Stochastic Manufacturing & Service Systems

Jim Dai and Hyunwoo Park

School of Industrial and Systems Engineering
Georgia Institute of Technology

October 19, 2011

Contents

Chapter 1

Newsvendor Problem

In this course, we will learn how to design, analyze, and manage a manufacturing or service system with uncertainty. Our first step is to understand how to solve a single period decision problem containing uncertainty or randomness.

1.1 Profit Maximization

We will start with the simplest case: selling perishable items. Suppose we are running a business retailing newspaper to Georgia Tech campus. We have to order a specific number of copies from the publisher every evening and sell those copies the next day. One day, if there is a big news, the number of GT people who want to buy and read a paper from you may be very high. Another day, people may just not be interested in reading a paper at all. Hence, you as a retailer, will encounter the demand variability and it is the primary uncertainty you need to handle to keep your business sustainable. To do that, you want to know what is the optimal number of copies you need to order every day. By intuition, you know that there will be a few other factors than demand you need to consider.

- **Selling price (p):** How much will you charge per paper?
- **Buying price (c_v):** How much will the publisher charge per paper? This is a variable cost, meaning that this cost is proportional to how many you order. That is why it is denoted by c_v.
- **Fixed ordering price (c_f):** How much should you pay just to place an order? Ordering cost is fixed regardless of how many you order.
- **Salvage value (s) or holding cost (h):** There are two cases about the leftover items. They could carry some monetary value even if expired. Otherwise, you have to pay to get rid of them or to storing them. If they have some value, it is called salvage value. If you have to pay, it is called

holding cost. Hence, the following relationship holds: $s = -h$. This is per-item value.

- **Backorder cost (b):** Whenever the actual demand is higher than how many you prepared, you lose sales. Loss-of-sales could cost you something. You may be bookkeeping those as backorders or your brand may be damaged. These costs will be represented by backorder cost. This is per-item cost.

- **Your order quantity (y):** You will decide how many papers to be ordered before you start a day. That quantity is represented by y. This is your decision variable.

As a business, you are assumed to want to maximize your profit. Expressing your profit as a function of these variables is the first step to obtain the optimal ordering policy. Profit can be interpreted in two ways: (1) revenue minus cost, or (2) money you earn minus money you lose.

Let us adopt the first interpretation first. Revenue is represented by selling price (p) multiplied by how many you actually sell. The actual sales is bounded by the realized demand and how many you prepared for the period. When you order too many, you can sell at most as many as the number of people who want to buy. When you order too few, you can only sell what you prepared. Hence, your revenue is minimum of D and y, i.e. $\min(D, y)$ or $D \wedge y$. Thinking about the cost, first of all, you have to pay something to the publisher when buying papers, i.e. $c_f + yc_v$. Two types of additional cost will be incurred to you depending on whether your order is above or below the actual demand. When it turns out you prepared less than the demand for the period, the backorder cost b per every missed sale will occur. The amount of missed sales cannot be negative, so it can be represented by $\max(D - y, 0)$ or $(D - y)^+$. When it turns out you prepared more, the quantity of left-over items also cannot go negative, so it can be expressed as $\max(y - D, 0)$ or $(y - D)^+$. In this way of thinking, we have the following formula.

$$\begin{aligned}\text{Profit} =& \text{Revenue} - \text{Cost} \\ =& \text{Revenue} - \text{Ordering cost} - \text{Holding cost} - \text{Backorder cost} \\ =& p(D \wedge y) - (c_f + yc_v) - h(y - D)^+ - b(D - y)^+ \end{aligned} \tag{1.1}$$

How about the second interpretation of profit? You earn $p - c_v$ dollars every time you sell a paper. For left-over items, you lose the price you bought in addition to the holding cost per paper, i.e. $c_v + h$. When the demand is higher than what you prepared, you lose b backorder cost. Of course, you also have to pay the fixed ordering cost c_f as well when you place an order. With this logic, we have the following profit function.

$$\begin{aligned}\text{Profit} =& \text{Earning} - \text{Loss} \\ =& (p - c_v)(D \wedge y) - (c_v + h)(y - D)^+ - b(D - y)^+ - c_f \end{aligned} \tag{1.2}$$

Since we used two different approaches to model the same profit function, (1.1) and (1.2) should be equivalent. Comparing the two equations, you will also notice that $(D \wedge y) + (y - D)^+ = y$.

Now our quest boils down to maximizing the profit function. However, (1.1) and (1.2) contain a random element, the demand D. We cannot maximize a function of random element if we allow the randomness to remain in our objective function. One day demand can be very high. Another day it is also possible nobody wants to buy a single paper. We have to figure out how to get rid of this randomness from our objective function. Let us denote profit for the nth period by g_n for further discussion.

Theorem 1.1 (Strong Law of Large Numbers)**.**

$$\mathbf{Pr}\left\{\lim_{n\to\infty}\frac{g_1+g_2+g_3+\cdots+g_n}{n}=\mathbf{E}[g_1]\right\}=1$$

The long-run average profit converges to the expected profit for a single period with probability 1.

