(T_y < \infty) = \Pr(T_y < \infty \mid X_0 = y)$ as the probability that the DTMC ever revisits state y if it starts at y.

3.2. Definition: A state $y \in S$ is called

- **recurrent** if $\rho_{yy} = 1$;
- **transient** if $\rho_{yy} < 1$.

In words, starting at a state y, it is recurrent if we always return to it after a finite number of steps; it is transient if there's a chance that we never revisit it again.

3.3. Example: Consider a DTMC with transition matrix

$$p = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \\ & \frac{1}{2} & \frac{1}{2} \\ & & 1 \end{bmatrix}$$

We claim that state 0 is transient. Starting at $X_0 = 0$, there are two outcomes for X_1:

- If $X_1 = 0$, then $T_0 = 1$.
- If $X_1 = 1$, then $T_0 = \infty$ as it is impossible to return to state 0.

Thus,

$$\rho_{00} = \Pr_0(T_0 < \infty) = \Pr_0(X_1 = 0) = \frac{1}{2} < 1 \implies \text{state 0 is transient.}$$

We claim that state 2 is recurrent. Starting at $X_0 = 2$, we can only go to 2 immediately, so

$$\rho_{22} = \Pr_2(T_2 < \infty) = \Pr_0(X_1 = 2) = 1 \implies \text{state 2 is recurrent.}$$

3.4. Remark: In practice, the definition of T_i is often hard to derive and it won't be as straightforward as the example above to identify recurrent and transient states. We will see better criteria in the next few sections.

Section 2. Communication

3.5. Definition: Let $x, y \in S$ (possibly the same state). Then x is said to **communicate** to y, or y is **accessible** from x, denoted by $x \to y$, if starting from x, the probability that the DTMC *eventually* (re)visits state y is positive, i.e.,

$$\rho_{xy} := \Pr_x(T_y < \infty) > 0.$$

Equivalently, x communicate to y if we can reach y from x in a finite number of steps, i.e.,

$$\exists n \geq 1 : P^n_{xy} > 0.$$

3.6. Remark: Note the notion of communication is one-way only!

$$(x \to y) \not\equiv (y \to x).$$

We will use $x \leftrightarrow y$ to denote double-way communication.

3.7. Lemma: *The relation $x \to y$ is transitive.*

Proof. Suppose $x \to y$ and $y \to z$. By the condition above,

$$\exists m \geq 1 : P^m_{xy} > 0, \qquad \exists n \geq 1 : P^n_{yz} > 0.$$

By CK equation,

$$P^{m+n}_{xz} = \sum_{k \in S} P^m_{xk} P^n_{kz} \geq P^m_{xy} P^n_{yz} > 0.$$

Indeed, $P^m_{xy}P^n_{yz}$ specifies one way to to from x to z in $m+n$ steps (through y) which P^{m+n}_{xz} is the total probability to go from x to z in $m+n$ steps. It follows that $x \to z$. □

3.8. Theorem: *If $\rho_{xy} > 0$ but $\rho_{yx} < 1$, then x is transient.*

Proof. Let $K := \min\{k \mid P^k_{xy} > 0\}$ to be the length of a shortest x, y-path. Since $P^K_{xy} > 0$, there exists states $y_1, \ldots, y_{K-1}$ such that $P_{xy_1}P_{y_1y_2} \cdots P_{y_{K-1}y} > 0$. Note this product of probabilities is just the probability that we go from x to y via this path. Moreover, none of the intermediate states is x, or we obtain a shorter path, contradicting the definition of K. Once we are in state y, with probability $1 - \rho_{yx} > 0$, we will never go back to x.

Above we described one case where we started from x, go to y, then never go back to x. The total probability that we start at x, go to y, and never go back to x, is at least this much:

$$\Pr_x(T_x = \infty) \geq P_{xy_1}P_{y_1y_2} \cdots P_{y_{K-1}y} \cdot (1 - P_{yx}) > 0.$$

Thus, $\rho_{xx} = \Pr_x(T_x < \infty) = 1 - \Pr_x(T_x = \infty) < 1$ and x is not recurrent. □

3.9. Corollary: *If x is recurrent and $\rho_{xy} > 0$, then $\rho_{yx} = 1$.*

Proof. Omitted. □

Section 3. Communicating Classes

3.10. Definition: A set of states $C \subseteq S$ is called a **communicating class** if it satisfies the following conditions:

1. $\forall i, j \in C : i \to j$ and $j \to i$.
2. $\forall i \in C, j \notin C, i \not\to j$ or $j \not\to i$.

In words, states in the same class communicate with each other and states in different classes do not communicate in both ways.

3.11. Note: The condition $i \to j$ there exists a i, j-dipath in the graphical representation of DTMC. Thus, identifying classes can be reduced to identifying the circuits. Moreover, since the relation is transitive, we can "merge" classes together.

3.12. Remark: The set of communicating classes does not partition the set of states. In particular, if a state has only out-arcs (and no in-arcs) then by our definition it does not belong to any communicating class.

3.13. Definition: A DTMC is called **irreducible** if all its states are in the same class, i.e., $i \leftrightarrow j$ for all $i, j \in S$. A set B is called **irreducible** if $i \leftrightarrow j$ for all $i, j \in B$.

3.14. Note: We will later show that recurrence and transience are class properties. Because of this, we can classify a class to be recurrent or transient. Moreover, to classify a class, it suffices to check only one state from the class.

3.15. Definition: A set of states A is called **closed** if

$$i \in A, j \notin A \implies P_{ij} = 0.$$

Equivalently, $A \subseteq S$ is closed if

$$i \in A, j \notin A \implies i \not\to j.$$

3.16. Intuition: Once the chain goes into a closed set A, it cannot get out. Graphically, the subgraph with vertex set A has no out-arcs (but in-arcs are fine).

Section 4. Decomposition of the State Space

3.17. Theorem: *The state space S can be written as a disjoint union*

$$S = T \,\dot{\cup}\, R_1 \,\dot{\cup}\, R_2 \,\dot{\cup}\, \cdots$$

where T is the set of all transient states (not necessarily one class) and R_i's are closed recurrent classes.

Proof. First, collect all the transient states and put them into the set T. Next, each recurrent state belongs to some recurrent class at it communicates to itself at least. For each recurrent state, take one class containing it, put these classes together, and remove the repeated ones. Denote these by $R_1, R_2, \dots$ Then by construction, $S = T \,\dot{\cup}\, R_1 \cup R_2 \cup \cdots$. It remains to show that the R_i's are closed and disjoint from each other.

First, suppose there exist two sets $R_m \cap R_n \neq \varnothing$ with $m \neq n$. Let $i \in R_m \cap R_n$. Then for any $j \in R_m$ and $k \in R_n$,

$$i \leftrightarrow j, \quad i \leftrightarrow k \implies j \leftrightarrow k.$$

Thus, j, k are in the same class. Since this holds for every $j \in R_m$ and $i \in R_k$, R_m and R_n are the same class. Contradiction.

Now suppose there is R_k not closed, so there exists $i \in R_k$, $j \notin R_k$, and $i \to j$, or equivalently, $\rho_{ij} > 0$. However, $j \notin R_k$, so $i \not\leftrightarrow j$, in particular $j \not\to i \iff \rho_{ji} = 0 < 1$. By definition, $\rho_{ij} > 0$ but $\rho_{ji} < 1$ means i is transient, contradicting the definition of R_k. □

Section 5. Recurrence and Transience II

3.18. Motivation: Recall that if a DTMC is time-homogeneous, then

$$\forall n \geq 1 : \Pr(X_{n+1} = j \mid X_n = i) = P_{i,j}.$$

Let $T_y := \min\{n \geq 1 : X_n = y\} \in \mathbb{Z}_+$ denote the time of the first (re)visit to state y. Consider the stochastic process (in fact, a DTMC, as it satisfies Markov property) given by

$$\{X_{T_y}, X_{T_y+1}, X_{T_y+2}, X_{T_y+3}, \ldots\}.$$

Intuitively, we ignored the behaviours of the chain until it first (re)visits state y at time T_y. Suppose $X_{T_y} = y$ and $X_{T_y+1} = z$. By time-homogeneity, the transition probability from y to z is equal to $P_{y,z}$, so the transition probabilities from time T_y to $T_y + 1$ is the same as if the original chain was starting from state y. Therefore, we should be able "forget" about the history and let the chain "restart" from state y. The following theorem, known as the **strong Markov property**, proves that our intuition is correct.

3.19. Theorem: *The process $\{X_{T_y+k}\}_{k=0,1,\ldots}$ behaves like a DTMC with initial state y.*

Proof. We wish to show that, if we are at state y at time $T_y = n$, then no matter what history $X_{n-1} = x_{n-1}, \ldots, X_0 = x_0$ we had, the transition probability going from $X_{T_y} = y$ to $X_{T_y+1} = z$ is equal to the transition probability P_{yz} given by $\Pr(X_{n+1} = z \mid X_n = y)$. In other words, we are proving that for all $n \in \mathbb{N}$ and for all states $\{x_{n-1}, \ldots, x_1, x_0\} \subseteq S$ with $x_{n-1}, \ldots, x_1 \neq y$ (as $T_y = n$ is the first time we (re)visit y), the following condition holds:

$$\Pr(X_{T_y+1} = z \mid X_{T_y} = y, T_y = n, X_{n-1} = x_{n-1}, \ldots, X_n = x_0) = \Pr(X_{n+1} = z \mid X_n = y).$$

But this is pretty straightforward:

$$\begin{aligned} &\Pr(X_{T_y+1} = z \mid X_{T_y} = y, T_y = n, X_{n-1} = x_{n-1}, \ldots, X_n = x_0) \\ &= \Pr(X_{n+1} = z \mid X_n = y, X_{n-1} = x_{n-1}, \ldots, X_n = x_0) && T_y = n \\ &= \Pr(X_{n+1} = z \mid X_n = y). && \text{Markov property} \end{aligned}$$

□

3.20. Motivation: Recall that a state $y \in S$ is recurrent if starting at y, we will eventually go back to y. Now by strong Markov property (as we assumed time-homogeneity), we can forget about the history and restart the chain at time T_y, so by the same argument we will visit y again, and again, and again. Thus, it makes sense for us to define the kth visit time to state y, T_y^k, and the total number of visits to y, $N(y)$. It's easy to see that if y is recurrent, then $N(y) = \infty$ as the chain will (re)visit it for an infinite number of times.

3.21. Definition: Let $T_y^1 = T_y$ and recursively define

$$T_y^k = \min\{n \geq T_y^{k-1} \mid X_n = y\}.$$

to be the time we arrive at state y (for the kth time) after the previous ($(k-1)$th) visit.

3.22. Theorem: *Starting at y, the probability that the chain revisits y for at least k times is equal to the kth power of the probability of revisiting y for the first time, i.e.,*

$$\Pr_y(T_y^k < \infty) = (\rho_{yy})^k.$$

Proof. This is a direct consequence of the strong Markov property. Recall that ρ_{yy} is the probability that we revisit y after starting at y. The strong Markov property allows us to split the chain into independent phases based on the (re)visits to y, so visiting the state y for at least k times is equal to the kth power of this probability. □

3.23. Note: Note this theorem gives us more ways of classifying states. Let $y \in S$.

- If y is transient, then $\rho_{yy} < 1$ and $\rho_{yy}^k \to 0$ as $k \to \infty$. Then:
  - the chain visits y for a finite number of times;
  - there exists a last visit to y.
- If y is recurrent, then $\rho_{yy} = 1$ and $\rho_{yy}^k = 1$ as $k \to \infty$. Then:
  - the chain visits y for an infinite number of times;
  - there does not exist a last visit to y.

These statements should be quite intuitive.

3.24. Definition: Let $N(y)$ be the total number of visits to state $y \in S$.

3.25. Note: By definition,

$$\begin{array}{rrcl} & \Pr_y(N(y) \geq k) & = & \Pr_y(T_y^k < \infty) = \rho_{yy}^k \\ \implies & \Pr_y(N(y) \geq k+1) & = & \rho_{yy}^{k+1} \\ \implies & \Pr_y(N(y) \leq k) & = & 1 - \rho_{yy}^{k+1}. \end{array}$$

Observe this is the cdf of Geometric$(1-\rho_{yy})$. Hence, the number of visits to state y conditioned on $X_0 = y$ follows a geometric distribution:

$$(N(y) \mid X_0 = y) \sim \text{Geometric}(1-\rho_{yy}).$$

Intuitively, we can interpret this as "keep trying until leaving y's communicating class", i.e., we define success as *not revisiting* y, or equivalently *leaving* y*'s class.*

3.26. (Cont'd): Since this is a geometric random variable, the expectation is given by

$$\mathbb{E}[N(y) \mid X_0 = y] =: \mathbb{E}_y[N(y)] = \frac{\rho_{yy}}{1-\rho_{yy}}.$$

Thus, we derive the same conclusion as in Note 3.23:

$$\begin{array}{rcccl} y \text{ is transient} & \iff & \rho_{yy} < 1 & \iff & \mathbb{E}_y[N(y)] < \infty. \\ y \text{ is recurrent} & \iff & \rho_{yy} = 1 & \iff & \mathbb{E}_y[N(y)] = \infty. \end{array}$$

More generally, we have the following result, which gives the total number of visits to state y given an arbitrary starting state $x \in S$.

3.27. Lemma: *For any $x, y \in S$,*

$$\mathbb{E}[N(y) \mid X_0 = x] =: \mathbb{E}_x[N(y)] = \frac{\rho_{xy}}{1-\rho_{yy}}.$$

Proof. We present a pure computational proof:

$$\begin{aligned}
\mathbb{E}_x[N(y)] &= \mathbb{E}_x(N(y) \mid T_y < \infty) \cdot \Pr_x(T_y < \infty) + \mathbb{E}_x(N(y) \mid T_y = \infty) \cdot \Pr_x(T_y = \infty) \\
&= \mathbb{E}_x(N(y) \mid T_y < \infty) \cdot \Pr_x(T_y < \infty) \\
&= (\mathbb{E}_y[N(y)] + 1) \cdot \rho_{xy} \\
&= \left(\frac{\rho_{xy}}{1-\rho_{yy}} + 1\right) \cdot \rho_{xy} \\
&= \rho_{xy} 1 - \rho_{yy}.
\end{aligned}$$

- Line 1 to 2: $\Pr_x(T_y = \infty)$ means we never visit y, so $\mathbb{E}_x[N(y) \mid T_y = \infty] = 0$.
- Line 2 to 3: $\Pr_x(T_y < \infty) = \rho_{xy}$ by definition. Next, $\mathbb{E}_x[N(y) \mid T_y < \infty]$ counts the expected number of visits to y after we've visited y for the first time. By the strong Markov property, this value is equal to $\mathbb{E}_x[N(y)] + 1$, where the plus 1 represents our first visit (at time T_y).

□

Section 6. Recurrence and Transience III

3.28. Motivation: In this section, we derive more criteria for recurrence and transience using the indicator function.

3.29. Definition: Let A be an event. Then $\mathbf{1}_A$ is a random variable defined as

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

This is known as the **indicator function** (for event A). Moreover, the expectation of an indicator random variable for event A is equal to the probability of event A:

$$\mathbb{E}[\mathbf{1}_A] = 1 \cdot \Pr(A) + 0 \cdot \Pr(A^C) = \Pr(A).$$

3.30. Lemma: $\mathbb{E}_x[N(y)] = \sum_{n=1}^{\infty} P^n_{xy}$.

Proof. By definition, we can write

$$N(y) = \sum_{i=1}^{\infty} \mathbf{1}_{\{X_n = y\}}.$$

Taking expectation, we have

$$\begin{aligned} \mathbb{E}_x[N(y)] &= \mathbb{E}\left[\sum_{i=1}^{\infty} \mathbf{1}_{\{X_n = y\}}\right] \\ &= \sum_{n=1}^{\infty} \mathbb{E}_x[\mathbf{1}_{\{X_n = y\}}] \\ &= \sum_{n=1}^{\infty} \Pr_x(X_n = y) \\ &= \sum_{n=1}^{\infty} P^n_{xy}. \end{aligned}$$

□

3.31. Note: Combining this with previous results, we see that

$$y \text{ is transient} \iff \mathbb{E}_y[N(y)] < \infty \iff \sum_{n=1}^{\infty} P^n_{yy} < \infty.$$
$$y \text{ is recurrent} \iff \mathbb{E}_y[N(y)] = \infty \iff \sum_{n=1}^{\infty} P^n_{yy} = \infty.$$

Section 7. Recurrence and Transience IV

3.32. Motivation: In this section, we show that recurrence and transience are class properties, and we provide some easy but useful results in classifying states.