Based on Theorem 1.1, we can change our objective function from just profit to expected profit. In other words, by maximizing the expected profit, it is guaranteed that the long-run average profit is maximized because of Theorem 1.1. Theorem 1.1 is the foundational assumption for the entire course. When we will talk about the long-run average something, it involves Theorem 1.1 in most cases. Taking expectations, we obtain the following equations corresponding to (1.1) and (1.2).

$$\begin{aligned}\mathbf{E}[g(D,y)] =& p\mathbf{E}[D\wedge y]-(c_f+yc_v)-h\mathbf{E}[(y-D)^+]-b\mathbf{E}[(D-y)^+] &(1.3)\\ =&(p-c_v)\mathbf{E}[D\wedge y]\\ &-(c_v+h)\mathbf{E}[(y-D)^+]-b\mathbf{E}[(D-y)^+]-c_f &(1.4)\end{aligned}$$

Since (1.3) and (1.4) are equivalent, we can choose either one of them for further discussion and (1.4) will be used.

Before moving on, it is important for you to understand what $\mathbf{E}[D\wedge y], \mathbf{E}[(y-D)^+], \mathbf{E}[(D-y)^+]$ are and how to compute them.

Example 1.1. Compute $\mathbf{E}[D\wedge 18], \mathbf{E}[(18-D)^+], \mathbf{E}[(D-18)^+]$ for the demand having the following distributions.

1. D is a discrete random variable. Probability mass function (pmf) is as follows.

d	10	15	20	25	30
$\mathbf{Pr}\{D=d\}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$

Answer: For a discrete random variable, you first compute $D\wedge 18, (18-D)^+, (D-18)^+$ for each of possible D values.

d	10	15	20	25	30
$\mathbf{Pr}\{D=d\}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$
$D \wedge 18$	10	15	18	18	18
$(18-D)^+$	8	3	0	0	0
$(D-18)^+$	0	0	2	7	12

Then, you take the weighted average using corresponding $\mathbf{Pr}\{D=d\}$ for each possible D.

$$\mathbf{E}[D \wedge 18] = \frac{1}{4}(10) + \frac{1}{8}(15) + \frac{1}{8}(18) + \frac{1}{4}(18) + \frac{1}{4}(18) = \frac{125}{8}$$
$$\mathbf{E}[(18-D)^+] = \frac{1}{4}(8) + \frac{1}{8}(3) + \frac{1}{8}(0) + \frac{1}{4}(0) + \frac{1}{4}(0) = \frac{19}{8}$$
$$\mathbf{E}[(D-18)^+] = \frac{1}{4}(0) + \frac{1}{8}(0) + \frac{1}{8}(2) + \frac{1}{4}(7) + \frac{1}{4}(12) = 5$$

2. D is a continuous random variable following uniform distribution between 10 and 30, i.e. $D \sim \text{Uniform}(10, 30)$.

   Answer: Computing expectation of continuous random variable involves integration. A continuous random variable has probability density function usually denoted by f. This will be also needed to compute the expectation. In this case,

$$f_D(x) = \begin{cases} \frac{1}{20}, & \text{if } x \in [10, 30] \\ 0, & \text{otherwise} \end{cases}$$

   Using this information, compute the expectations directly by integration.

$$\begin{aligned}
\mathbf{E}[D \wedge 18] &= \int_{-\infty}^{\infty} (x \wedge 18) f_D(x) dx \\
&= \int_{10}^{30} (x \wedge 18) \frac{1}{20} dx \\
&= \int_{10}^{18} (x \wedge 18) \left(\frac{1}{20}\right) dx + \int_{18}^{30} (x \wedge 18) \left(\frac{1}{20}\right) dx \\
&= \int_{10}^{18} x \left(\frac{1}{20}\right) dx + \int_{18}^{30} 18 \left(\frac{1}{20}\right) dx \\
&= \frac{x^2}{40}\Big|_{x=10}^{x=18} + \frac{18x}{20}\Big|_{x=18}^{x=30}
\end{aligned}$$

   The key idea is to remove the $\wedge$ operator that we cannot handle by separating the integration interval into two. The other two expectations can

be computed in a similar way.

$$\begin{aligned}
\mathbf{E}[(18-D)^+] &= \int_{-\infty}^{\infty} (18-x)^+ f_D(x)dx \\
&= \int_{10}^{30} (18-x)^+ \frac{1}{20} dx \\
&= \int_{10}^{18} (18-x)^+ \left(\frac{1}{20}\right) dx + \int_{18}^{30} (18-x)^+ \left(\frac{1}{20}\right) dx \\
&= \int_{10}^{18} (18-x) \left(\frac{1}{20}\right) dx + \int_{18}^{30} 0 \left(\frac{1}{20}\right) dx \\
&= \left. \frac{18x - \frac{x^2}{2}}{20} \right|_{x=10}^{x=18} + 0
\end{aligned}$$

$$\begin{aligned}
\mathbf{E}[(D-18)^+] &= \int_{-\infty}^{\infty} (18-x)^+ f_D(x)dx \\
&= \int_{10}^{30} (x-18)^+ \frac{1}{20} dx \\
&= \int_{10}^{18} (x-18)^+ \left(\frac{1}{20}\right) dx + \int_{18}^{30} (x-18)^+ \left(\frac{1}{20}\right) dx \\
&= \int_{10}^{18} 0 \left(\frac{1}{20}\right) dx + \int_{18}^{30} (x-18) \left(\frac{1}{20}\right) dx \\
&= 0 + \left. \frac{\frac{x^2}{2} - 18x}{20} \right|_{x=18}^{x=30}
\end{aligned}$$

Now that we have learned how to compute $\mathbf{E}[D \wedge y], \mathbf{E}[(y-D)^+], \mathbf{E}[(D-y)^+]$, we have acquired the basic toolkit to obtain the order quantity that maximizes the expected profit. First of all, we need to turn these expectations of the profit function formula (1.4) into integration forms. For now, assume that the demand is a nonnegative continuous random variable.

$$\begin{aligned}
\mathbf{E}[g(D,y)] =&(p-c_v)\mathbf{E}[D\wedge y]-(c_v+h)\mathbf{E}[(y-D)^+]-b\mathbf{E}[(D-y)^+]-c_f\\
=&(p-c_v)\int_0^\infty (x\wedge y)f_D(x)dx\\
&-(c_v+h)\int_0^\infty (y-x)^+f_D(x)dx\\
&-b\int_0^\infty (x-y)^+f_D(x)dx-c_f\\
=&(p-c_v)\left(\int_0^y xf_D(x)dx+\int_y^\infty yf_D(x)dx\right)\\
&-(c_v+h)\int_0^y (y-x)f_D(x)dx\\
&-b\int_y^\infty (x-y)f_D(x)dx-c_f\\
=&(p-c_v)\left(\int_0^y xf_D(x)dx+y\left(1-\int_0^y f_D(x)dx\right)\right)\\
&-(c_v+h)\left(y\int_0^y f_D(x)dx-\int_0^y xf_D(x)dx\right)\\
&-b\left(\mathbf{E}[D]-\int_0^y xf_D(x)dx-y\left(1-\int_0^y f_D(x)dx\right)\right)-c_f \quad (1.5)
\end{aligned}$$

There can be many ways to obtain the maximum point of a function. Here we will take the derivative of (1.5) and set it to zero. y that makes the derivative equal to zero will make $\mathbf{E}[g(D,y)]$ either maximized or minimized depending on the second derivative. For now, assume that such y will maximize $\mathbf{E}[g(D,y)]$. We will check this later. Taking the derivative of (1.5) will involve differentiating an integral. Let us review an important result from Calculus.

Theorem 1.2 (Fundamental Theorem of Calculus)**.** *For a function*

$$H(y)=\int_c^y h(x)dx,$$

we have $H'(y)=h(y)$, where c is a constant.

Theorem 1.2 can be translated as follows for our case.

$$\frac{d}{dy}\left(\int_0^y xf_D(x)dx\right)=yf_D(y) \quad (1.6)$$

$$\frac{d}{dy}\left(\int_0^y f_D(x)dx\right)=f_D(y) \quad (1.7)$$

Also remember the relationship between cdf and pdf of a continuous random variable.

$$F_D(y)=\int_{-\infty}^y f_D(x)dx \quad (1.8)$$

Use (1.6), (1.7), (1.8) to take the derivative of (1.5).

$$\begin{aligned}\frac{d}{dy}\mathbf{E}[g(D,y)] =&(p-c_v)\left(yf_D(y)+1-F_D(y)-yf_D(y)\right)\\ &-(c_v+h)\left(F_D(y)+yf_D(y)-yf_D(y)\right)\\ &-b\left(-yf_D(y)-1+F_D(y)+yf_D(y)\right)\\ =&(p+b-c_v)(1-F_D(y))-(c_v+h)F_D(y)\\ =&(p+b-c_v)-(p+b+h)F_D(y)=0 \qquad (1.9)\end{aligned}$$

If we differentiate (1.9) one more time to obtain the second derivative,

$$\frac{d^2}{dy^2}\mathbf{E}[g(D,y)] = -(p+b+h)f_D(y)$$

which is always nonpositive because $p, b, h, f_D(y) \geq 0$. Hence, taking the derivative and setting it to zero will give us the maximum point not the minimum point. Therefore, we obtain the following result.

Theorem 1.3 (Optimal Order Quantity)**.** *The optimal order quantity y^* is the smallest y such that*

$$F_D(y) = \frac{p+b-c_v}{p+b+h} \text{ or } y = F_D^{-1}\left(\frac{p+b-c_v}{p+b+h}\right).$$

for continuous demand D.

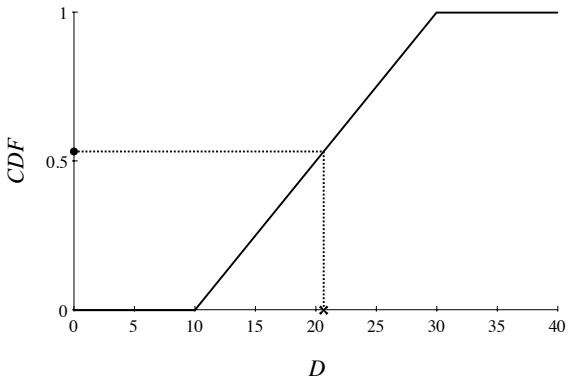
Looking at Theorem 1.3, it provides the following intuitions.