3.33. Theorem: *Recurrence and transience are class properties.*

Proof. Assume x, y are in the same class, i.e., $x \leftrightarrow y$. Suppose x is recurrent. Since $x \to y$ and $y \to x$, there exist $m, n \in \mathbb{Z}_+$ such that $P_{xy}^{(m)} > 0$ and $P_{yx}^{(n)} > 0$. Note that

$$\begin{aligned} P_{yy}^{m+n+k} &= \Pr(X_{m+n+k} = y \mid X_0 = y) \\ &\geq \Pr(X_{m+n+k} = y, X_{n+k} = x, X_n = x \mid X_0 = y) \\ &\overset{\star}{=} P_{yx}^n P_{xx}^k P_{xy}^m \end{aligned}$$

Make sure you understand where $\star$ comes from! Basically, we have

- $X_0 = y \hookrightarrow X_n = x$: P_{yx}^n.
- $X_n = x \hookrightarrow X_{n+k} = x$: P_{xx}^k.
- $X_{n+k} = y \hookrightarrow X_{n+m+k} = y$: P_{xy}^m.

Next, observe that

$$\begin{aligned} \sum_{\ell=1}^{\infty} P_{yy}^{\ell} &\geq \sum_{\ell=m+n+1}^{\infty} P_{yy}^{\ell} \\ &= \sum_{k=1}^{\infty} P_{yy}^{m+n+k} && k \mapsto \ell - m - n \\ &\geq \sum_{k=1}^{\infty} P_{yx}^n P_{xx}^k P_{xy}^m \\ &= P_{yx}^n P_{xy}^m \sum_{k=1}^{\infty} P_{xx}^k \\ &= \infty \end{aligned}$$

Note the last line follows from the fact that $P_{yx}^n > 0, P_{xy}^m > 0$, and x is recurrent. Thus, y is recurrent. It follows that recurrence and transience are class properties. □

3.34. Motivation: Recall each transient state can be visited for a finite number of times. Thus a finite closed set cannot have all states being transient.

3.35. Theorem:

1. *A finite closed set has at least one recurrent state.*
2. *A finite closed class must be recurrent.*
3. *An irreducible DTMC with finite state space is recurrent.*

Proof. It suffices to prove statement 1 as the other two statements are direct consequences.

Let C be a finite closed set and suppose all states in C are transient. By our previous observations, this tells that for any states $x, y \in C$,

$$\mathbb{E}_x[N(y)] = \frac{\rho_{xy}}{1-\rho_{yy}} < \infty.$$

Since $|C| < \infty$, this implies that

$$\sum_{y \in C} \mathbb{E}_x[N(y)] < \infty.$$

However,

$$\begin{aligned}\sum_{y \in C} \mathbb{E}_x[N(y)] &= \mathbb{E}_x\left[\sum_{y \in C} N(y)\right] \\ &= \mathbb{E}_x\left[\sum_{y \in C}\sum_{n=1}^{\infty} \mathbf{1}_{\{X_n = y\}}\right] \\ &= \mathbb{E}_x\left[\sum_{n=1}^{\infty}\sum_{y \in C} \mathbf{1}_{\{X_n = y\}}\right].\end{aligned}$$

Since C is closed, starting from x, at any time n, X_n must be in one of the states in C. Hence, exactly one indicator takes value 1 and the rest are 0, i.e.,

$$\sum_{y \in C} \mathbf{1}_{\{X_n = y\}} = 1.$$

Summing up these 1's for an infinite number of times, we have

$$\mathbb{E}_x\left[\sum_{n=1}^{\infty}\sum_{y \in C} \mathbf{1}_{\{X_n = y\}}\right] = \mathbb{E}_x\left[\sum_{n=1}^{\infty} 1\right] = \infty.$$

Contradiction. Hence, we conclude that there must be at least one recurrent state. □

Chapter 4. Properties of DTMCs

Section 8. Existence of Stationary Distribution

4.1. In this section, we show that an irreducible and recurrent DTMC "almost" has a stationary distribution. If the state space is finite, then it has a stationary distribution. However, if the state space is infinite, then we need something stronger than recurrence (hint: positive recurrence) to guarantee the existence of a stationary distribution.

4.2. Definition: A row vector $\mu^* = [\mu^*(0), \mu^*(1), \ldots, \mu^*(i), \ldots]$ is called a **stationary measure** or **invariant measure**, if $\mu^* \geq \mathbf{0}$ and $\mu^* P = \mu^*$.

4.3. Remark: Comparing this with Definition 2.24, we see that a stationary measure is an "un-normalized" stationary distribution, i.e., its entries do not sum to 1. Now if the sum of entries is finite (which will always be the case when the state space is finite), then we can normalize it to get a stationary distribution:

$$\mu(i) = \frac{\mu^*(i)}{\sum_{j \in S} \mu^*(j)}, \quad \text{provided that } \sum_{j \in S} \mu^*(j) < \infty.$$

Otherwise, it's impossible to get a stationary distribution from the stationary measure.

4.4. The following result tells us that every irreducible and recurrent DTMC has a stationary measure. Don't get too scared by its construction; we care more about its *existence*.

4.5. Theorem: *Let $\{X_n\}_{n=0}^{\infty}$ be an irreducible and recurrent DTMC with transition matrix P. Let $x \in S$ and $T_x := \min\{n \geq 1 : X_n = x\}$ be the time of the first (re)visit to x. Then the row vector given by*

$$\mu_x(y) = \sum_{n=0}^{\infty} \Pr_x(X_n = y, T_x > n), \quad y \in S.$$

defines a stationary measure with $0 < \mu_x(y) < \infty$ for all $y \in S$.

4.6. Intuition: Let's digest the probability first. Let $x, y \in S$ and fix $n \in \mathbb{N}$. Then

$$\Pr_x(X_n = y, T_x > n)$$

represents the probability that we

- start at state x,
- visit y at time n, and that
- we have not revisited x yet prior to our visit to y at time n

Equivalently, it can be expressed as the expectation of the product of two indicator functions, one for each event:

$$\mathbb{E}_x[\mathbf{1}_{\{X_n=y\}} \cdot \mathbf{1}_{\{T_x>n\}}].$$

Thus, we can rewrite the expression in the theorem as

$$\begin{aligned}
\mu_x(y) &= \sum_{n=0}^{\infty} \Pr_x(X_n = y, T_x > n) \\
&= \sum_{n=0}^{\infty} \mathbb{E}_x[\mathbf{1}_{\{X_n=y\}} \cdot \mathbf{1}_{\{T_x>n\}}] \\
&= \mathbb{E}_x\left[\sum_{n=0}^{T_x-1} \mathbf{1}_{\{X_n=y\}} \cdot \mathbf{1}_{\{T_x>n\}} + \sum_{n=T_x}^{\infty} \mathbf{1}_{\{X_n=y\}} \cdot \mathbf{1}_{\{T_x>n\}}\right] \\
&\overset{1}{=} \mathbb{E}_x\left[\sum_{n=0}^{T_x-1} \mathbf{1}_{\{X_n=y\}}\right] \\
&\overset{2}{=} \mathbb{E}_x[\text{number of visits to } y \text{ before returning to } x]
\end{aligned}$$

- $\overset{1}{=}$: Once n exceeds T_x, the second indicator equals zero, so the product evaluates to zero. Now when $n = T_x$, $X_{T_x} \neq y$ as by definition we are at x at time T_x.
- $\overset{2}{=}$: We are counting the number of times we visit state y among $X_0, X_1, \dots, X_{T_x-1}$. $T_x - 1$ by definition is the time right before returning to y.

Intuitively speaking, we are cutting the MC into different "cycles" according to visits to state x. Thus, *$\mu_x(y)$ can be interpreted as the expected number of visits to y before returning to x*. In particular, *$\mu_x(x) = 1$ as you revisit x exactly once in each cycle*.

4.7. We now give a formal proof to the theorem. Note the proof might look pretty long but the previous paragraph has already done a lot of work explaining the theorem.

Proof. For $x, y \in S$, define

$$\bar{P}^n_{xy} := \Pr_x(X_n = y, T_x > n),$$

which represents the probability we start at x, visit y at time n, and that we have not revisited y prior to our visit to y at time n. Then we can write

$$\mu_x(y) = \sum_{n=0}^{\infty} \Pr_x(X_n = y, T_x > n) = \sum_{n=0}^{\infty} \bar{P}^n_{xy}.$$

We wish to show that

$$\forall z \in S : (\mu_x P)(z) = \mu_x(z),$$

which proves that μ_x is indeed a stationary measure.

Fix $z \in S$. We need to consider two cases: $z \neq x$ and $z = x$.

Case 1. If $z \neq x$, then

$$\begin{aligned}
(\mu_x P)(z) &= \sum_y \mu_x(y) P_{yz} \\
&= \sum_y \sum_{n=0}^{\infty} \bar{P}^n_{xy} P_{yz} \\
&= \sum_{n=0}^{\infty} \sum_y \bar{P}^n_{xy} P_{yz} \\
&\overset{\star}{=} \sum_{n=0}^{\infty} \bar{P}^{n+1}_{xz} \\
&= \sum_{n=0}^{\infty} \bar{P}^n_{xz} && \bar{P}^0_{xz} = \Pr_x(X_0 = z, T_x > 0) = 0 \\
&= \mu_x(z),
\end{aligned}$$

where $\star$ holds because

$$\begin{aligned}
\sum_y \bar{P}^n_{xy} P_{yz} &= \sum_y \Pr_x(X_n = y, T_x > n, X_{n+1} = z) \\
&= \Pr_x(T_x > n, X_{n+1} = z) && \text{summing out } y \\
&= \Pr_x(T_x > n+1, X_{n+1} = z) && X_{n+1} \neq x \implies T_x > n+1 \\
&= \bar{P}^{n+1}_{xz}.
\end{aligned}$$

Case 2. If $z = x$, then

$$\begin{aligned}
\sum_y \bar{P}^n_{xy} P_{yx} &= \sum_y \Pr_x(X_n = y, T_x > n, X_{n+1} = x) \\
&= \Pr_x(T_x = n+1).
\end{aligned}$$

Then we have

$$\begin{aligned}
(\mu_x P)(x) &= \sum_{n=0}^{\infty} \sum_y \bar{P}^n_{xy} P_{yx} \\
&= \sum_{n=0}^{\infty} \Pr_x(T_x = n+1) \\
&\overset{\star}{=} 1 \\
&= \mu_x(x).
\end{aligned}$$

Note that the LHS of $\star$ represents the probability that we eventually return to x, which is 1 as x is recurrent. It follows that $\mu_x P = \mu_x$. as desired.

It remains to show that $\mathbf{0} < \mu_x < \infty$. First, for any n, we have

$$\begin{aligned} 1 &= \mu_x(x) && x \text{ is recurrent} \\ &= (\mu_x P^n)(x) && \mu_x = \mu_x P^n \\ &= \sum_z \mu_x(z) P^n_{zx} && \text{CK, taking } z \text{ as the intermediate state} \\ &\geq \mu_x(y) P^n_{yx} && \text{taking a specific } y \text{ as the intermediate state} \end{aligned}$$

Since the chain is irreducible, there exists some n^* such that $P^{n^*}_{yx} > 0$. Then we have

$$\mu_x(y) P^{n^*}_{yx} \leq 1 \implies \mu_x < \infty.$$

Secondly, recall that we have proved earlier that if $x \to y$, then there is a way to visit y before returning to x, i.e.,

$$\Pr_x(\text{the number of visits to } y \text{ before returning to } x \geq 1) > 0.$$

Then the expectation of this quantity will also be strictly positive. But by the previous Intuition block,

$$\mu_x(y) = \mathbb{E}_x[\text{number of visits to } y \text{ before returning to } x]$$

is exactly this expectation! This concludes the proof. □

Section 9. Periodicity

4.8. Definition: The **period** of a state x is defined as

$$d(x) = \gcd\{n \geq 1 : P^n_{xx} > 0\}.$$

4.9. Remark: Periodicity in a deterministic setting means after this number of steps, you are guaranteed to be back to this state. This is not the case in a stochastic setting. In the above definition we are taking the gcd of time stamps by which the process *can* return to x (i.e., with positive probabilities), not *must* return to x (i.e., with probability of 1). In fact, there is no guarantee that the chain will be in state x at time $d(x)$. Indeed, $\gcd(A)$ is not necessarily in A, so $d(x)$ is not necessarily a time stamp satisfying $P^{d(x)}_{xx} > 0$.

4.10. (Cont'd): Let $d = d(x)$. From above, we see that the statement

$$\textit{if } n \textit{ is a multiple of } d\textit{, then } P^n_{xx} > 0$$

is FALSE. The correct way to interpret this is that

$$d \nmid n \implies P^n_{xx} = 0,$$

i.e., *if n is not a multiple of d, then it is impossible to go back to x in n steps after starting at x*. Note this is the contrapositive of the reverse of the wrong interpretation.

4.11. Definition: If $d(x) = 1$, the state x is said to be **aperiodic**. A Markov chain is said to be **aperiodic** if all of its states are aperiodic.

4.12. Note: If $P_{xx} > 0$, then x is automatically aperiodic. The converse does not hold, i.e., knowing that x is aperiodic does not give you $P_{xx} > 0$. Indeed, if $\{n \geq 1 : P^n_{xx} > 0\}$ contains co-primes, the x could be aperiodic without having $P_{xx} > 0$.

4.13. Example: Consider the symmetric simple random walk example. We have

$$P^n_{00} = \begin{cases} 0 & n \text{ is odd} \\ \binom{n}{n/2}\left(\frac{1}{2}\right)^{n/2}\left(\frac{1}{2}\right)^{n/2} = \binom{n}{n/2}\left(\frac{1}{2}\right)^{n} & n \text{ is even} \end{cases}$$

Indeed, walking an odd number of steps cannot get you back to the origin. Now let n be even. To get back to the origin after n steps, you must have walked $n/2$ steps toward the left and $n/2$ toward the right. Thus, you are looking at a binomial r.v. with $k = n/2$ and $p = 1/2$. Taking the gcd of the positive even integers (as they give $P^n_{00} > 0$), we get

$$d(0) = \gcd\{2\mathbb{Z}_{\geq 0}\} = 2.$$

Since this argument holds for any starting state i, we have $d(i) = 2$ for all $i \in \mathbb{Z}$. Also, the period does not depend on p, so *every state $i \in \mathbb{Z}$ has a period of 2 in any SLR*.

4.14. Lemma: *States in the same class have the same period, i.e.,*

$$x \leftrightarrow y \implies d(x) = d(y).$$

Proof. Since $x \leftrightarrow y$, there exist $m, n \in \mathbb{Z}_+$ such that $P_{xy}^m > 0$ and $P_{yx}^n > 0$. Then by CK

$$P_{xx}^{m+n} \geq P_{xy}^n P_{yx}^n > 0.$$

Moreover, for any ℓ such that $P_{yy}^\ell > 0$, we have

$$P_{xx}^{m+n+\ell} \geq P_{xy}^m P_{yy}^\ell P_{yx}^n > 0.$$

As a result, $d(x) \mid (m+n)$ and $d(x) \mid (m+n+\ell)$. Thus, $d(x) \mid \ell$. Since this holds for all ℓ such that $P_{yy}^\ell > 0$, $d(x)$ is a common divisor of all ℓ's. Recall that $d(y)$ is the greatest common divisor of all ℓ's. Thus, $d(x) \leq d(y)$ (in fact, $d(x) \mid d(y)$). A symmetric derivation gives $d(y) \leq d(x)$. It follows that $d(x) = d(y)$ as desired. □

Section 10. Detailed Balance Condition

4.15. Definition: A distribution $\pi = \{\pi(x)\}_{x \in S}$ is said to satisfy the **detailed balance condition** if $\pi(x) \cdot P_{xy} = \pi(y) \cdot P_{yx}$ for all $x, y \in S$.

4.16. Proposition: *If π satisfies the detailed balance condition, then it is a stationary distribution.*

Proof. Suppose $\pi_i P_{ij} = \pi_j P_{ji}$ for all $i, j \in S$. Let P_j denote the j-th column of the transition matrix P. Then we have $\pi P = (\pi \cdot P_1, \pi \cdot P_2, \dots, \pi \cdot P_n)$. In particular, the j-th entry of πP is given by

$$\begin{aligned} [\pi P]_j = \pi \cdot P_j &= \sum_i \pi_i P_{ij} \\ &= \sum_i \pi_j P_{ji} && \text{detailed balance condition} \\ &= \pi_j \sum_i P_{ji} = \pi_j \end{aligned}$$

Note the last equality holds as $\sum_i P_{ji}$ is a row-sum and thus equals 1. Therefore, $\pi P = \pi$ and by definition it is a stationary distribution. □

4.17. Remark: The converse is not true; the detailed balance condition is stronger than stationarity. Let us compare these two conditions with the following example.