- Fixed cost c_f does not affect the optimal quantity you need to order.
- If you can procure items for free and there is no holding cost, you will prepare as many as you can.
- If $b \gg h, b \gg c_v$, you will also prepare as many as you can.
- If the buying cost is almost as same as the selling price plus backorder cost, i.e. $c_v \approx p+b$, you will prepare nothing. You will prepare only upon you receive an order.

Example 1.2. Suppose $p = 10, c_f = 100, c_v = 5, h = 2, b = 3, D \sim \text{Uniform}(10, 30)$. How many should you order for every period to maximize your long-run average profit?
Answer: First of all, we need to compute the criterion value.

$$\frac{p+b-c_v}{p+b+h} = \frac{10+3-5}{10+3+2} = \frac{8}{15}$$

Then, we will look up the smallest y value that makes $F_D(y) = 8/15$.



Therefore, we can conclude that the optimal order quantity

$$y^* = 10 + 20\frac{8}{15} = \frac{62}{3} \text{ units.}$$

Although we derived the optimal order quantity solution for the continuous demand case, Theorem 1.3 applies to the discrete demand case as well. I will fill in the derivation for discrete case later.

Example 1.3. Suppose $p = 10, c_f = 100, c_v = 5, h = 2, b = 3$. Now, D is a discrete random variable having the following pmf.

d	10	15	20	25	30
$\mathbf{Pr}\{D = d\}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$

What is the optimal order quantity for every period?
Answer: We will use the same value 8/15 from the previous example and look up the smallest y that makes $F_D(y) = 8/15$. We start with $y = 10$.

$$\begin{aligned}
F_D(10) &= \frac{1}{4} &&< \frac{8}{15} \\
F_D(15) &= \frac{1}{4} + \frac{1}{8} = \frac{3}{8} &&< \frac{8}{15} \\
F_D(20) &= \frac{1}{4} + \frac{1}{8} + \frac{1}{8} = \frac{1}{2} &&< \frac{8}{15} \\
F_D(25) &= \frac{1}{4} + \frac{1}{8} + \frac{1}{8} + \frac{1}{4} = \frac{3}{4} &&\geq \frac{8}{15}
\end{aligned}$$

Hence, the optimal order quantity $y^* = 25$ units.

1.2 Cost Minimization

Suppose you are a production manager of a large company in charge of operating manufacturing lines. You are expected to run the factory to minimize the cost. Revenue is another person's responsibility, so all you care is the cost. To model the cost of factory operation, let us set up variables in a slightly different way.

- **Understock cost (c_u):** It occurs when your production is not sufficient to meet the market demand.
- **Overstock cost (c_o):** It occurs when you produce more than the market demand. In this case, you may have to rent a space to store the excess items.
- **Unit production cost (c_v):** It is the cost you should pay whenever you manufacture one unit of products. Material cost is one of this category.
- **Fixed operating cost (c_f):** It is the cost you should pay whenever you decide to start running the factory.

As in the profit maximization case, the formula for cost expressed in terms of c_u, c_o, c_v, c_f should be developed. Given random demand D, we have the following equation.

$$\begin{aligned}\text{Cost} =&\text{Manufacturing Cost}\\ &+\text{Cost associated with Understock Risk}\\ &+\text{Cost associated with Overstock Risk}\\ =&(c_f + yc_v) + c_u(D-y)^+ + c_o(y-D)^+ \end{aligned} \tag{1.10}$$

(1.10) obviously also contains randomness from D. We cannot minimize a random objective itself. Instead, based on Theorem 1.1, we will minimize expected cost then the long-run average cost will be also guaranteed to be minimized. Hence, (1.10) will be transformed into the following.

$$\begin{aligned}\mathbf{E}[\text{Cost}] =&(c_f + yc_v) + c_u\mathbf{E}[(D-y)^+] + c_o\mathbf{E}[(y-D)^+]\\ =&(c_f + yc_v) + c_u\int_0^\infty (x-y)^+ f_D(x)dx + c_o\int_0^\infty (y-x)^+ f_D(x)dx\\ =&(c_f + yc_v) + c_u\int_y^\infty (x-y) f_D(x)dx + c_o\int_0^y (y-x) f_D(x)dx \end{aligned} \tag{1.11}$$

Again, we will take the derivative of (1.11) and set it to zero to obtain y that makes $\mathbf{E}[\text{Cost}]$ minimized. We will verify the second derivative is positive in this case. Let g here denote the cost function and use Theorem 1.2 to take the derivative of integrals.

$$\begin{aligned}\frac{d}{dy}\mathbf{E}[g(D,y)] =&c_v + c_u(-yf_D(y) - 1 + F_D(y) + yf_D(y))\\ &+ c_o(F_D(y) + yf_D(y) - yf_D(y))\\ =&c_v + c_u(F_D(y) - 1) + c_oF_D(y) \end{aligned} \tag{1.12}$$

The optimal production quantity y^* is obtained by setting (1.12) to be zero.

Theorem 1.4 (Optimal Production Quantity)**.** *The optimal production quantity that minimizes the long-run average cost is the smallest y such that*

$$F_D(y) = \frac{c_u - c_v}{c_u + c_o} \text{ or } y = F^{-1}\left(\frac{c_u - c_v}{c_u + c_o}\right).$$

Theorem 1.4 can be also applied to discrete demand. Several intuitions can be obtained from Theorem 1.4.

- Fixed cost (c_f) again does not affect the optimal production quantity.
- If understock cost (c_u) is equal to unit production cost (c_v), which makes $c_u - c_v = 0$, then you will not produce anything.
- If unit production cost and overstock cost are negligible compared to understock cost, meaning $c_u \gg c_v, c_o$, you will prepare as much as you can.

To verify the second derivative of (1.11) is indeed positive, take the derivative of (1.12).

$$\frac{d^2}{dy^2}\mathbf{E}[g(D,y)] = (c_u + c_o)f_D(y) \tag{1.13}$$

(1.13) is always nonnegative because $c_u, c_o \geq 0$. Hence, y^* obtained from Theorem 1.4 minimizes the cost instead of maximizing it.

Before moving on, let us compare criteria from Theorem 1.3 and Theorem 1.4.

$$\frac{p+b-c_v}{p+b+h} \quad \text{and} \quad \frac{c_u - c_v}{c_u + c_o}$$

Since the profit maximization problem solved previously and the cost minimization problem solved now share the same logic, these two criteria should be somewhat equivalent. We can see the connection by matching $c_u = p + b, c_o = h$. In the profit maximization problem, whenever you lose a sale due to underpreparation, it costs you the opportunity cost which is the selling price of an item and the backorder cost. Hence, $c_u = p + b$ makes sense. When you overprepare, you should pay the holding cost for each left-over item, so $c_o = h$ also makes sense. In sum, Theorem 1.3 and Theorem 1.4 are indeed the same result in different forms.

Example 1.4. Suppose demand follows Poisson distribution with parameter 3. The cost parameters are $c_u = 10, c_v = 5, c_o = 15$. Note that $e^{-3} \approx 0.0498$.
Answer: The criterion value is

$$\frac{c_u - c_v}{c_u + c_o} = \frac{10-5}{10+15} = 0.2,$$

so we need to find the smallest y such that makes $F_D(y) \geq 0.2$. Compute the probability of possible demands.

$$\begin{aligned}
\mathbf{Pr}\{D=0\} &= \frac{3^0}{0!}e^{-3} = 0.0498 \\
\mathbf{Pr}\{D=1\} &= \frac{3^1}{1!}e^{-3} = 0.1494 \\
\mathbf{Pr}\{D=2\} &= \frac{3^2}{2!}e^{-3} = 0.2241
\end{aligned}$$

Interpret these values into $F_D(y)$.

$$\begin{aligned}
F_D(0) =& \mathbf{Pr}\{D=0\} = 0.0498 < 0.2 \\
F_D(1) =& \mathbf{Pr}\{D=0\} + \mathbf{Pr}\{D=1\} = 0.1992 < 0.2 \\
F_D(2) =& \mathbf{Pr}\{D=0\} + \mathbf{Pr}\{D=1\} + \mathbf{Pr}\{D=2\} = 0.4233 \geq 0.2
\end{aligned}$$

Hence, the optimal production quantity here is 2.

1.3 Initial Inventory

Now let us extend our model a bit further. As opposed to the assumption that we had no inventory at the beginning, suppose that we have m items when we decide how many we need to order. The solutions we have developed in previous sections assumed that we had no inventory when placing an order. If we had m items, we should order $y^* - m$ items instead of y^* items. In other words, the optimal order or production quantity is in fact the optimal order-up-to or production-up-to quantity.

We had another implicit assumption that we should order, so the fixed cost did not matter in the previous model. However, if c_f is very large, meaning that starting off a production line or placing an order is very expensive, we may want to consider not to order. In such case, we have two scenarios: to order or not to order. We will compare the expected cost for the two scenarios and choose the option with lower expected cost.

Example 1.5. Suppose understock cost is \$10, overstock cost is \$2, unit purchasing cost is \$4 and fixed ordering cost is \$30. In other words, $c_u = 10, c_o = 2, c_v = 4, c_f = 30$. Assume that $D \sim \text{Uniform}(10, 20)$ and we already possess 10 items. Should we order or not? If we should, how many items should we order?
Answer: First, we need to compute the optimal amount of items we need to prepare for each day. Since

$$\frac{c_u - c_v}{c_u + c_o} = \frac{10-4}{10+2} = \frac{1}{2},$$

the optimal order-up-to quantity $y^* = 15$ units. Hence, if we need to order, we should order $5 = y^* - m = 15 - 10$ items. Let us examine whether we should actually order or not.

1. Scenario 1: Not To Order

   If we decide not to order, we will not have to pay c_f and c_v since we order nothing actually. We just need to consider understock and overstock risks. We will operate tomorrow with 10 items that we currently have if we decide not to order.

$$\begin{aligned}
\mathbf{E}[\text{Cost}] =& c_u \mathbf{E}[(D-10)^+] + c_o \mathbf{E}[(10-D)^+] \\
=& 10(\mathbf{E}[D] - 10) + 2(0) = \$50
\end{aligned}$$

Note that in this case $\mathbf{E}[(10-D)^+] = 0$ because D is always greater than 10.