4.18. (Cont'd): Consider a DTMC with $S = \{x, y, z\}$. Suppose π is a stationary distribution. Then the total "flow" going into x should be equal to the total "flow" leaving x. Indeed, since π is stationary, the probability (at all time) that the chain is at state $w \in S$ is always $\pi(w)$. Then the probability going from w to x at any transition step is $\pi(w)P_{wx}$. Summing across all $w \in S$, the total probability flow into x is given by

$$\sum_w \pi(w)P_{wx} = (\pi P)_x = \pi(x) = \sum_w \pi(x)P_{xw}.$$

But the very last expression in this equality is the total probability leaving x.

4.19. (Cont'd): Now suppose π is a detailed balance condition, i.e., for all $x, y \in S$,

$$\pi(x)P_{xy} = \pi(y)P_{yx}.$$

Using the same logic as above, $\pi(x)$ is the probability that we are at state x, and P_{xy} is the probability that we are going from x to y. Thus, LHS is the probability flow from x to y. Similarly, the RHS is the probability flow from y to x. The detailed balance condition tells us that these two arrows have the same edge weight! In other words, the flow between each pair of states are balanced.

Section 11. Time Reversibility

4.20. Definition: Let $\{X_n\}_{n=0}^\infty$ be a DTMC. Fix n, the process $\{Y_m\}_{m=0}^\infty$ where

$$Y_m := X_{n-m}$$

is called the **reversed process** of $\{X_n\}_{n=0}^\infty$.

4.21. The reversed process of a DTMC is not necessarily a DTMC, but it is in the following case, where the DTMC starts at a stationary distribution that is strictly positive.

4.22. Proposition: *If $\{X_n\}_{n=0}^\infty$ starts from a stationary distribution π with $\pi(i) > 0$ for any $i \in S$, then its reversed process $\{Y_m\}_{m=0}^\infty$ is a DTMC with transition probabilities*

$$\hat{P}_{ij} = \Pr(Y_{m+1} = j \mid Y_m = i) = \frac{\pi(j) P_{ji}}{\pi(i)}.$$

Proof. We derive the formula for $\hat{P}_{ij}$ then show that the Markov property is satisfied.

$$\begin{aligned}
&\Pr\left(Y_{m+1} = i_{m+1} \mid Y_m = i_m, \ldots, Y_0 = i_0\right) \\
&= \frac{\Pr\left(Y_{m+1} = i_{m+1}, Y_m = i_m, \ldots, Y_0 = i_0\right)}{\Pr\left(Y_m = i_m, \ldots, Y_0 = i_0\right)} && \text{definition of conditional probability} \\
&= \frac{\Pr\left(X_{n-(m+1)} = i_{m+1}, X_{n-m} = i_m, \ldots, X_n = i_0\right)}{\Pr\left(X_{n-m} = i_m, \ldots, X_n = i_0\right)} && \text{definition of } Y \\
&= \frac{\Pr\left(X_{n-(m+1)} = i_{m+1}\right) P_{i_{m+1},i_m} P_{i_m,i_{m-1}} \ldots P_{i_1,i_0}}{\Pr\left(X_{n-m} = i_m\right) P_{i_m,i_{m-1}} \ldots P_{i_1,i_0}} && \text{definition of transition probabilities} \\
&= \frac{\Pr\left(X_{n-(m+1)} = i_{m+1}\right) P_{i_{m+1},i_m}}{\Pr\left(X_{n-m} = i_m\right) P_{i_m,i_{m-1}}} && \text{cancel the same terms}
\end{aligned}$$

Since $\{X_n\}$ starts from a stationary distribution π, we have

$$\Pr(X_{n-(m+1)} = i_{m+1}) = \pi(i_{m+1}), \qquad \Pr(X_{n-m} = i_m) = \pi(i_m),$$

so the above expression can be rewritten as

$$\Pr\left(Y_{m+1} = i_{m+1} \mid Y_m = i_m, \ldots, Y_0 = i_0\right) = \frac{\pi\left(i_{m+1}\right) P_{i_{m+1},i_m}}{\pi\left(i_m\right)}.$$

Observe this transition probability does not depend on the history $i_{m-1}, \ldots, i_0$, so Markov property holds and $\{Y_m\}_{m=0}^\infty$ is indeed a DTMC. Moreover, replacing i_{m+1} by j and i_m by i, we see that the transition probabilities of the reversed DTMC is given by

$$\hat{P}_{ij} = \Pr(Y_{m+1} = j \mid Y_m = i) = \frac{\pi(j) P_{ji}}{\pi(i)}.$$

□

4.23. Remark: Let us verify that $\hat{P} = \{\hat{P}_{ij}\}_{i,j\in S}$ is indeed a valid transition matrix:

$$\hat{P}_{ij} = \frac{\pi(j)P_{ji}}{\pi(i)} \geq 0$$
$$\sum_{j\in S} \hat{P}_{ij} = \frac{\sum_{j\in S}\pi(j)P_{ji}}{\pi(i)} = \frac{(\pi P)_i}{\pi(i)} = \frac{\pi(i)}{\pi(i)} = 1$$

4.24. Definition: A DTMC $\{X_n\}_{n=0,1,...}$ is called **time-reversible** if its reversed chain $\{Y_m := X_{n-m}\}_{m=0}^n$ has the same distribution as $\{X_m\}_{m=0}^n$ for all n.

4.25. Remark: Note the "same distribution" here does not just refer to the marginal distribution, but also the joint distributions. Now recall that the distribution of a DTMC is complete determined by its initial distribution as well as the transition matrix.

4.26. Remark: If a DTMC is time-reversible, then its reversed process is clearly a DTMC. The converse is not true, i.e., the reversed process of a DTMC is a DTMC does not guarantee it to be time-reversible, i.e., the reversed DTMC is not guaranteed to have the same distribution as the original DTMC. Intuitively, this difference is related to the difference between the detailed balanced condition and the stationarity condition. Compare this with Remark 4.17. Below we show that time-reversibility is equivalent to the detailed balance condition.

4.27. Proposition: *A DTMC $\{X_n\}_{n=0,1,...}$ is time-reversible iff it satisfies the detailed balance condition.*

Proof. ($\Leftarrow$) Assume the DTMC satisfies the detailed balance condition. Then $\{X_n\}$ starts from the stationary distribution π and $\pi(i)P_{ij} = \pi(j)P_{ji}$. Thus, $\{Y_m\}_{m=0}^\infty$ is a DTMC and $Y_0 = X_n \sim \pi$. The transition probability of the reversed DTMC $\{Y_m\}_{m=0}^\infty$ is given by

$$\hat{P}_{ij} = \frac{\pi(j)P_{ji}}{\pi(i)} = \frac{\pi(i)P_{ij}}{\pi(i)} = P_{ij}.$$

Thus, $\{X_n\}$ and $\{Y_m\}$ are two DTMCs with same initial distribution and transition matrix, hence have the same distribution.

($\Rightarrow$) Assume the DTMC is time-reversible. By definition, X_0 and $X_n = Y_0$ have the same distribution. This holds for all n, so X_0 follows a stationary distribution π. Moreover, by time-reversibility,

$$P_{ij} = \hat{P}_{ij} = \frac{\pi(j)P_{ji}}{\pi(i)} \implies \pi(i)P_{ij} = \pi(j) = P_{ji}$$

for all $i, j \in S$, which is exactly the detailed balance condition. □

Chapter 5. Limiting Behaviours of DTMCs

Section 1. Main Theorems: Preparation

5.1. Note: Let us define the following abbreviations.

I : The MC is ***I****rreducible.* A : The MC is ***A****periodic.*
R : The MC is ***R****ecurrent.* S : The MC has a ***S****tationary* distribution π.

5.2. Lemma: *If π is a stationary distribution with $\pi(y) > 0$, then y is recurrent.*

Proof. Assume the DTMC $\{X_n\}_{n=0}^{\infty}$ starts from the stationary distribution π. Then for each $n \in \mathbb{Z}_{\geq 0}$, $\Pr(X_n = y) = \pi(y)$. Adding them up, we have

$$\begin{aligned}
\infty &= \sum_{n=1}^{\infty} \Pr(X_n = y) && \text{summing up positive numbers} \\
&= \sum_{n=1}^{\infty} \mathbb{E}[\mathbf{1}_{\{X_n = y\}}] && \text{probability} = \text{expectation of indicator} \\
&= \mathbb{E}\left[\sum_{n=1}^{\infty} \mathbf{1}_{\{X_n = y\}}\right] \\
&= \mathbb{E}[N(y)] \\
&= \sum_{x \in S} \mathbb{E}_x[N(y)]\pi(x) && \mathbb{E}_x[A] = \mathbb{E}[A \mid X_0 = x] \\
&= \sum_{x \in S} \pi(x) \frac{\rho_{xy}}{1 - \rho_{yy}} && \text{definition of } \mathbb{E}_x[N(y)] \\
&\leq \sum_{x \in S} \pi(x) \frac{1}{1 - \rho_{yy}} && \rho_{xy} \leq 1 \\
&= \frac{1}{1 - \rho_{yy}} \implies \rho_{yy} = 1. && \sum_{x \in S} \pi(x) = 1
\end{aligned}$$

□

5.3. Corollary: *If y is transient, then $\pi(y) = 0$ for any stationary distribution π.*

Proof. Contrapositive of the Lemma above. □

5.4. Corollary: *An irreducible MC with a stationary distribution π is recurrent.*

Proof. Since π exists, there exists some $y \in S$ such that $\pi(y) > 0$. By the Lemma above, y is recurrent. Since the MC is irreducible, all states are recurrent. □

Section 2. Main Theorems: Convergence Theorem

5.5. Motivation: In this section, we show that for an irreducible, aperiodic DTMC with a stationary distribution π, the transition probabilities converges to π as $n \to \infty$. In particular, the starting state x is irrelevant, i.e., the limiting transition probability, hence also the limiting distribution, does not depend on the starting state:

$$\lim_{n\to\infty} P^n_{xy} = \pi(y) \implies \lim_{n\to\infty} \Pr(X_n = y) = \pi(y).$$

As a corollary, this implies that the stationary distribution of a MC, if exists, is unique.

5.6. Lemma: *If y is aperiodic, then there is $n_0 \in \mathbb{N}$ such that $P^n_{yy} > 0$ for all $n \geq n_0$.*

Proof. We use a corollary from Bezout's Lemma: Given a set of co-primes I, there exist a subset $\{i_1, \dots, i_m\} \subseteq I$ and $n_0 \in \mathbb{N}$ such that for any $n \geq n_0$, n can be written as $n = a_1 i_i + \cdots + a_m i_m$ for some $a_1, \dots, a_m \in \mathbb{Z}_+$. Take $I = \{n \geq 1 : P^n_{yy} > 0\}$. By aperiodicity of y, the elements in I are co-primes, so we can find $n_0 \in \mathbb{Z}_+$ such that for any $n \geq 0$, there exist $a_1, \dots, a_m \in \mathbb{Z}_+$ and $i_1, \dots, i_m \in I$ such that $n = a_1 i_1 + \cdots + a_m i_m$. Therefore,

$$P^n_{yy} \geq \overbrace{P^{i_1}_{yy}P^{i_1}_{yy}\cdots P^{i_1}_{yy}}^{a_i\text{terms}} \cdot \overbrace{P^{i_2}_{yy}P^{i_2}_{yy}\cdots P^{i_2}_{yy}}^{a_2\text{terms}} \cdots \overbrace{P^{i_m}_{yy}P^{i_m}_{yy}\cdots P^{i_m}_{yy}}^{a_m\text{terms}} > 0$$

as desired. □

5.7. Theorem (Convergence)**:** *If a MC is irreducible, aperiodic, and has a stationary distribution π, then the transition probabilities converges to π as $n \to \infty$, i.e.,*

$$I \wedge A \wedge S \implies \forall x, y \in S : P^n_{xy} \overset{n\to\infty}{\to} \pi(y).$$

In particular, the starting state x is irrelevant.

Proof. Consider two independent DTMCs $\{X_n\}_{n=0,1,\dots}$ and $\{Y_n\}_{n=0,1,\dots}$ with the same transition matrix P but arbitrary initial distributions. Define $Z_n = (X_n, Y_n)$ for $n = 0, 1, \dots$ It's easy to see that $\{Z_n\}_{n=0,1,\dots}$ is also a DTMC (check Markov property) with transition matrix

$$\bar{P}_{(x_1,y_1)(x_2,y_2)} = P_{x_1x_2} \cdot P_{y_1y_2}.$$

Claim 1. The MC $\{Z_n\}_{n=0,1,\dots}$ is irreducible.

Proof 1. Since $\{X_n\}_{n=0,1,\dots}$ and $\{Y_n\}_{n=0,1,\dots}$ are irreducible, for any $x_1, x_2, y_1, y_2 \in S$, there exists $k, \ell \in \mathbb{N}$ such that $P^k_{x_1x_2} > 0$ and $P^\ell_{y_1y_2} > 0$. Since the DTMC is aperiodic, by the Lemma above, $P^m_{x_1x_2} > 0$ and $P^m_{y_1y_2} > 0$ for all m larger than some threshold $M \in \mathbb{N}$. Then for $n \geq M + \max\{k, \ell\}$, $P^n_{x_1x_2} \geq P^k_{x_1x_2}P^{n-k}_{x_2x_2} > 0$ and $P^n_{y_1y_2} \geq P^\ell_{y_1y_2}P^{n-\ell}_{y_2y_2} > 0$. Thus,

$$\bar{P}^n_{(x_1,y_1)(x_2,y_2)} = P^n_{x_1x_2}P^n_{y_1y_2} > 0.$$

Since this holds for all $(x_1, y_1), (x_2, y_2) \in S \times S$, $\{Z_n\}_{n=0,1,\dots}$ is irreducible. ■

Claim 2. $\{Z_n\}$ is recurrent.

Proof 2. Note that $\bar{\pi}(x,y) = \pi(x)\pi(y)$ is a stationary distribution of $\{Z_n\}_{n=0,1,\dots}$. Take x such that $\pi(x) > 0$, then $\bar{\pi}(x,x) = \pi(x)^2 > 0$. By Lemma 5.2, (x,x) is recurrent. Since $\{Z_n\}_{n=0,1,\dots}$ is irreducible, all the states in $\{Z_n\}_{n=0,1,\dots}$ are recurrent. ■

Now define $T = \min\{n \geq 0 : X_n = T_n\}$, the first time that the two chains meet. Also define $V_{(x,x)} = \min\{n \geq 0 : X_n = Y_n = x\}$, the first time that the two chains meet at x. By definition, $T \leq V_{(x,x)}$. Thus,

$$\begin{aligned} V_{(x,x)} &= \min\{x \geq 0 : X_n = Y_n = x\} \\ &= \min\{n \geq 0 : Z_n = (x,x)\} \\ &\leq \min\{n \geq 1 : Z_n = (x,x)\} \\ &= T_{(x,x)} \end{aligned}$$

where $T_{(x,x)}$ denotes the time of the first (re)visit to state (x,x).

Claim 3. $T_{(x,x)} < \infty$ and the two Markov chains will eventually meet.