2. Scenario 2: To Order

   If we decide to order, we will order 5 items. We should pay c_f and c_v accordingly. Understock and overstock risks also exist in this case. Since we will order 5 items to lift up the inventory level to 15, we will run tomorrow with 15 items instead of 10 items if we decide to order.

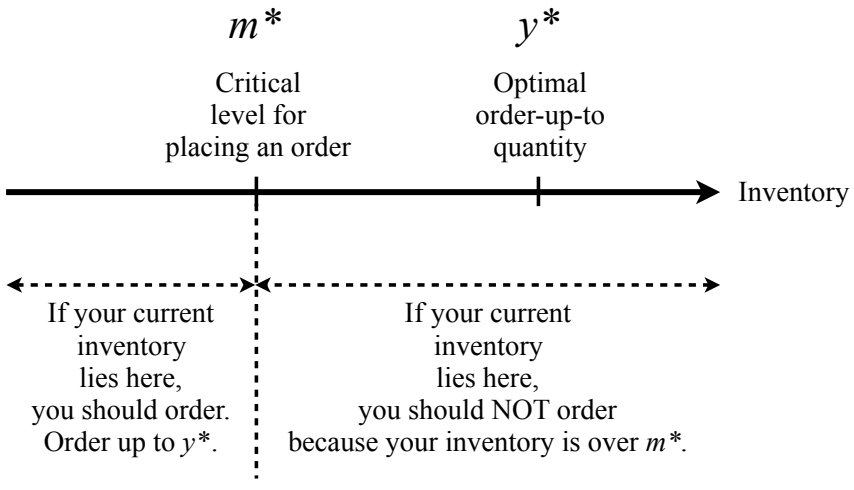
$$\begin{aligned}\mathbf{E}[\text{Cost}] =& c_f + (15-10)c_v + c_u\mathbf{E}[(D-15)^+] + c_o\mathbf{E}[(15-D)^+] \\ =& 30 + 20 + 10(1.25) + 2(1.25) = \$65\end{aligned}$$

Since the expected cost of not ordering is lower than that of ordering, we should not order if we already have 10 items.

It is obvious that if we have y^* items at hands right now, we should order nothing since we already possess the optimal amount of items for tomorrow's operation. It is also obvious that if we have nothing currently, we should order y^* items to prepare y^* items for tomorrow. There should be a point between 0 and y^* where you are indifferent between order and not ordering.

Suppose you as a manager should give instruction to your assistant on when he/she should place an order and when should not. Instead of providing instructions for every possible current inventory level, it is easier to give your assistant just one number that separates the decision. Let us call that number the critical level of current inventory m^*. If we have more than m^* items at hands, the expected cost of not ordering will be lower than the expected cost of ordering, so we should not order. Conversely, if we have less than m^* items currently, we should order. Therefore, when we have exactly m^* items at hands right now, the expected cost of ordering should be equal to that of not ordering. We will use this intuition to obtain m^* value.

The decision process is summarized in the following figure.



Example 1.6. Given the same settings with the previous example ($c_u = 10, c_v = 4, c_o = 2, c_f = 30$), what is the critical level of current inventory m^* that determines whether you should order or not?
Answer: From the answer of the previous example, we can infer that the critical value should be less than 10, i.e. $0 < m^* < 10$. Suppose we currently own m^* items. Now, evaluate the expected costs of the two scenarios: ordering and not ordering.

1. Scenario 1: Not Ordering

$$\begin{aligned}\mathbf{E}[\text{Cost}] =& c_u\mathbf{E}[(D-m^*)^+] + c_o\mathbf{E}[(m^*-D)^+] \\ =& 10(\mathbf{E}[D]-m^*) + 2(0) = 150 - 10m^* \end{aligned} \tag{1.14}$$

2. Scenario 2: Ordering

   In this case, we will order. Given that we will order, we will order $y^* - m^* = 15 - m^*$ items. Therefore, we will start tomorrow with 15 items.

$$\begin{aligned}\mathbf{E}[\text{Cost}] =& c_f + (15-10)c_v + c_u\mathbf{E}[(D-15)^+] + c_o\mathbf{E}[(15-D)^+] \\ =& 30 + 4(15-m^*) + 10(1.25) + 2(1.25) = 105 - 4m^* \end{aligned} \tag{1.15}$$

At m^*, (1.14) and (1.15) should be equal.

$$150 - 10m^* = 105 - 4m^* \quad \Rightarrow \quad m^* = 7.5 \text{ units}$$

The critical value is 7.5 units. If your current inventory is below 7.5, you should order for tomorrow. If the current inventory is above 7.5, you should not order.

1.4 Simulation

Generate 100 random demands from Uniform$(10, 30)$.
$p = 10, c_f = 30, c_v = 4, h = 5, b = 3$

$$\frac{p+b-c_v}{p+b+h} = \frac{10+3-4}{10+3+5} = \frac{1}{2}$$

The optimal order-up-to quantity from Theorem 1.3 is 20. We will compare the performance between the policies of $y = 15, 20, 25$.

Listing 1.1: Continuous Uniform Demand Simulation

```
# Set up parameters
p=10;cf=30;cv=4;h=5;b=3

# How many random demands will be generated?
n=100

# Generate n random demands from the uniform distribution
```

```
Dmd=runif(n,min=10,max=30)

# Test the policy where we order 15 items for every period
y=15
mean(p*pmin(Dmd,y)-cf-y*cv-h*pmax(y-Dmd,0)-b*pmax(Dmd-y,0))
> 33.54218

# Test the policy where we order 20 items for every period
y=20
mean(p*pmin(Dmd,y)-cf-y*cv-h*pmax(y-Dmd,0)-b*pmax(Dmd-y,0))
> 44.37095

# Test the policy where we order 25 items for every period
y=25
mean(p*pmin(Dmd,y)-cf-y*cv-h*pmax(y-Dmd,0)-b*pmax(Dmd-y,0))
> 32.62382
```

You can see the policy with $y = 20$ maximizes the 100-period average profit as promised by the theory. In fact, if n is relatively small, it is not guaranteed that we have maximized profit even if we run based on the optimal policy obtained from this section. The underlying assumption is that we should operate with this policy for a long time. Then, Theorem 1.1 guarantees that the average profit will be maximized when we use the optimal ordering policy.

Discrete demand case can also be simulated. Suppose the demand has the following distribution. All other parameters remain same.

d	10	15	20	25	30
$\mathbf{Pr}\{D = d\}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$

The theoretic optimal order-up-to quantity in this case is also 20. Let us test three policies: $y = 15, 20, 25$.

Listing 1.2: Discrete Demand Simulation

```
# Set up parameters
p=10;cf=30;cv=4;h=5;b=3

# How many random demands will be generated?
n=100

# Generate n random demands from the discrete demand distribution
Dmd=sample(c(10,15,20,25,30),n,replace=TRUE,c(1/4,1/8,1/4,1/8,1/4))

# Test the policy where we order 15 items for every period
y=15
mean(p*pmin(Dmd,y)-cf-y*cv-h*pmax(y-Dmd,0)-b*pmax(Dmd-y,0))
> 19.35

# Test the policy where we order 20 items for every period
y=20
mean(p*pmin(Dmd,y)-cf-y*cv-h*pmax(y-Dmd,0)-b*pmax(Dmd-y,0))
> 31.05

# Test the policy where we order 25 items for every period
```

```
y=25
mean(p*pmin(Dmd,y)-cf-y*cv-h*pmax(y-Dmd,0)-b*pmax(Dmd-y,0))
> 26.55
```

There are other distributions such as triangular, normal, Poisson or binomial distributions available in R. When you do your senior project, for example, you will observe the demand for a department or a factory. You first approximate the demand using these theoretically established distributions. Then, you can simulate the performance of possible operation policies.

1.5 Exercise

1. Show that $(D \wedge y) + (y - D)^+ = y$.

2. Let D be a discrete random variable with the following pmf.

d	5	6	7	8	9
$\mathbf{Pr}\{D = d\}$	$\frac{1}{10}$	$\frac{3}{10}$	$\frac{4}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

   Find

   (a) $\mathbf{E}[\min(D, 7)]$

   (b) $\mathbf{E}[(7 - D)^+]$

   where $x^+ = \max(x, 0)$.

3. Let D be a Poisson random variable with parameter 3. Find

   (a) $\mathbf{E}[\min(D, 2)]$

   (b) $\mathbf{E}[(3 - D)^+]$.

   Note that pmf of a Poisson random variable with parameter λ is

$$\mathbf{Pr}\{D = k\} = \frac{\lambda^k}{k!} e^{-\lambda}.$$

4. Let D be a continuous random variable and uniformly distributed between 5 and 10. Find

   (a) $\mathbf{E}[\max(D, 8)]$

   (b) $\mathbf{E}[(D - 8)^-]$

   where $x^- = \min(x, 0)$.

5. Let D be an exponential random variable with parameter 7. Find

   (a) $\mathbf{E}[\max(D, 3)]$

(b) $\mathbf{E}[(D-4)^-]$.

Note that pdf of an exponential random variable with parameter λ is

$$f_D(x) = \lambda e^{-\lambda x} \text{ for } x \geq 0.$$

6. David buys fruits and vegetables wholesale and retails them at Davids Produce on La Vista Road. One of the more difficult decisions is the amount of bananas to buy. Let us make some simplifying assumptions, and assume that David purchases bananas once a week at 10 cents per pound and retails them at 30 cents per pound during the week. Bananas that are more than a week old are too ripe and are sold for 5 cents per pound.

   (a) Suppose the demand for the good bananas follows the same distribution as D given in Problem 2. What is the expected profit of David in a week if he buys 7 pounds of banana?

   (b) Now assume that the demand for the good bananas is uniformly distributed between 5 and 10 like in Problem 4. What is the expected profit of David in a week if he buys 7 pounds of banana?

   (c) Find the expected profit if David's demand for the good bananas follows an exponential distribution with mean 7 and if he buys 7 pounds of banana.