Proof 3. By definition, $\Pr(T_{(x,x)} < \infty) = \mathbb{E}[\Pr(T_{(x,x)} < \infty \mid (X_0, Y_0))]$. For any (x_0, y_0),

$$\Pr(T_{(x,x)} < \infty \mid X_0 = x_0, Y_0 = y_0) = \rho_{(x_0,y_0)}(x,x)$$

Since $\{Z_n\}_{n=0,1,\dots}$ is irreducible, $(x,x) \to (x_0,y_0)$, so $\rho_{(x,x)(x_0 y_0)} > 0$. Since (x,x) is recurrent, $\rho_{(x,x)(x_0,y_0)} = 1$. Taking expectation over all (x_0,y_0), we have $\Pr(T_{(x,x)} < \infty) = 1$. Therefore,

$$T \leq V_{(x,x)} \leq T_{(x,x)} < \infty.$$

Hence, the two independent Markov chains $\{X_n\}_{n=0}^{\infty}$ and $\{Y_n\}_{n=0}^{\infty}$ will eventually meet. ■

We are ready to finish the proof. For any state $y \in S$ and any $n \in \mathbb{N}$,

$$\begin{aligned} \Pr(X_n = y, T \leq n) &= \sum_{m=0}^{n} \sum_{x \in S} \Pr(T = m, X_m = x, X_n = y) \\ &= \sum_{m=0}^{n} \sum_{x \in S} \Pr(T = m, X_m = x) \Pr(X_n = y \mid X_m = x, T = m) \\ &= \sum_{m=0}^{n} \sum_{x \in S} \Pr(T = m, X_m = x) \Pr(X_n = y \mid X_m = x) && \text{Markov property} \\ &= \sum_{m=0}^{n} \sum_{x \in S} \Pr(T = m, X_m = s) \cdot P_{xy}^{n-m} \\ &= \sum_{m=0}^{n} \sum_{x \in S} \Pr(T = m, Y_m = x) \Pr(Y_n = y \mid Y_m = x) \end{aligned}$$

$$\begin{aligned}
&= \sum_{m=0}^{n} \sum_{x \in S} \Pr(T = m, Y_m = x) \Pr(Y_n = y \mid Y_m = x, T = m) \\
&= \sum_{m=0}^{n} \sum_{x \in S} \Pr(T = m, Y_m = x, Y_n = y) \\
&= \Pr(Y_n = y, T \leq n)
\end{aligned}$$

Intuitively, this says that $\{X_n\}_{n=0,1,\dots}$ and $\{Y_n\}_{n=0,1,\dots}$ have the same distribution once they meet. Then

$$\begin{aligned}
&|\Pr(X_n = y) - \Pr(Y_n = y)| \\
&= |\cancel{\Pr(X_n = y, T \leq n)} + \Pr(X_n = y, T > n) - \cancel{\Pr(Y_n = y, T \leq n)} - \Pr(Y_n = y, T > n)| \\
&= |\Pr(X_n = y, T > n) - \Pr(Y_n = y, T > n)| \\
&\leq \Pr(X_n = y, T > n) + \Pr(Y_n = y, T > n) \\
&\leq 2 \Pr(T > n) \to 0 \text{ as } n \to \infty.
\end{aligned}$$

Recall that this holds for any initial distribution of $\{X_n\}_{n=0}^{\infty}$ and $\{Y_n\}_{n=0}^{\infty}$. Take $X_0 = x$ and $Y_0 \sim \pi$. Then by this argument,

$$|P_{xy}^n - \pi(y)| = |\Pr(X_n = y) - \pi(y)| = |\Pr(X_n = y) - \Pr(Y_n = y)| \overset{n\to\infty}{\to} 0.$$

Equivalently, we get

$$P_{xy}^n \to \pi(y) \overset{n\to\infty}{\to} n \to \infty.$$

□

Section 3. Main Theorems: Asymptotic Frequency

5.8. Motivation: In this section, we show that the long-fun fraction of time a DTMC spends in state y is given by

$$\frac{N_n(y)}{n} \to \frac{1}{\mathbb{E}_y[T_y]},$$

where we can interpret $\mathbb{E}_y[T_y]$ as the expected cycle length.

5.9. Theorem (Asymptotic Frequency)**:** *Consider an irreducible and recurrent DTMC. If $N_n(y)$ is the number of visits to y up to time n, then*

$$\frac{N_n(y)}{n} \to \frac{1}{\mathbb{E}_y[T_y]}.$$

where $T_y = \min\{n \geq 1 : X_n = y\}$.

Proof. We chop the timeline into different cycles. Let $T_y^{(k)}, k \geq 1$ be the time that the chain (re)visits y for the k-th time after time 0. By the strong Markov property, the cycle lengths

$$T_y^{(k+1)} - T_y^{(k)}$$

are iid variables for all $k \geq 1$. The strong law of large number states that given iid $X_1, \dots, X_n$,

$$\frac{\sum_{i=1}^n X_i}{n} \to \mathbb{E}[X_1].$$

Thus,

$$\frac{T_y^{(k)} - T_y^{(1)}}{k-1} = \frac{\sum_{i=1}^{k-1}(T_y^{(i+1)} - T_y^{(i)})}{k-1} \to \mathbb{E}(T_y^{(i+1)} - T_y^{(i)}) = \mathbb{E}_y[T_y].$$

With negligible changes, this implies that

$$\frac{T_y^{(k)}}{k} \xrightarrow{k\to\infty} \mathbb{E}_y[T_y].$$

Recall $N_n(y)$ is the number of visits to y up to time n. Thus,

$$\begin{aligned}
T_y^{(N_n(y))} \leq n < T_y^{(N_n(y)+1)} &\implies \frac{T_y^{(N_n(y))}}{N_n(y)} \leq \frac{n}{N_n(y)} < \frac{T_n^{(N_n(y)+1)}}{N_n(y)+1}\frac{N_n(y)+1}{N_n(y)} \\
&\implies \mathbb{E}_y[T_y] \leq \frac{n}{N_n(y)} < \mathbb{E}_y[T_y] \cdot 1 \qquad n \to \infty \\
&\implies \frac{n}{N_n(y)} \xrightarrow{n\to\infty} E_y(T_y) \\
&\implies \frac{N_n(y)}{n} \xrightarrow{n\to\infty} \frac{1}{\mathbb{E}_y[T_y]}.
\end{aligned}$$

□

Section 4. Main Theorems: Stationarity and Nicest Case

5.10. Theorem: *For an irreducible DTMC with a stationary distribution π, we have*

$$\pi(y) = \frac{1}{\mathbb{E}_y[T_y]}.$$

In particular, the stationary distribution is unique.

Proof. First, irreducibility and stationarity implies recurrence, so we can apply the previous theorem to obtain

$$\frac{N_n(y)}{n} \to \frac{1}{\mathbb{E}_y[T_y]}.$$

Taking expectation of both sides (using dominated convergence theorem):

$$\mathbb{E}\left[\frac{N_n(y)}{n}\right] \to \frac{1}{\mathbb{E}_y[T_y]}$$

For any initial distribution, we have

$$\begin{aligned}\mathbb{E}[N_n(Y)] &= \mathbb{E}\left[\sum_{m=1}^{n} \mathbf{1}_{\{X_m=y\}}\right] \\ &= \sum_{m=1}^{n} \mathbb{E}[\mathbf{1}_{\{X_m=y\}}] \\ &= \sum_{m=1}^{n} \Pr(X_m = y).\end{aligned}$$

Now assume the MC starts from the stationary distribution π. Then

$$\begin{aligned}\Pr(X_m = y) = \Pr(X_0 = y) = \pi(y) &\implies \mathbb{E}[N_n(y)] = \sum_{m=1}^{n} \Pr(X_m = y) = n\pi(y) \\ &\implies \mathbb{E}\left[\frac{N_n(y)}{n}\right] = \pi(y) = \frac{1}{\mathbb{E}_y[T_y]}.\end{aligned}$$

□

5.11. Corollary (Nicest Case)**:** *Suppose I, A, S (which together implies R). Then*

$$\pi(y) = \lim_{n\to\infty} P_{xy}^n = \lim_{n\to\infty} \frac{N_n(y)}{n} = \frac{1}{\mathbb{E}_y[T_y]}.$$

In words, everything exists and equal to each other:

$$\begin{aligned}\textit{stationary distribution} &= \textit{limiting transition probability} \\ &= \textit{long-run fraction of time} \\ &= \textit{1/expected cycle length.}\end{aligned}$$

Section 5. Main Theorems: Long-Run Average

5.12. Intuition: Let $f(X_n)$ be the reward/penalty function at X_n. Then

$$\frac{1}{n}\sum_{m=1}^{n} f(X_n)$$

is the average reward per step from 1 to n. Taking the limit as $n \to \infty$, we obtain the long run average reward per step. We have proven that the stationary distribution gives you the long-run fraction of time, so the long-run average should just be the reward weighted by π.

5.13. Lemma: *For an irreducible DTMC with stationary distribution π, the stationary distribution is a strictly positive vector:*

$$I \wedge S \implies \forall x \in S : \pi(x) > 0.$$

Proof. Since π is a stationary distribution,

$$\sum_z \pi(z) = 1 \implies \exists y \in S : \pi(y) > 0.$$

Since the MC is irreducible, for any state $x \in S$, $y \to x$, so

$$\exists n \in \mathbb{Z}_+ : P^n_{yx} > 0.$$

Since π is stationary,

$$\pi = \pi \cdot P^n \implies \pi(x) = \sum_{z \in S} \pi(z) P^n_{zx} \geq \pi(y) P^n_{yx} > 0$$

as both arguments are positive. □

5.14. Theorem (Long-Run Average)**:** *Let $\{X_i\}_{i=0}^{\infty}$ be an irreducible DTMC with stationary distribution π. Let $f : S \to \mathbb{R}$ be a function satisfying*

$$\sum_x |f(x)|\pi(x) < \infty.$$

Then the long-run average of the DTMC is given by

$$\lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} f(X_n) = \sum_x f(x)\pi(x) = \pi \cdot f^T \in \mathbb{R}.$$

Proof. Recall that irreducibility and stationarity implies recurrence. Now irreducibility and recurrence implies the existence of a stationary measure. Thus,

$$\begin{aligned}\mu_x(y) &= \sum_{n=0}^{\infty} \Pr_x(X_n = y, T_x > n), \quad y \in S \\ &= \mathbb{E}_x[\text{number of visists to } y \text{ before returning to } x] = \mathbb{E}_x[N_{T_x}(y)].\end{aligned}$$

Moreover, note that

$$\sum_y \mathbb{E}_x[N_{T_x}(y)] = \mathbb{E}_x[T_x] \overset{1}{=} \frac{1}{\pi(x)} < \infty,$$

where $\overset{1}{=}$ follows from Theorem 5.10. Now since $\{\pi_x(y)\} = \{\mathbb{E}_x[N_{T_x}(y)]\}_{y\in S}$ is normalizable,

$$\left\{\frac{\mathbb{E}_x[N_{T_x}(y)]}{\mathbb{E}_x[T_x]}\right\}_{y\in S}$$

is a stationary distribution. It is also unique due to the irreducibility and stationarity of the DTMC (Theorem 5.10). Hence, we must have

$$\frac{\mathbb{E}_x[N_{T_x}(y)]}{\mathbb{E}_x[T_x]} = \pi(y).$$

Multiplying both sides by $\mathbb{E}_x[T_x]$, we have

$$\mathbb{E}_x[T_x] = \frac{1}{\pi(x)} \implies \mathbb{E}_x[N_{T_x}(y)] = \frac{\pi(y)}{\pi(x)}.$$

Note that *this gives you a way to find the expected number of visits to state y in one cycle (starting and ending at x)*. Again, we chop the DTMC into different cycles. The reward collected in the kth cycle (defined by revisits to x) and its expectation are given by

$$Y_k := \sum_{m=T_{k-1}+1}^{T_k} f(X_m)$$

$$\mathbb{E}[Y_k] = \sum_{y\in S} \mathbb{E}_x[N_{T_x}(y)] \cdot f(y) = \frac{\sum_{y\in S}\pi(y)f(y)}{\pi(x)} = \frac{\pi f^T}{\pi(x)}$$

Note the expectation is calculated by multiplying the expected number of visits to each state by the reward to be collected at that state. We implicitly use the strong Markov property. Now the average reward over time is

$$\frac{\sum_{k=2}^{N_n(x)} Y_k + C}{\sum_{k=2}^{N_n(x)}(T_k - T_{k-1}) + D}.$$

where $T_k = T_x^{(k)}$ is the time of the k-th (re)visit to x. The negligible terms C and D come from the first last cycles Y_1 and $Y_{N_n(x)}$; they are negligible because we let $n \to \infty$. By strong law of large numbers,

$$\frac{\sum_{k=2}^{N_n(x)} Y_k + C}{\sum_{k=2}^{N_n(x)}(T_k - T_{k-1}) + D} = \frac{\frac{1}{N_n(x)-1}\sum_{k=2}^{N_n(x)} Y_k + C}{\frac{1}{N_n(x)-1}\sum_{k=2}^{N_n(x)}(T_k - T_{k-1}) + D}$$

$$\xrightarrow{n\to\infty} \frac{\mathbb{E}[Y_k]}{\mathbb{E}_x[T_x]} = \frac{\pi f^T/\pi(x)}{1/\pi(x)} = \pi f^T.$$

This concludes the proof. □

Section 6. Application of Main Theorems

5.15. Example: Consider the DTMC represented by the following transition matrix:

$$P = \begin{bmatrix} 0.1 & 0.2 & 0.4 & 0.3 \\ 0.1 & 0.2 & 0.4 & 0.3 \\ 0.3 & 0.4 & 0.3 & 0 \\ 0.1 & 0.2 & 0.4 & 0.3 \end{bmatrix}$$

We make the following elementary observations:

- Irreducible: Observe that $P_{03}, P_{32}, P_{21}, P_{10} > 0$. Thus, all states are in the same class.
- Aperiodic: $P_{00} > 0$ so 0 is aperiodic and thus all states are aperiodic.
- Recurrent: This is an irreducible DTMC with finite states.
- Solve for stationary distribution: $\pi P = \pi$, $\pi \mathbf{1} = 1$ gives
$$\pi = \left[\frac{19}{110}, \frac{30}{110}, \frac{40}{110}, \frac{21}{110}. \right].$$

By previous results, we know that

- The limiting transition probabilities are given by
$$\lim_{n\to\infty} P_{xy}^n = \pi(y).$$
- Long-run fraction of time visiting y is given by
$$\lim_{n\to\infty} \frac{N_n(y)}{n} = \pi(y).$$
- The expected time that the chain revisits y after starting at y is
$$\mathbb{E}_y[T_y] = \frac{1}{\pi(y)}.$$
- Long-run average is given by πf^T.

5.16. Remark: The stationary distribution is typically easy to find, so the above results are usually used to find the other related quantities.

Section 7. Roles of Different Conditions in Main Theorems

5.17. Motivation: In this section, we explore the roles of different conditions in main theorems. In particular, we look at a set of counterexamples, examine what could go wrong if one condition is missing.

5.18. Note (Missing Irreducibility)**:** *Irreducibility is related to the uniqueness of the stationary distribution.* Recall an irreducible chain has a unique stationary distribution (if exists) π. As a counter example, if $P = I$ (so the chain has 2 classes), then every π satisfying $\pi\mathbf{1} = 1$ is a stationary distributions. As a result, $\lim_{n\to\infty} P_{xy}^n$ and $\lim_{n\to\infty} \Pr(X_n = y)$ will depend on the initial state and the initial distribution μ.

5.19. Note (Missing Aperiodicity)**:** *Aperiodicity is related to the existence of the limiting distribution* $\lim_{n\to\infty} P_{xy}^n$*.* Recall if y is aperiodic, then $\lim_{n\to\infty} P_{xy}^n$ exists. As a counterexample, consider

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

so that $d(0) = d(1) = 2$. Note that $P^2 = I$, so that $P^{2n} = I$ and $P^{2n+1} = P \neq I$ for all $n \in \mathbb{N}$. In particular, the sequence $\{P^k\}_{k\in\mathbb{N}}$ does not converge and there is no limiting distribution for this DTMC.

5.20. Note (Missing Recurrence)**:** *Recurrence is related to the existence of the stationary distribution* π*.* More precisely, if the MC is irreducible and recurrent, then a stationary measure exists. We will later introduce positive recurrence which is equivalent to the existence of a stationary distribution.

Chapter 6. Exit Distribution and Exit Time

Section 1. Exit Distribution

6.1. Motivation: We now look at the temporary behaviours of a DTMC.

- If a DTMC starts from a transient state and will eventually enter a recurrent class, when will this happen? This time is known as the **exit time** or **absorption time**.
- If there are more than one recurrent classes, which one will the chain enter? This probability is known as the **exit probability** or **absorption probability**.

6.2. Note (Exit Probability)**:** Let $A, B \subseteq S$ and $C := S \setminus A \cup B$ be finite. Starting from a state in C, *what is the probability that the chain exits C by entering A*? Define

$$V_A = \min\{n \geq 0 : X_n \in A\}$$
$$V_B = \min\{n \geq 0 : X_n \in B\}$$

to be the time that the MC first visits A, B, respectively. We are interested in

$$\Pr_x(V_A < V_B).$$

6.3. Example: Consider the following example.

$$P = \begin{pmatrix} 0.25 & 0.6 & 0 & 0.15 \\ 0 & 0.2 & 0.7 & 0.1 \\ & & 1 & \\ & & & 1 \end{pmatrix}$$

Define $A = \{3\}, B = \{4\}, C = \{1, 2\}$. Note that $P_{33} = P_{44} = 1$ is irrelevant here as we are only interested in the chain before it hits 3 or 4. Define $h(1) = \Pr_1(V_3 < V_4)$ and $h(2) = \Pr_2(V_3 < V_4)$. Then we have

$$h(1) = \Pr_1 (V_3 < V_4) = \sum_{x=1}^{4} \Pr (V_3 < V_4 \mid X_1 = x, X_0 = 1).$$

Since

$$\Pr (V_3 < V_4 \mid X_1 = x, X_0 = 1) = \begin{cases} \Pr_1 (V_3 < V_4) = h(1) & x = 1 \\ \Pr_2 (V_3 < V_4) = h(2) & x = 2 \\ 1 & x = 3 \\ 0 & x = 4 \end{cases},$$

we have $h(1) = 0.25h(1) + 0.6h(2)$. A similar analysis on state 2 yields $h(2) = 0.2h(2) + 0.7$. Solving this system of equations, we have $h(1) = 0.7$ and $h(2) = 0.875$. The idea we used here is called *first-step analysis* and will be used to derive general results below.

6.4. Theorem (Exit Probability)**:** *Let $S = A \cup B \cup C$ where A, B, C are disjoint and C is finite. If the chain will eventually visit A or B after starting at any state $x \in C$, i.e.,*

$$\forall x \in C : \Pr_x(\min\{V_A \wedge V_B\} < \infty) > 0,$$

then (the probability that we hit A before B after starting at x)

$$h(x) := \Pr_x(V_A < V_B)$$

is the unique solution of the system of equations

$$h(x) = \sum_y P_{xy} h(y), \quad x \in C$$

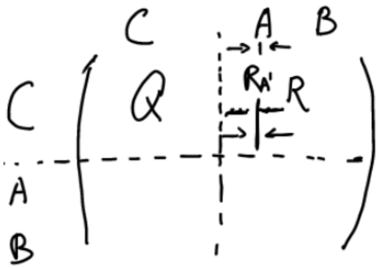
with boundary conditions $h(a) = 1$ for all $a \in A$ and $h(b) = 0$ for all $b \in B$.

Proof. We show existence first. By first-step analysis,

$$\begin{aligned} h(x) &= \Pr\left(V_A < V_B \mid X_0 = X\right) \\ &= \sum_{y \in S} \Pr\left(V_A < V_B \mid X_0 = x, X_1 = y\right) \cdot \Pr\left(X_0 = x \mid X_1 = y\right) \\ &= \sum_{y \in S} P_{xy} h(y) \end{aligned}$$

and the boundary conditions hold trivially. Hence, we only need to look at the uniqueness. Let us rewrite the system of equations as

$$h^T = Qh^T + R_A^T$$



where

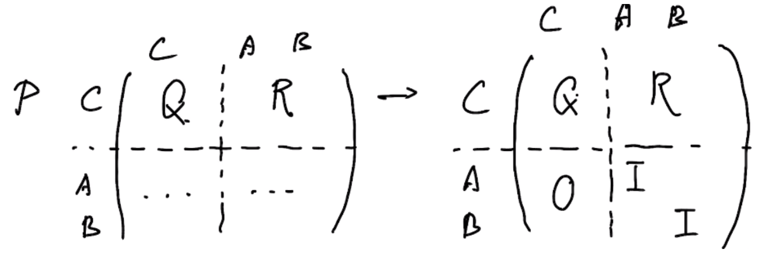
- $h = (h(x_1), h(x_2), \ldots)^T$ for $x_1, x_2, \ldots \in C$
- Q: the submatrix corresponding to the transition probabilities from $c_1 \in C$ to $c_2 \in C$;
- R: the submatrix corresponding to the transition probabilities from $c \in C$ to $a \in A$ or $b \in B$.
- R_A^T: the column vector corrsponding to the sum of transition probabilities from each $c \in C$ to all states in A, i.e.,

$$R_A^T = \begin{pmatrix} \sum_{y \in A} P_{x_1,y} \\ \sum_{y \in A} P_{x_2,y} \\ \vdots \end{pmatrix}$$

To justify the equation above, observe that

$$\begin{aligned}
h(x) &= \sum_{y\in S} P_{xy}h(y) \\
&= \sum_{y\in C} P_{xy}h(y) + \sum_{y\in A} P_{xy}h(y) + \sum_{y\in B} P_{xy}h(y) \\
&= \sum_{y\in C} P_{xy}h(y) + \sum_{y\in A} P_{xy} + 0 && \forall a \in A : h(a) = 1; \forall b \in B : h(b) = 0 \\
&= [Qh^T](x) + [R_A^T](x).
\end{aligned}$$

Then $Ih^T = Qh^T + R_A^T \implies (I-Q)h^T = R_A^T$, which implies that h^T is unique as long as $I-Q$ is invertible. Now modify the transition matrix to obtain P':



Since we are only interested in observing the chain before it enters A or B, changing the transition probabilities going out of states in A or B will not change the result of this problem. After this change, A and B become absorbing and all states in C become transient (as $\Pr_x(V_A \wedge V_B < \infty) > 0$). Let X' denote the modified chain with transition matrix P'. To show $I-Q$ is invertible, note that since the states in C are transient (in P'), we have

$$0 = \lim_{n\to\infty} \Pr_x(X'_n \in C) = \lim_{n\to\infty} \sum_{y\in C} ((P')^n)_{xy} = \lim_{n\to\infty} \sum_{y\in C} (Q^n)_{xy}.$$

The last equality holds because of the block structure of P':

$$P' = \begin{bmatrix} Q & R \\ 0 & I \end{bmatrix}.$$

This corresponds to the fact that in order to have $X'_n \in C$, we must have $X'_0, \ldots, X'_{n-1} \in C$ which implies that $\lim_{n\to\infty} Q^n = O$ (the zero matrix). Then, all the eigenvalues of Q have norm smaller than 1. Thus, there does not exist a non-zero column vector f^T such that $If^T = Qf^T$, so there is no f^T such that $(I-Q)f^T = 0$. It follows that $I-Q$ is invertible. □

6.5. Remark: We see that the function h in the above theorem satisfies

$$h(x) = \sum_{y} P_{xy}h(y) = \mathbb{E}_x[h(X_1)] := h^{(1)}(x), \quad x \in C,$$

i.e., it is a *harmonic function.*

6.6. Definition: A function f is called **harmonic** at state x if

$$f(x) = \sum_y P_{xy} f(y) = \mathbb{E}_x[f(X_1)] = f^{(1)}(x).$$

The function f is **harmonic** in $A \subseteq S$ it is harmonic at every state in A. The function f is **harmonic** if it is harmonic in S, or equivalently, $f^T = P f^T = (f^{(1)})^T$.

6.7. Remark: In the proof, we have seen that

$$h^T = (I - Q)^{-1} R_A^T.$$

This is the matrix formula to calculate $h(x) = \Pr_x(V_A < V_B)$.

Section 2. Exit Time

6.8. Motivation: We are now interested in the **expected time** that the chain exits part of the state space. Let $S = A \cup C$ where A, C are disjoint and C is finite. Define

$$V_A := \min\{n \geq 0 : X_n \in A\}$$

to be the first time the chain exists C, or equivalently, the first time the chain visits A. We are interested in $\mathbb{E}_x(V_A) = \mathbb{E}[V_A \mid X_0 = x]$ for $x \in C$.

6.9. Example: Consider the same example.

$$P = \begin{pmatrix} 0.25 & 0.6 & 0 & 0.15 \\ 0 & 0.2 & 0.7 & 0.1 \\ & & 1 & \\ & & & 1 \end{pmatrix}$$

Define $C = \{1, 2\}, A = \{3, 4\}$. We are interested in $g(1)$ and $g(2)$. Note that $g(3) = g(4) = 0$. Using first-step analysis, we obtain

$$\begin{aligned} g(1) &= 1 + 0.25g(1) + 0.6g(2) \\ g(2) &= 1 + 0.2g(2) \end{aligned}$$

which gives us $g(1), g(2) = (7/3, 5/4)$.

6.10. Theorem: *Let $S = A \cup C$ where A, C are disjoint and C is finite. If the chain eventually visits A after starting at x, i.e.,* $\Pr_x(V_A < \infty)$ *for all $x \in C$, then*

$$g(x) = \mathbb{E}_x[V_A], \quad x \in C$$

is the unique solution to the system of equations

$$g(x) = 1 + \sum_{y \in S} P_{xy} g(y), \quad x \in C$$

with the boundary conditions $g(a) = 0$ for all $a \in A$.

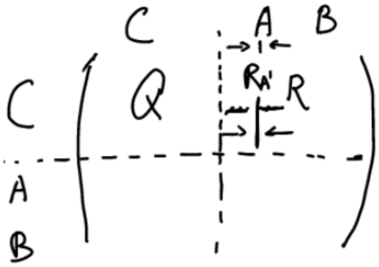
Proof. Existence is simple:

$$\begin{aligned}
g(x) &= \mathbb{E}\left[V_A \mid X_0 = x\right] \\
&= \sum_{y \in C} \left(\mathbb{E}\left[V_A \mid X_1 = y, X_0 = x\right] + 1\right) \mathbb{P}\left(X_1 = y \mid X_0 = x\right) \\
&\qquad + \sum_{y \in A} \left(\mathbb{E}\left[V_A \mid X_1 = y, X_0 = x\right] + 1\right) \mathbb{P}\left(X_1 = y \mid X_0 = x\right) \\
&= \sum_{y \in C} (g(y) + 1) P_{x,y} + \sum_{y \in A} P_{x,y} \\
&= 1 + \sum_{y \in C} P_{x,y} g(y) \\
&= 1 + \sum_{y \in \mathcal{S}} P_{x,y} g(y) && \forall a \in A : g(a) = 0.
\end{aligned}$$

Note the 1 in $(g(y) + 1)$ reflects the fact that we have walked one step. We now look at uniqueness. For $x \in S$,

$$g(x) = 1 + \sum_{y \in S} P_{xy} g(y) \implies g^T = \mathbf{1}^T + Q g^T$$

where Q corresponds to the $|C| \times |C|$ submatrix of P and R corresponds to the $|C| \times |A|$ submatrix of P as before.



Then

$$I g^T = \mathbf{1}^T + Q g^T \implies (I - Q) g^T = \mathbf{1}^T \implies g^T = (I - Q)^{-1} \mathbf{1}^T.$$

We are looking at exactly the same matrix $I - Q$ as in the exit distribution part. By the previous theorem, when $I - Q$ is invertible, g' is the unique solution. □

Chapter 7. DTMCs with Infinite State Space

Section 3. Null Recurrence and Positive Recurrence

7.1. Note: Unless otherwise specified, all the results covered so far hold for all DTMCs (finite and infinite state space). We now discuss a pair of notions which only makes sense for infinite state spaces.

7.2. Definition: A state x is called **positive recurrent** if

$$\mathbb{E}_x[T_x] = \mathbb{E}[T_x \mid X_0 = x] < \infty.$$

A recurrent state x is called **null recurrent** if

$$\mathbb{E}_x[T_x] = \mathbb{E}[T_x \mid X_0 = x] = \infty.$$

7.3. Intuition: Recall a state is recurrent iff $\Pr_x(T_x < \infty) = 1$, i.e., starting at x, you will always go back to x. A state is positive recurrent if the expected time it takes to go back to x is finite. Thus, a positive recurrent state is always recurrent (note we didn't put "recurrent" in the definition of positive recurrence). On the other hand, a state is null recurrent if it is recurrent, meaning we will eventually go back to it, but the expected amount of time is infinite.

7.4. How can it be possible for a random variable which is finite with probability 1 to have an infinite first moment? See the next example.

7.5. Example (St. Petersburg Paradox)**:** We now demonstrate how a random variable could be finite with probability 1 but with infinite mean. Let $X = 2^n$ with probability 2^{-n} for $n = 1, 2, \ldots$. Then

$$\sum_{n=1}^{\infty} 2^{-n} = 1 \implies \Pr(X < \infty) = 1.$$

But

$$\begin{aligned}\mathbb{E}[X] &= 2 \times \frac{1}{2} + 4 \times \frac{1}{4} + 8 \times \frac{1}{8} + \cdots \\ &= 1 + 1 + \cdots \\ &= \infty.\end{aligned}$$

Section 4. Positive Recurrence and Stationarity

7.6. Theorem: *For an irreducible DTMC, the following are equivalent:*

1. *Some state is positive recurrent.*
2. *There exists a stationary distribution π.*
3. *All the states are positive recurrent.*

Proof. $(3 \Rightarrow 1)$: Trivial.

$(1 \Rightarrow 2)$: Let x be a positive recurrent state. By Theorem 4.5, we know that a recurrent state can give us a stationary measure

$$\mu_x(y) = \sum_{n=0}^{\infty} \Pr_x(X_n = y, T_x > n), \quad y \in S.$$

which corresponds to "starting from x, the expected number of visits to y before returning to x". It can be normalized to become a stationary distribution iff

$$\sum_y \mu_x(y) < \infty.$$

Note that

$$\begin{aligned}
\sum_{y \in S} \mu(y) &= \sum_{y \in S} \sum_{n=0}^{\infty} \Pr_x\left(X_n = y, T_x > n\right) \\
&= \sum_{y \in S} \sum_{n=0}^{\infty} \mathbb{E}_x\left(\mathbf{1}_{\{X_n=y\}} \mathbf{1}_{\{T_x>n\}}\right) \\
&= \mathbb{E}_x\left[\sum_{n=0}^{\infty} \mathbf{1}_{\{T_x>n\}} \sum_{y \in S} \mathbf{1}_{\{X_n=y\}}\right] \\
&= \mathbb{E}_x\left[\sum_{n=0}^{\infty} \mathbf{1}_{\{T_x>n\}}\right] & \sum_{y \in S} \mathbf{1}_{\{X_n=y\}} = 1 \\
&= \mathbb{E}_x\left[T_x\right] < \infty \Longrightarrow \pi(y) = \frac{\mu(y)}{\mathbb{E}_x\left(T_x\right)}
\end{aligned}$$

as x is positive recurrent.

$(2 \Rightarrow 3)$: Since the DTMC is irreducible with a stationary distribution π, we have $\pi(x) > 0$ for all $x \in S$. Since the DTMC is irreducible and recurrent,

$$\pi(x) = \frac{1}{\mathbb{E}_x[T_x]} \implies \mathbb{E}_x[T_x] = \frac{1}{\pi(x)} < \infty.$$

It follows that all states $x \in S$ are positive recurrent. □

7.7. Corollary: *Positive recurrence and null recurrence are class properties.*

Proof. Let $x \in S$ be positive recurrent and C be the class containing x. Since C is recurrent, it is closed. Since for any $y \in C$, the chain starting from y will only move in C, we can focus on C and consider the chain restricted on C, with transition matrix $P|_C = \{P_{xy}\}_{x,y\in C}$. Note that

$$\begin{array}{cc} & \begin{array}{cc} C & C^c \end{array} \\ \begin{array}{c} C \\ C^c \end{array} & \begin{pmatrix} P|_C & 0 \\ & \end{pmatrix} \end{array}$$

where the top right block equals 0 since C is closed. This restricted chain is irreducible and has a positive recurrent state 0. Another remark is that

$$\mathbb{E}_x[T_x]|_P = \mathbb{E}_x[T_x]|_{P|_C}$$

as the left quantity is the expected time of the first revisit in the original chain and the right is that in the restricted chain. By the previous theorem, x is positive recurrent implies all its states are positive recurrent, so the states in C are positive recurrence. Since both positive recurrence and recurrence are class properties, so is null recurrence. □

7.8. Corollary: *A state x is positive recurrent iff there exists a stationary distribution π such that $\pi(x) > 0$.*

Proof. Note that x is recurrent in both directions, so it suffices to prove the result for the case where the chain is irreducible. (Otherwise, we can consider the chain restricted on the closed class containing x.) The backward direction is given by the previous theorem. For the forward direction, by the previous theorem, since x is positive recurrent, there exists a stationary distribution π. Recall that an irreducible, stationary MC implies that $\pi(x) > 0$ for all x, hence $\pi(x) > 0$. □

7.9. Corollary: *A DTMC with finite state space must have at least one positive recurrent state.*

Proof. WLOG, assume the MC is irreducible. We already know it must be recurrent. Take $x \in S$. Then $\mu_x(y) = \sum_{n=0}^{\infty} \Pr_x(X_n = y, T_x > n)$ gives a stationary measure. Moreover, since there are only finitely many terms, the summation $\sum_{y\in S} \mu_x(y)$ is trivially finite, which implies that $\{\mu_x(y)\}_{y\in S}$ is normalizable and

$$\left\{ \pi(y) := \frac{\pi_x(y)}{\sum_{y\in S} \mu_x(y)} \right\}_{y\in S}$$

is a stationary distribution. Thus, x must be positive recurrent. □

7.10. Corollary: *A DTMC with finite state space does not have any null recurrent state.*

Proof. Suppose there is a null recurrent state, so a null recurrent class exists. Since it is recurrent, it is closed. Consider the chain restricted to this class. The restricted chain is irreducible and null recurrent. However, since it only has a finite number of states, it must have a positive recurrent state. Contradiction. Hence, there is no null recurrent state. □

7.11. We can rephrase the statement above as, *a null recurrent class must have an infinite number of states.* Intuitively, recall

$$\frac{1}{\mathbb{E}_x[T_x]} = \lim_{n\to\infty} \frac{N_n(x)}{n}$$

is the long-run fraction of time spent in state x. Then $\mathbb{E}_x[T_x] = \infty$ says the long-run fraction is 0. This can happen only if there are infinitely many such states.

Section 5. Example: Simple Random Walks

7.12. Example: Let us revisit the simple random walk example, where $S = \mathbb{Z}$, $P_{i,i+1} = p$ and $P_{i,i-1} = 1 - p = q$ with $p \in (0,1)$. We already know that this DTMC is irreducible with period 2. We here show that the MC is transient for $p \neq 1/2$ and null recurrent for $p = 1/2$.

Proof. First assume $p \neq 1/2$. WLOG, assume $p > 1/2$ (symmetry). Write $X_n = Y_1 + \cdots + Y_n$ where Y_i's are iid random variables with probability distribution

$$Y_n = \begin{cases} 1 & p \\ -1 & 1-p \end{cases}$$

Then $\mathbb{E}[Y_n] = 1 \cdot p + (-1)(1-p) = 2p - 1 > 0$. By Strong Law of Large Numbers,[1]

$$\frac{X_n}{n} = \frac{1}{n}\sum_{m=1}^{n} Y_m \xrightarrow{n\to\infty}_{a.s.} \mathbb{E}[Y_1] = 2p - 1.$$

Since the average converges to a constant, the sum goes to infinity:

$$X_n \xrightarrow{n\to\infty}_{a.s.} \infty.$$

Since X_n keeps going towards the right direction, for any state $i \geq 0$ (in particular, state 0), there is a last visit time to i. This implies that 0 is transient and $\{X_n\}$ is transient.

Now suppose $p = 1/2$. We wish to show that the chain is recurrent but not positive recurrent. Recall a state i is recurrent iff $\sum_{n=0}^{\infty} P_{ii}^{(n)} = \infty$. For state 0, we have $P_{00}^{2n+1} = 0$ as the chain has a period of 2 and

$$P_{00}^{2n} = \binom{2n}{n}\left(\frac{1}{2}\right)^n\left(\frac{1}{2}\right)^n = \binom{2n}{n}\left(\frac{1}{4}\right)^n.$$

Using Stirling's formula, which says that the factorial of n has the following behavior:

$$n \to \infty \implies n! \sim \sqrt{2\pi}e^{-n}n^{n+(1/2)},$$

we have (omit computational details)

$$\binom{2n}{n}\left(\frac{1}{4}\right)^n \sim \frac{1}{\sqrt{\pi n}} \implies \sum_{n=1}^{\infty}\binom{2n}{n}\left(\frac{1}{4}\right)^n = \infty$$

and state 0 is recurrent. Next, we show that 0 is not positive recurrent by showing there does not exist a stationary distribution. Consider the system of equations $\pi P = \pi$. We have

$$\pi(x) = \frac{1}{2}\pi(x-1) + \frac{1}{2}\pi(x+1)$$
$$\vdots$$
$$\pi(x+1) - \pi(x) = \pi(x) - \pi(x-1)$$

[1] a.s.: Almost surely, or with probability of 1.

Since this holds for all $x \in \mathbb{Z}$, $\{\pi(x)\}$ is an arithmetic series. The general form is

$$\pi(x) = \pi(0) + ax$$

where $a = \pi(1) - \pi(0)$. Also, $\pi(x) \in [0, 1]$ for all $x \in \mathbb{Z}$. This forces $a = 0$ which implies that $\pi(x) = \pi(0)$ for all $x \in \mathbb{Z}$. If $\pi(0) = 0$, then π is a zero vector and the sum of entries equals $0 \neq 1$. If $\pi(0) > 0$, then the sum of entries is $\infty \neq 1$. Thus, the normalization condition $\sum_x \pi(x) = 1$ can never hold. It follows that no stationary distribution exists and the chain is not positive recurrent. Hence it is null recurrent. □

7.13. Example: Now consider a SRW with a reflecting barrier, i.e., $S = \{0, 1, 2, \ldots\}$ with $P_{x,x+1} = p$, $P_{x,x-1} = 1 - p$ for $x \geq 1$ and $P_{01} = 1$. Intuitively, whenever you visits 0, your next step will always take you back to 1. We show that if $p < 1/2$, then the chain is positive recurrent. This shows that *a positive recurrent class can have infinitely many states*.

Proof. We solve for the stationary distribution to figure out the requirements for its existence. Since only $P_{x,x+1}$ and $P_{x,x-1}$ are non-zero, we can use the detailed balance condition:

$$\pi(0) \cdot 1 = \pi(1)(1-p) \Longrightarrow \pi(1) = \frac{1}{1-p}\pi(0)$$
$$\pi(i) \cdot p = \pi(i+1)(1-p) \Longrightarrow \pi(i+1) = \frac{p}{1-p}\pi(i), \qquad i \geq 1$$

This gives

$$\begin{aligned}\pi(i) = \frac{p}{1-p}\pi(i-1) &= \left(\frac{p}{1-p}\right)^2 \pi(i-2) = \cdots \\ &= \left(\frac{p}{1-p}\right)^{i-1} \pi(i) \\ &= \left(\frac{p}{1-p}\right)^{i-1} \frac{1}{1-p}\pi(0)\end{aligned}$$

Note that $\pi(1), \pi(2), \ldots$ form a geometric series with ratio

$$\frac{p}{1-p}.$$

Thus, $\sum_{x=0}^{\infty} \pi(x) < \infty$ and a stationary distribution exists iff the ratio satisfies

$$\frac{p}{1-p} < 1 \iff p < 1/2.$$

□

7.14. Remark: To summarize, the reflected simple random walk is

- positive recurrent iff $p < 1/2$;
- null recurrent iff $p = 1/2$;
- transient iff $p > 1/2$.

Chapter 8. Miscellaneous

Section 6. The Metropolis-Hastings Algorithm

8.1. Metropolis-Hastings is a *Markov chain Monte-Carlo* (MCMC) method for obtaining a sequence of random samples from a probability distribution from which direct sampling is difficult. This sequence can be used to approximate the distribution (e.g., to generate a histogram) or to compute an integral (e.g., an expected value)

8.2. (Cont'd): The MH algorithm can draw samples from any probability distribution $P(x)$, provided that we know a function $f(x)$ propositional to the density of P and the values of $f(x)$ can be calculated. The requirement that $f(x)$ must only be proportional to the density, rather than exactly equal to it, makes the algorithm particularly useful, because calculating the necessary normalization factor is often extremely difficult in practice.

8.3. (Cont'd): The MH algorithm works by generating a sequence of sample values in such a way that, as more and more sample values are produced, the distribution of values more closely approximates the desired distribution $P(x)$. These sample values are produced iteratively, with the distribution of the next sample being dependent only on the current sample value (thus making the sequence of samples into a MC). Specifically, at each iteration, the algorithm picks a candidate for the next sample value based on the current sample value. Then, with some probability, the candidate is either accepted (in which case the candidate value is used in the next iteration) or rejected (in which case the candidate value is discarded, and current value is reused in the next iteration) – the probability of acceptance is determined by comparing the values of the function $f(x)$ of the current and candidate sample values with respect to the desired distribution $P(x)$.

8.4. Motivation: Suppose we wish to sample from a probability distribution

$$\pi = \{\pi(x)\}_{x \in S}$$

but direct sampling is hard to implement. We can construct a DTMC (which is easy to simulate), then modify it to get another DTMC, for which π is the unique stationary distribution of that DTMC. Then after long enough, the distribution of the DTMC will converge/approach the stationary distribution π. Here's a summary:

1. Given a distribution π for which direct sampling is hard to implement.
2. Construct a DTMC $\mathcal{X}_1$.
3. Modify it to another DTMC $\mathcal{X}_2$, for which π is its unique stationary distribution.
4. Wait for long enough, until the distribution of $\mathcal{X}_2$ converges to π.
5. Now follow the trace of $\mathcal{X}_2$; the results can be viewed as sampling from π.

8.5. The Metropolis-Hastings Algorithm.

1. Start with an irreducible DTMC with transition matrix
$$Q = \{Q_{xy}\}_{x,y\in S}$$
and certain initial distribution (typically an initial state).

2. At each step,
   (a) Propose a move from the current state x to some state $y \in S$ according to Q_{xy}.
   (b) Accept the proposed move with probability
   $$r_{xy} = \min\left\{\frac{\pi(y)Q_{yx}}{\pi(x)Q_{xy}}, 1\right\}.$$
   (c) If the move is rejected, stay in x.

3. Wait for long enough, then start sampling from this MC.

8.6. Proposition: *The modified DTMC has π as its unique stationary distribution.*

Proof. By construction, the transition matrix of the modified MC is given by
$$P_{xy} = \begin{cases} Q_{xy} \cdot r_{xy} & x \neq y \\ 1 - \sum_{y\neq x} P_{xy} & x = y \end{cases}$$
We show that π satisfies the detailed balance condition, which implies that π is a stationary distribution. Fix $x, y \in S$. WLOG (rename if necessary), assume that $\pi(y)Q_{yx} \leq \pi(x)Q_{xy}$. This gives us
$$r_{xy} = \min\left\{\frac{\pi(y)Q_{yx}}{\pi(x)Q_{xy}}, 1\right\} = 1, \qquad r_{yx} = \frac{\pi(x)Q_{x,y}}{\pi(y)Q_{yx}}.$$
Therefore,
$$P_{xy} = Q_{xy} \cdot 1 = Q_{xy}$$
$$P_{yx} = Q_{yx} \cdot r_{yx} = \frac{\pi(x)Q_{xy}}{\pi(y)}$$
and thus the detailed balance condition is satisfied and π is a stationary distribution:
$$\pi(x)P_{xy} = \pi(x)Q_{xy} = \pi(y)P_{yx}.$$
It remains to show that DTMC converges. To use the convergence theorem, we need irreducibility and aperiodicity. Irreducibility is guaranteed by construction. Aperiodicity is almost always satisfied as the rejection rate is typically non-zero, meaning we have positive probability of staying in the same state, hence $P_{xx} > 0$. Thus, by the convergence theorem,
$$\lim_{n\to\infty} \Pr(X_n = x) = \pi(x).$$
□

Section 7. Branching Process (Galton Watson Process)

8.7. Motivation: Suppose we are studying a population. Starting from one common ancestor, each individual, at the end of its life, produces a random number Y of offspring. Moreover, the number of offspring of different individuals are independent. Suppose Y has probability distribution

$$\Pr(Y = k) = P_k, \quad k = 0, 1, \ldots,$$

with $P_k \geq 0$ and $\sum_{k=1}^{\infty} P_k = 1$. Let X_n denote the number of individuals in the nth generation. Then we have

- $X_0 = 1$
- $X_{n+1} = Y_1^{(n)} + \cdots + Y_{X_n}^{(n)}$,

where $Y_1^{(n)}, \ldots, Y_{X_n}^{(n)}$ are independent copies of Y and $Y_i^{(n)}$ is the number of offspring of the ith individual in the nth generation. We first look at the expected number of offspring in generation n. As expected, it is the nth power of the average number of offspring per individual.

8.8. Proposition: *The expected number of offspring in generation n is $(\mathbb{E}[Y])^n$.*

Proof. Assuming $\mathbb{E}[Y] = \mu$. Then

$$\begin{aligned}
\mathbb{E}\left(X_{n+1}\right) &= \mathbb{E}\left(Y_1^n + Y_2^{(n)} + \cdots + Y_{X_n}^{(n)}\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(Y_1^n + Y_2^{(n)} + \cdots + Y_{X_n}^{(n)} \mid X_n\right)\right) \\
&= \mathbb{E}\left(\mu X_n\right) = \mu \mathbb{E}\left(X_n\right)
\end{aligned}$$

Note that $\mathbb{E}[X_1 + \cdots + X_N] = \mathbb{E}[N]\mathbb{E}[X_1]$ is known as **Wald's identity**. Continuing inductively, we see that $\mathbb{E}[X_n] = \mu^n \mathbb{E}[X_0] = \mu^n$ with $n = 0, 1, \ldots$ □

8.9. Motivation: What's the probability that the population dies out? Observe that $\{X_n\}_{n=0}^{\infty}$ is a DTMC with state 0 being absorbing (the population is no longer able to produce any new individual) and all other states are transient, as long as $P_0 > 0$. (Indeed, if $P_0 = 0$, then the population is guaranteed to exist forever.) However, $P_0 > 0$ does not imply that the population will extinct for sure. In particular, if the probability that the population size goes to ∞ is positive, then the probability of extinction is smaller than 1. To find the extinction probability, we introduce the concept of *generating functions.*

8.10. Definition: Let $\{P_0, P_1, \ldots\}$ be a distribution on $\{0, 1, \ldots\}$. Let η be a random variable following $\{P_0, P_1, \ldots\}$, that is, $\Pr(\eta = i) = P_i$. The **generating function** of η, or of the distribution $\{P_0, P_1, \ldots\}$, is defined by

$$\phi(s) = \mathbb{E}[s^\eta] = \sum_{k=0}^{\infty} P_k s^k, \qquad 0 \leq s \leq 1.$$

8.11. Proposition: *Let ϕ be a generating function of η, or of the distribution $\{P_0, P_1, \dots\}$.*

1. *$\phi(0) = P_0$, $\phi(1) = 1$.*
2. *The generating function completely determines the distribution. In particular,*
$$\Pr(\eta = k) = P_k = \frac{1}{k!} \frac{d^k \phi(s)}{ds^k}\bigg|_{s=0}.$$
3. *Let $\eta_1, \dots, \eta_n$ be independent random variables with generating functions $\phi_1, \dots, \phi_n$, then the generating function of $X = \eta_1 + \cdots + \eta_n$ is given by $\phi_X(s) = \phi_1(s) \cdots \phi_n(s)$.*
4. *The kth moment of ϕ is given by*
$$\begin{aligned}\frac{d^k \phi(s)}{ds^k}\bigg|_{s=1} = \frac{d^k \mathbb{E}\left[s^\eta\right]}{ds^k}\bigg|_{s=1} &= \mathbb{E}\left[\frac{d^k s^\eta}{ds^k}\bigg|_{s=1}\right] \\ &= \mathbb{E}\left[(\eta(\eta-1)\cdots(\eta-k+1)s^{\eta-k}\big|_{s=1}\right] \\ &= \mathbb{E}[\eta(\eta-1)\cdots(\eta-k+1)]\end{aligned}$$
In particular,
$$\begin{aligned}\mathbb{E}(\eta) &= \phi'(1) \\ \mathrm{Var}(\eta) &= \varphi''(1) + \varphi'(1) - (\varphi'(1))^2\end{aligned}$$

Proof. For statement 2, consider the Taylor expansion of $\phi(s)$ at 0:
$$\varphi(s) = P_0 + P_1 s^1 + \cdots + P_{k-1}s^{k-1} + P_k s^k + P_{k+1}s^{k+1} + \dots.$$
Taking the kth derivative of both sides, we have
$$\frac{d^k \varphi(s)}{ds^k} = k! P_k + \underbrace{(\cdots)s + (\cdots)s^2 + \dots}_{\text{non-negative coefficients}}$$
which is non-negative for all $k \in \mathbb{Z}_{\geq 0}$ and $s \in [0,1]$¿ Evaluating at $s = 0$, all but the first term are gone:
$$\frac{d^k \varphi(s)}{ds^k}\bigg|_{s=0} = k! P_k \Longrightarrow P_k = \frac{1}{k!}\frac{d^k \varphi(s)}{ds^k}\bigg|_{s=0}$$
Since the derivatives are non-negative, we see that $\varphi(s)$ is increasing and convex.

For statement 3, observe that
$$\begin{aligned}\varphi_x(s) &= \mathbb{E}(s^X) \\ &= \mathbb{E}[s^{\eta_1 + \cdots + \eta^n}] \\ &= \mathbb{E}(s^{\eta_1} \cdots s^{\eta_n}) \\ &= \mathbb{E}(s^{\eta_1}) \cdots \mathbb{E}(s^{\eta_n}) \quad \text{independence} \\ &= \varphi_1(s) \dots \varphi_n(s)\end{aligned}$$

□

8.12. Note: Back to extinction probability. Define $N = \min\{n : X_n = 0\}$ to be the **extinction time** and $u_n = \Pr(N \le n) = \Pr(X_n = 0)$ to be the probability that extinction happens before or at time n. Note that u_n is non-decreasing and bounded from above by monotone convergence theorem. Thus, the notion of extinction u probability is well-defined:

$$u := \lim_{n\to\infty} u_n = \Pr(N < \infty)$$

The key step is to note that we have the following relation between u_n and u_{n-1}:

$$u_n = \sum_{k=0}^{\infty} P_k u_{n-1}^k = \phi(u_{n-1}),$$

where ϕ is the generating function of $\{P_0, P_1, \ldots\}$, or equivalently, the generating function of Y. To justify this, note that each sub-population has the same distribution as the entire population. The entire population dies in n steps iff each sub-population initiated by an individual in generation 1 dies out in $n-1$ steps. In other words,

$$\begin{aligned} u_n &= \Pr(N \le n) \\ &= \sum_k \Pr\left(N \le n \mid X_1 = k\right) \Pr\left(X_1 = k\right) \\ &= \sum_k \Pr\left(N_1 \le n-1, \ldots, N_k \le n-1 \mid X_1 = k\right) P_k \\ &= \sum_k P_k u_{n-1}^k = \varphi\left(u_{n-1}\right) \end{aligned}$$

where N_m is the number of steps for the subpopulation m to die out. We can thus reduce the problem to the following: with initial value $u_0 = 0$ (since $X_0 = 1$) and relation $u_n = \phi(u_{n-1})$, what is $\lim_{n\to\infty} u_n = u$?

8.13. Theorem: *The extinction probability u is the smallest intersection of $\phi(s)$ and $f(s) = s$, or equivalently, it is the smallest solution of $\phi(s) = s$ between 0 and 1.*

Proof. Recall that $\phi(0) = p_0 > 0$ and $\phi(1) = 1$, and since ϕ is non-decreasing and convex, $\phi(s) = s$ always has a solution between 0 and 1. □

8.14. Note: There are two possible cases: $u < 1$ or $u = 1$, where the second case means the extinction is inevitable. How to tell whether we are in Case 1 or Case 2? The answer is to check the derivative of ϕ at $s = 1$. In particular, if $\phi'(1) > 1$ then we are at Case 1. Otherwise, we are at Case 2. Moreover, recall that we know $\phi'(1) = \mathbb{E}[Y]$. Thus, we conclude that if $\mathbb{E}[Y] = 1$, then extinction happens with certain probability smaller than 1.

8.15. Intuition: On average, having more than 1 offspring for each individual implies that the population will probably explode, which diminish the chance to wipe out the whole population. On the other hand, if each individual has less than 1 offspring on average, then there is a risk that the population will die out.

Part II

Poisson Processes

Chapter 9. Poisson Processes

Section 1. Review: Exponential Distribution

9.1. Note: Let $T \sim \mathrm{Exp}(\lambda)$ be an exponential random variable. Then:

$$\begin{aligned} F_T(t) &= 1 - e^{-\lambda t} \cdot \mathbf{1}_{t \geq 0} \\ f_T(t) &= \lambda e^{-\lambda t} \cdot \mathbf{1}_{t \geq 0} \\ \mathbb{E}[T] &= 1/\lambda \\ \mathrm{Var}[T] &= 1/\lambda^2 \end{aligned}$$

9.2. Proposition (Rescaling Property of Exponential Distribution)**:**

- *If* $S \sim \mathrm{Exp}(1)$, *then* $S/\lambda \sim \mathrm{Exp}(\lambda)$.
- *If* $T \sim \mathrm{Exp}(\lambda)$, *then* $\lambda T \sim \mathrm{Exp}(1)$.

9.3. Theorem (Memoryless)**:** $\Pr(T > t + s \mid T > t) = \Pr(T > s)$.

Proof.

$$\Pr(T > t + s \mid T > t) = \frac{\Pr(T > t + s)}{\Pr(T > t)} = \frac{1 - (1 - e^{-\lambda(t+s)})}{1 - (1 - e^{-\lambda t})} = e^{-\lambda s} = \Pr(T > s).$$

□

9.4. Theorem (Min of Independent Exponentials)**:** *Let* $S \sim \mathrm{Exp}(\lambda)$, $T \sim \mathrm{Exp}(\mu)$, *and* $S \perp\!\!\!\perp T$. *Define* $Z := \min\{S, T\}$. *Then* $Z \sim \mathrm{Exp}(\lambda + \mu)$ *and*

$$\Pr(S = Z) = \Pr(S \leq T) = \frac{\lambda}{\lambda + \mu}.$$

Proof. For distribution of Z:

$$\begin{aligned} \Pr(Z > t) &= \Pr(S > t, T > t) \\ &= \Pr(S > t) \cdot \Pr(T > t) = e^{-\lambda t} e^{-\mu t} = e^{-(\lambda + \mu)t} \implies Z \sim \mathrm{Exp}(\lambda + \mu). \end{aligned}$$

For the second statement:

$$\begin{aligned} \Pr(S \leq T) &= \mathbb{E}[\Pr(S \leq T \mid S)] (= \mathbb{E}[\mathbf{1}_{\{S \leq T\}} \mid S]) \\ &= \int_0^\infty \Pr(S \leq T \mid S = s) \cdot f_S(s) \, ds \\ &= \int_0^\infty e^{-\mu s} \cdot \lambda e^{-\lambda s} \, ds = \frac{\lambda}{\lambda + \mu} \end{aligned}$$

□

9.5. Corollary: *Let $\{T_i\}_{i=1}^n$ be independent random variables with $T_i \sim \operatorname{Exp}(\lambda_i)$. Then $Z := \min\{T_1, \dots, T_n\} \sim \operatorname{Exp}(\lambda_1 + \cdots + \lambda_n)$ and*

$$\Pr(Z = T_i) = \Pr(T_i \leq T_1, \dots, T_i \leq T_n) = \frac{\lambda_i}{\lambda_i + \cdots + \lambda_n}.$$

Proof. Omitted. □

9.6. Intuition: "Competition" among independent exponential random variables will result in an exponential random variable, with the parameter being the sum of the parameters. The probability that the ith variable wins is

$$\frac{\lambda_i}{\lambda_i + \cdots + \lambda_n}.$$

9.7. To summarize, we have looked at three properties of the exponential distribution:

1. rescaling property;
2. memorylessness;
3. minimum of independent exponentials.

9.8. Note: If $W_1, \dots, W_n \overset{\text{iid}}{\sim} \operatorname{Exp}(\lambda)$, then $W_1 + \cdots + W_n \sim \operatorname{Erlang}(n, \lambda)$ with cdf

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

We will occasionally work with the Erlang distribution in this chapter.

Section 2. Review: Poisson Distribution

9.9. Note: Let $X \sim \text{Poi}(\lambda)$. Then

$$\begin{aligned} \Pr(X = n) &= e^{-\lambda}\frac{\lambda^n}{n!}, \quad n = 0, 1, \dots \\ \mathbb{E}[X] &= \lambda \\ \text{Var}[X] &= \lambda \end{aligned}$$

9.10. Theorem: *Let* $\{X_i\}_{i=1}^n$ *be independent Poisson random variables with* $X_i \sim \text{Poi}(\lambda_i)$*. Then their sum follows a Poisson distribution:*

$$X_1 + \cdots + X_n \sim \text{Poi}(\lambda_1 + \cdots + \lambda_n).$$

Proof. Consider the generating function ϕ_{X_i} of X_i:

$$\begin{aligned} \phi_{X_i}(s) &= \sum_{k=0}^{\infty} \Pr(X_i = k) \cdot s^k \\ &= \sum_{k=0}^{\infty} e^{-\lambda i}\frac{\lambda_i^k}{k!}s^k \\ &= \sum_{k=0}^{\infty} e^{-\lambda i}\frac{(\lambda_i s)^k}{k!} \\ &= e^{-\lambda_i}e^{\lambda_i s}\underbrace{\sum_{k=0}^{\infty} e^{-\lambda_i s}\frac{(\lambda_i s)^k}{k!}}_{=1} \\ &= e^{-\lambda_i(1-s)} \end{aligned}$$

as the summation is exactly the pmf of $\text{Poi}(\lambda_i s)$. By independence, the generating function of $X = X_1 + \cdots + X_n$, $\phi(s)$, is

$$\begin{aligned} \phi(s) &= \prod_{i=1}^{n} \phi_{X_i}(s) \\ &= \prod_{i=1}^{n} e^{-\lambda_i(1-s)} \\ &= e^{-(\lambda_1 + \cdots + \lambda_n)(1-s)}. \end{aligned}$$

This is the generating function of $\text{Poi}(\lambda_1 + \cdots + \lambda_n)$. By uniqueness of generating function, the result follows. □

Section 3. Poisson Process

9.11. Motivation: Recall that a DTMC is a discrete-time process. We now consider the simplest form of continuous-time process, the **counting process**, which counts the number of occurrences of certain events up to time $t \in \mathbb{R}_{\geq 0}$.

9.12. Definition: Let $0 \leq S_1 \leq S_2 \leq \cdots$ be the time of occurrence of some events. Then the process $\{N(t)\}_{t\geq 0}$ given by

$$N(t) := \Big|\{n : S_n \leq t\}\Big| = \sum_{n=1}^{\infty} \mathbf{1}_{\{S_n \leq t\}}$$

is called a **counting process** of the events $\{S_i\}_{i=1}^{\infty}$. Equivalently, $N(t) = n$ iff $S_n \leq t < S_{n+1}$.

9.13. Example: We can model the calls arrived at a call center using a counting process, where S_n are the arrival times and $N(t)$ is the number of calls received before time t.

9.14. Note: We have the following properties and assumptions:

- Non-negativity: $\forall t \geq 0 : N(t) \in \mathbb{Z}_{\geq 0}$.
- Non-decreasing: $\forall t_1 \leq t_2 : N(t_1) \leq N(t_2)$.
- Right-continuity: $N(t) = \lim_{s \downarrow t} N(s)$.

We also assume that $N(0) = 0$ and $N(t)$ increases by 1 each time.

9.15. Definition: Let $S_0 = 0$. Define

$$W_i = S_i - S_{i-1}, \quad i = 1, \ldots, n$$

to be the **interarrival time** between the $(i-1)$th event and the ith event.

9.16. Definition: A **renewal process** is a counting process for which the interarrival times $W_1, W_2, \ldots$ are iid.

9.17. Definition: A **(homogeneous) Poisson process** $\{N(t)\}_{t\geq 0}$ is a renewal process for which the interarrival times are exponentially distributed, i.e.,

$$\{W_i\}_{i=1}^{n} \overset{\text{iid}}{\sim} \text{Exp}(\lambda).$$

The parameter λ is called the **intensity** or **rate** of the process and we write

$$\{N(t)\}_{t\geq 0} \sim \text{Poi}(\lambda t).$$

Section 4. Basic Properties of Poisson Processes

9.18. Motivation: In this section, we will be looking at three basic properties of Poisson processes, which mainly result from the memoryless property of exponential random variables:

1. **continuous-time Markov property**: knowing the history (past time of occurrences) does not give you more information;
2. **independent increments**: the increments in counts over different time periods, $\{N(t_i) - N(t_{i-1})\}_{i\geq 1}$, are independent from each other;
3. **Poisson increments**: the increments in counts follow a Poisson distribution, i.e., $\{N(t_i) - N(t_{i-1})\}_{i\geq 1} \sim \text{Poi}(\lambda(t_i - t_{i-1}))$.

9.19. Theorem (Continuous-Time Markov Property)**:** *For any index $m \in \mathbb{Z}_+$, time $t_1 < \cdots < t_m$, and states $i_1, i_2, \ldots, i, j \in S$, the following condition holds:*

$$\begin{aligned}&\Pr(N(t_m) = j \mid N(t_{m-1}) = i, N(t_{m-2}) = i_{m-2}, \ldots, N(t_1) = i_1)\\ =&\Pr(N(t_m) = j \mid N(t_{m-1}) = i).\end{aligned}$$

Proof. We will give an informal proof. Suppose we stand at time t_{m-1} and we are given that $N(t_{m-1}) = i$. This only tells us that i events happened before time t_{m-1}, but no information about the actual time of occurrence is given. The "history", $N(t_{m-2}) = i_{m-2}, \ldots, N(t_1) = i_1$, does give us more information. In particular, we now know the number of occurrences during each time period $t_k - t_{k-1}$ for $k = 1, \ldots, m-1$.

The Markov property states that, given how many events occurred before, when the last event occurred has no influence on when the next event will occur. In other words, how long we have waited for the next event has no influence on how long we still need to wait. But this is exactly the *memoryless property* of the exponential distribution. Since the exponential distribution is the only continuous-time distribution which is memoryless, we also know that *the Poisson process is the only renewal process which has the Markov property.* □

9.20. Theorem (Independent Increments)**:** *Given times $t_1 < t_2 \leq t_3 < t_4$, we have*

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

Proof. This again follows from the memoryless property of the exponential distribution. □

9.21. Theorem (Poisson Increments)**:** *The increments in counts over time period $t_2 - t_1$ follows a Poisson distribution:*

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

In particular, for any $t \in \mathbb{R}_{\geq 0}$, the count follows a Poisson distribution:

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

Proof. By the memoryless property of exponential random variables, it suffices to prove one particular period satisfies this property. In other words, we wish to show that

$$N := N(t_2 - t_1) \sim \mathrm{Poi}(\lambda(t_2 - t_1)).$$

Now by definition,

$$\begin{aligned} N = n &\iff S_n \le t_2 - t_1 < S_{n+1} \\ &\iff (W_1 + \cdots + W_n \le t_2 - t_1) \wedge (W_1 + \cdots + W_n + W_{n+1} > t_2 - t_1) \end{aligned}$$

Sums of exponential random variables follow the Erlang distribution, so we have

$$\begin{aligned} \Pr(W_1 + \cdots + W_n \le t_2 - t_1) &= 1 - \sum_{k=1}^{n-1} \frac{1}{k!} e^{-\lambda(t_2-t_1)} (\lambda(t_2 - t_1))^k \\ \Pr(W_1 + \cdots + W_n + W_{n+1} \le t_2 - t_1) &= 1 - \sum_{k=1}^{n} \frac{1}{k!} e^{-\lambda(t_2-t_1)} (\lambda(t_2 - t_1))^k \end{aligned}$$

Combining them,

$$\begin{aligned} \Pr(N = n) &= \Pr(W_1 + \cdots + W_n \le t_2 - t_1) - \Pr(W_1 + \cdots + W_n + W_{n+1} \le t_2 - t_1) \\ &= \frac{1}{n!} e^{-\lambda(t_2-t_1)} (\lambda(t_2 - t_1))^n, \end{aligned}$$

which is exactly the pmf of $\mathrm{Poi}(\lambda(t_2 - t_1))$. □

9.22. Remark: As a result of the Poisson increment property, $N(1) = \mathrm{Poi}(\lambda)$ and $\mathbb{E}[N(1)] = \lambda$. In words, λ is the expected number of arrives/occurrence in one unit of time, and this is why it is called the **intensity** of the process.

9.23. Note: Note that the distribution of the increments, $\{N(t_i) - N(t_{i-1})\}_{i=1}^k$, together with the independence of these increments, uniquely determines the distribution of the process $\{N(t_i)\}_{i=1}^k$. This gives us the following alternative definition of Poisson processes, which is often easier to work with in proofs.

9.24. Definition: $\{N(t)\}_{t \ge 0}$ is a **Poisson process** if

- $N(t) = 0$;
- $\forall 0 \le s < t : N(t) - N(s) \sim \mathrm{Poi}(\lambda(t - s))$;
- $\forall t_0 < t_1 < \cdots < t_n : N(t_1) - N(t_0), N(t_2) - N(t_1), \ldots, N(t_n) - N(t_{n-1})$ are independent.

9.25. Remark: Using this definition, the original definition can be worded as a property: *Poisson processes are counting processes of events with iid exponential interarrival times.*

Section 5. Combining and Splitting Poisson Processes

9.26. Motivation: In this section, we look at how to *combine* two independent Poisson processes N_1, N_2 to obtain one Poisson process N, and how to *split* one Poisson process N into two independent Poisson processes N_1, N_2. Note that when we are only interested in N_1, *splitting* is also called *thinning*.

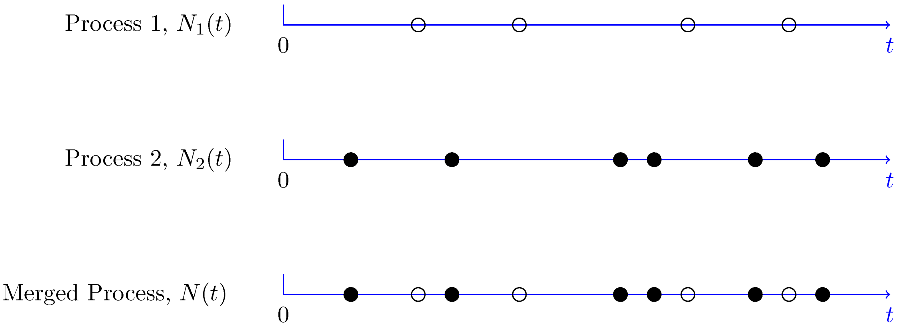


Figure 9.1: Merging two Poisson processes.

9.27. Theorem: *Let $\{N_i(t)\}_{t\geq 0}$, $i = 1, 2$ be two independent Poisson processes with intensities λ_1, λ_2, respectively. Then $N(t) = N_1(t) + N_2(t)$ is a Poisson process with intensity $\lambda = \lambda_1 + \lambda_2$.*

Proof. We check the alternative definition.

- $N(0) = N_1(0) + N_2(0) = 0$. Good.
- For $0 \leq s < t$,

$$N(t) - N(s) = \underbrace{N_1(t) - N_1(s)}_{\sim \mathrm{Poi}(\lambda_1(t-s))} + \underbrace{N_2(t) - N_2(s)}_{\sim \mathrm{Poi}(\lambda_2(t-s))}.$$

  Since these two random variables are independent, adding two Poisson random variables gives us $N(t) - N(s) \sim \mathrm{Poi}((\lambda_1 + \lambda_2)(t - s))$. Good.
- The random variables $\{N(t_i) - N(t_{i-1}) = N_1(t_i) - N_1(t_{i-1}) + N_2(t_i) - N_2(t_{i-1})\}_{i=1}^n$ are jointly independent since $N_1(t_i) - N_1(t_{i-1})$'s and $N_2(t_i) - N_2(t_{i-1})$'s are independent for all i and the N_1 terms are independent of N_2 terms.

The result follows. □

9.28. Corollary: *Let $\{N_i(t)\}_{t\geq 0}$, $i = 1, \ldots, k$ be k independent Poisson processes with intensities $\lambda_1, \ldots, \lambda_n$, then $N(t) = \sum_{i=1}^k N_i(t)$ is a Poisson process with intensity $\sum_{i=1}^k \lambda_i$.*

Proof. Easy induction. □

9.29. Note: We now look at splitting Poisson processes. Consider a Poisson process with intensity λ as the counting process of the events with iid exponential interarrival times. For each event, mark it with a 1 with probability P, with 2 with probability $1-p$. By construction, the marking of different events are independent.

9.30. (Cont'd): Let N_1 and N_2 be the counting processes of the events with marks 1 and 2, respectively. Then $\{N_1(t)\}$ and $\{N_2(t)\}$ are independent Poisson processes with intensities $p\lambda$ and $(1-p)\lambda$, respectively. Note this is the inverse procedure of combining two independent Poisson processes into one Poisson process.

Proof. Again, we check the alternative definition.

- $N_1(0) = N_2(0) = 0$. Good.
- Since $N(t) - N(s) = N(t-s) \sim \text{Poi}(\lambda(t-s))$, and the splitting rule does not change over time, it suffices to consider the case where $s = 0$. Consider the joint distribution:

$$\begin{aligned}
&\Pr(N_1(t) = m, N_2(t) = n) \\
&= \Pr(N_1(t) = m, N_2(t) = n \mid N(t) = m+n) \cdot \Pr(N(t) = m+n) \\
&= \binom{m+n}{m} p^m (1-p)^n \cdot e^{-\lambda t} \frac{(\lambda t)^{m+n}}{(m+n)!} \\
&= \frac{(m+n)!}{m!n!} (p\lambda t)^m ((1-p)\lambda t)^n e^{-p\lambda t} \cdot e^{-(1-p)\lambda t} \frac{1}{(m+n)!} \\
&= e^{-p\lambda t} \frac{(p\lambda t)^m}{m!} \cdot e^{-(1-p)\lambda t} \frac{(1-p)\lambda t)^n}{n!}
\end{aligned}$$

  Note that the left term $\sim \text{Poi}(p\lambda t)$ at m and the right term $\sim \text{Poi}((1-p)\lambda t)$ at n. This implies that

  1. $N_1(t) \perp\!\!\!\perp N_2(t)$.
  2. $N_1(t) \sim \text{Poi}(p\lambda t)$, $N_2(t) \sim \text{Poi}((1-p)\lambda t)$.

  The second is the Poisson increment property that we want.
- For independence increments, since $N(t)$ has independent increments and the marking is independent, $N_1(t)$ and $N_2(t)$ also have independent increments. Thus, $\{N_1(t)\}$ and $\{N_2(t)\}$ are Poisson processes with intensities $p\lambda$ and $(1-p)\lambda$, respectively. Note that $N_1(t) \perp\!\!\!\perp N_2(t)$, $t \geq 0$, is not yet enough for two processes N_1 and N_2 to be independent. Instead, we need

$$(N_1(t_0), N_1(t_1), \ldots, N_1(t_n)) \perp\!\!\!\perp (N_2(t_0), N_2(t_1), \ldots, N_2(t_n))$$

  for all n and $t_0 < t_1 < \cdots < t_n$. This follows from the independent increment property.

□

Section 6. Order Statistics Property

9.31. Definition: Let $X_1, \dots, X_n$ be (typically iid) random variables. The order statistics of $\{X_1, \dots, X_n\}$ is a permutation of $\{X_1, \dots, X_n\}$ arranged in *non-decreasing* order. In particular, $X_{(1)} = \min\{X_1, \dots, X_n\}$ and $X_{(n)} = \max\{X_1, \dots, X_n\}$.

9.32. Motivation: Conditioned on $N(t) = n$, the occurrence times before t are distributed as the order statistics of n iid Uniform$(0,t)$ random variables.

9.33. Theorem: *Let $\{N(t)\}_{t \geq 0}$ be a Poisson process with with intensity λ. Conditioned on $N(t) = n$, the occurrence times of the events in time period $[0,t]$ are distributed as the order statistics of n iid uniformly distributed random variables on $[0,t]$. That is,*

$$(S_1, \dots, S_n \mid N(t) = n) \overset{d}{=} (U_{(1)}, \dots, U_{(n)}),$$

where S_i denotes the time of occurrence of the ith event, $\{U_i\}_{i=1}^n \sim$ Uniform$(0,t)$, and $\{U_{(i)}\}_{i=1}^n$ are the order statistics of $\{U_i\}_{i=1}^n$.

Proof. Consider n intervals $\{[a_i, b_i]\}_{i=1}^n$ with $0 \leq a_1 < b_1 < a_2 < b_2 < \cdots < a_n < b_n < t$.

$$\begin{aligned}
&\Pr(S_i \in (a_i, b_i], i = 1, \dots, n \mid N(t) = n) \\
&= \frac{\Pr(S_i \in (a_i, b_i], i = 1, \dots, n, N(t) = n)}{\Pr(N(t) = n)} \\
&= \frac{\Pr(N(a_1) = 0, N(b_1) - N(a_1) = 1, N(a_2) - N(b_1) = 0, N(b_2) - N(a_2) = 1, \dots, N(t) - N(b_n) = 0)}{\Pr(N(t) = n)}
\end{aligned}$$

Now $N(t_p) - N(t_q) \sim \text{Poi}(\lambda(t_q - t_p))$ and all terms are independent, so the numerator can be written as

$$\begin{aligned}
&(e^{-\lambda a_1}) \cdot (\lambda(b_1 - a_1) \cdot e^{-\lambda(b_1 - a_1)}) \cdots (e^{-\lambda(t - b_n)}) \\
&= \exp\{-\lambda(a_1 + (b_1 - a_1) + (a_2 - b_1) + \cdots + (t - b_n))\} \cdot \lambda^n \prod_{i=1}^n (b_i - a_i) \\
&= e^{-\lambda t} \cdot \lambda^n \prod_{i=1}^n (b_i - a_i)
\end{aligned}$$

Next, we know that $N(t) \sim \text{Poi}(\lambda t)$. Combine them, we have

$$\Pr(S_i \in (a_i, b_i], i = 1, \dots, n \mid N(t) = n) = \frac{e^{-\lambda t} \cdot \lambda^n \prod_{i=1}^n (b_i - a_i)}{e^{-\lambda t}} \frac{(\lambda t)^n}{n!} = \frac{n!}{t_n} \prod_{i=1}^n (b_i - a_i).$$

Divide both sides by $\prod_{i=1}^n (b_i - a_i)$ and take limits $b_i \downarrow a_i$, we obtain the conditional pdf

$$f_{S_1, \dots, S_n | N(t) = n}(a_1, \dots, a_n) = \frac{n!}{t^n} \mathbf{1}_{\{a_1 < a_2 < \cdots < a_n\}}.$$

This is also the pdf of $(U_{(1)}, \dots, U_{(n)})$. □

9.34. Note: We now have two methods of simulating the occurrence time of the events of a Poisson process until time t:

Method 1. Simulate iid Exp(λ) until the sum exceeds t.

Method 2. Simulate a Poi(λt), denote it as N (number of events occurred).
Now generate N iid Uniform$(0, t)$ random variables and sort them.

9.35. Corollary: *Let $s \le t$. Then*

$$(N(s) \mid N(t) = n) \sim \text{Binomial}\left(n, \frac{s}{t}\right).$$

Proof. By definition,

$$\begin{aligned}
\Pr(N(s) = k \mid N(t) = n) &= \Pr(S_1, \dots, S_k \le s; S_{k+1}, \dots, S_n > s \mid N(t) = n) \\
&= \Pr(U_{(1)}, \dots, U_{(k)} \le s; U_{(k+1)}, \dots, U_{(n)} > s) \\
&= \Pr(k \text{ out of } n \text{ iid Uniform}(0, t) \text{ are smaller than } s) \\
&= \binom{n}{k} \left(\frac{s}{t}\right)^k \left(1 - \frac{s}{t}\right)^{n-k}
\end{aligned}$$

which is the pmf of a binomial random variable with parameter n and $p = s/t$. □

Section 7. Nonhomogeneous Poisson Process

9.36. Definition: $\{N(t)\}_{t\geq 0}$ is a **non-homogeneous Poisson process** with rate $\lambda(t)$ if

1. $N(0) = 0$.
2. $\forall 0 \leq s \leq t : N(t) - N(s) \sim \text{Poi}(\int_s^t \lambda(r)\,dr)$.
3. $N(t)$ has independent increments, i.e., $\{N(t_i) - N(t_{i-1})\}_{i=1}^n$ are independent.

9.37. Remark: We see that this definition is a generalization of the alternative definition for the homogeneous Poisson process. The homogeneous Poisson process can be viewed as a special case where $\lambda(t)$ is constant for all t.

9.38. Note: For more intuition, consider the following quantity.

$$\begin{aligned}
&\text{Pr(there exists at least one event in a small interval } [t, t+\Delta t])\\
&= \Pr(N(t+\Delta t) - N(t) \geq 1)\\
&= 1 - \Pr(N(t+\Delta t) - N(t) = 0)\\
&= 1 - \exp\left(-\int_t^{t+\Delta t} \lambda r\,dr\right) \qquad\qquad N(t) + \Delta t - N(t) \sim \text{Poi}\left(\int_t^{t+\Delta t} \lambda r\,dr\right)
\end{aligned}$$

As $\Delta t \to 0$,

$$\Delta t \to 0 \implies \int_t^{t+\Delta t} \lambda(r)\,dr \to 0.$$

Since this number is small, we use Taylor expansion of $e^x = 1 + x + C$ at 0, which gives us

$$\exp\left(-\int_t^{t+\Delta t} \lambda r\,dr\right) = 1 - \int_t^{t+\Delta t} \lambda(r)\,dr + C,$$

where C represents the sum of the subsequent terms. Plug this in,

$$\Pr(N(t+\Delta t) - N(t) \geq 1) = \int_t^{t+\Delta t} \lambda(r)\,dr + C \approx \lambda(t)\Delta t$$

where we assumed that Δt is small and $\lambda(r)$ is continuous. We can interpret $\lambda(t)$ as the **attractiveness** of time t, i.e., some t^* are more attractive (more likely for events to happen) compared to other times. Homogeneous Poisson processes basically assumed that all times are *equally-attractive*, where non-homogeneous Poisson processes assign different attractiveness to different times.

9.39. Note: What properties are still valid for non-homogeneous Poisson processes?

- It is still a counting process.
- However, interarrival times are no longer iid exponential.
- Therefore, non-homogeneous Poisson processes are no longer renewal processes.
- Markov property still holds, i.e., history can be ignored given the current state.

- Independent increments and Poisson increments still hold by definition.
- Combining and splitting still work. More precisely,
  - $\lambda_1(r), \lambda_2(r) \implies \lambda(r) = \lambda_1(r) + \lambda_2(r)$.
  - $\lambda(r)$ with $p(r)$ and $1 - p(r) \implies \lambda_1(r) = \lambda(r)p(r)$ and $\lambda_2(r) = \lambda(r)(1 - p(r))$.
- Order statistics property still hold. No longer uniform, but iid with density

$$f(s) = \frac{\lambda(s)}{\int_0^t \lambda(s)\, ds}, \quad s \in [0, t].$$

Section 8. Compound Poisson Processes

9.40. Motivation: Suppose each arrival/occurrence is associated with a quantity which are assumed to be iid. We are interested in the total quantity up to time t:

$$S(t) = Y_1 + \cdots + Y_{N(t)}, \quad Y_i's \text{ iid.}$$

For example, consider the claims arrive at an insurance company. The number of claims can be modelled by a Poisson process and the total amount of claims can be modeled by a compound Poisson process. Let us first consider the mean and variance of $S(t)$.

9.41. Theorem: *Let $Y_1, Y_2, \ldots$ be iid random variables and N be a non-negative integer-valued random variable, independent of the Y_i's. Define $S = Y_1 + \cdots + Y_N$. Then*

- *If $\mathbb{E}[Y_i] = \mu$ and $\mathbb{E}[N] < \infty$, then $\mathbb{E}[S] = \mu \cdot \mathbb{E}[N]$.*
- *If $\mathrm{Var}[Y_i] = \sigma^2$ and $\mathrm{Var}[N] < \infty$, then $\mathrm{Var}[S] = \sigma^2\mathbb{E}[N] + \mu^2\mathrm{Var}[N]$.*
  In particular, if $N \sim \mathrm{Poi}(\lambda)$, then $\mathrm{Var}[S] = \lambda\mathbb{E}[Y_i^2] = \lambda(\mathbb{E}^2[Y_i] + \mathrm{Var}[Y_i])$.

Proof. Statement 1 is simply Wald's identity. For statement 2, we use the Law of Total Variance, which states that for two random variable X and Y,

$$\mathrm{Var}[X] = \mathbb{E}[\mathrm{Var}[X \mid Y]] + \mathrm{Var}[\mathbb{E}[X \mid Y]]$$

where $\mathrm{Var}[X \mid Y] = \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2 \mid Y]$. We are interested in $\mathrm{Var}(S) = \mathrm{Var}(Y_1 + \cdots + Y_N)$. The conditional mean and variance of S is

$$\mathbb{E}\left[\sum_{i=1}^{N} Y_i \,\middle|\, N\right] = N \cdot \mathbb{E}[Y_i] = \mu N$$
$$\mathrm{Var}\left[\sum_{i=1}^{N} Y_i \,\middle|\, N\right] = N \cdot \mathrm{Var}[Y_i] = \sigma^2 N.$$

Note the second statement holds as

$$\mathrm{Var}\left[\sum_{i=1}^{N} Y_i \,\middle|\, N = n\right] = \mathrm{Var}\left[\sum_{i=1}^{n} Y_i\right] = n\mathrm{Var}[Y_i] = n\sigma^2.$$

Now use Law of Total Variance, we have

$$\begin{aligned}\mathrm{Var}[S] &= \mathrm{Var}\left(\mathbb{E}\left[\sum_{i=1}^{N} Y_i \,\middle|\, N\right]\right) + \mathbb{E}\left[\mathrm{Var}\left(\sum_{i=1}^{N} Y_i \,\middle|\, N\right)\right] \\ &= \mathrm{Var}[\mu N] + \mathbb{E}[\sigma^2 N] = \mu^2\mathrm{Var}[n] + \sigma^2\mathbb{E}[N].\end{aligned}$$

In particular, when $N \sim \mathrm{Poi}(\lambda)$, $\mathbb{E}[N] = \mathrm{Var}[N] = \lambda$ so we get

$$\mathrm{Var}[S] = \mu^2\lambda + \sigma^2\lambda = \lambda(\mathbb{E}^2[Y_i] + \mathrm{Var}[Y_i]) = \lambda\mathbb{E}[Y_i^2]$$

as desired. □