7. Suppose we are selling lemonade during a football game. The lemonade sells for \$18 per gallon but only costs \$3 per gallon to make. If we run out of lemonade during the game, it will be impossible to get more. On the other hand, leftover lemonade has a value of \$1. Assume that we believe the fans would buy 10 gallons with probability 0.1, 11 gallons with probability 0.2, 12 gallons with probability 0.4, 13 gallons with probability 0.2, and 14 gallons with probability 0.1.

   (a) What is the mean demand?

   (b) If 11 gallons are prepared, what is the expected profit?

   (c) What is the best amount of lemonade to order before the game?

   (d) Instead, suppose that the demand was normally distributed with mean 1000 gallons and variance 200 gallons^2. How much lemonade should be ordered?

8. Suppose that a bakery specializes in chocolate cakes. Assume the cakes retail at \$20 per cake, but it takes \$10 to prepare each cake. Cakes cannot be sold after one week, and they have a negligible salvage value. It is estimated that the weekly demand for cakes is: 15 cakes in 5% of the weeks, 16 cakes in 20% of the weeks, 17 cakes in 30% of the weeks, 18 cakes in 25% of the weeks, 19 cakes in 10% of the weeks, and 20 cakes in 10% of the weeks. How many cakes should the bakery prepare each week? What is the bakery's expected optimal weekly profit?

9. A camera store specializes in a particular popular and fancy camera. Assume that these cameras become obsolete at the end of the month. They guarantee that if they are out of stock, they will special-order the camera and promise delivery the next day. In fact, what the store does is to purchase the camera from an out of state retailer and have it delivered through an express service. Thus, when the store is out of stock, they actually lose the sales price of the camera and the shipping charge, but they maintain their good reputation. The retail price of the camera is \$600, and the special delivery charge adds another \$50 to the cost. At the end of each month, there is an inventory holding cost of \$25 for each camera in stock (for doing inventory etc). Wholesale cost for the store to purchase the cameras is \$480 each. (Assume that the order can only be made at the beginning of the month.)

   (a) Assume that the demand has a discrete uniform distribution from 10 to 15 cameras a month (inclusive). If 12 cameras are ordered at the beginning of a month, what are the expected overstock cost and the expected understock or shortage cost? What is the expected total cost?

   (b) What is optimal number of cameras to order to minimize the expected total cost?

   (c) Assume that the demand can be approximated by a normal distribution with mean 1000 and standard deviation 100 cameras a month. What is the optimal number of cameras to order to minimize the expected total cost?

10. Next month's production at a manufacturing company will use a certain solvent for part of its production process. Assume that there is an ordering cost of \$1,000 incurred whenever an order for the solvent is placed and the solvent costs \$40 per liter. Due to short product life cycle, unused solvent cannot be used in following months. There will be a \$10 disposal charge for each liter of solvent left over at the end of the month. If there is a shortage of solvent, the production process is seriously disrupted at a cost of \$100 per liter short. Assume that the initial inventory level is m, where $m = 0, 100, 300, 500$ and 700 liters.

   (a) What is the optimal ordering quantity for each case when the demand is discrete with $\mathbf{Pr}\{D = 500\} = \mathbf{Pr}\{D = 800\} = 1/8$, $\mathbf{Pr}\{D = 600\} = 1/2$ and $\mathbf{Pr}\{D = 700\} = 1/4$?

   (b) What is the optimal ordering policy for arbitrary initial inventory level m? (You need to specify the critical value m^* in addition to the optimal order-up-to quantity y^*. When $m \leq m^*$, you make an order. Otherwise, do not order.)

   (c) Assume optimal quantity will be ordered. What is the total expected cost when the initial inventory $m = 0$? What is the total expected cost when the initial inventory $m = 700$?

11. Redo Problem 10 for the case where the demand is governed by the continuous uniform distribution varying between 400 and 800 liters.

12. An automotive company will make one last production run of parts for Part 947A and 947B, which are not interchangeable. These parts are no longer used in new cars, but will be needed as replacements for warranty work in existing cars. The demand during the warranty period for 947A is approximately normally distributed with mean 1,500,000 parts and standard deviation 500,000 parts, while the mean and standard deviation for 947B is 500,000 parts and 100,000 parts. (Assume that two demands are independent.) Ignoring the cost of setting up for producing the part, each part costs only 10 cents to produce. However, if additional parts are needed beyond what has been produced, they will be purchased at 90 cents per part (the same price for which the automotive company sells its parts). Parts remaining at the end of the warranty period have a salvage value of 8 cents per part. There has been a proposal to produce Part 947C, which can be used to replace either of the other two parts. The unit cost of 947C jumps from 10 to 14 cents, but all other costs remain the same.

    (a) Assuming 947C is not produced, how many 947A should be produced?

    (b) Assuming 947C is not produced, how many 947B should be produced?

    (c) How many 947C should be produced in order to satisfy the same fraction of demand from parts produced in-house as in the first two parts of this problem.

    (d) How much money would be saved or lost by producing 947C, but meeting the same fraction of demand in-house?

    (e) Is your answer to question (c), the optimal number of 947C to produce? If not, what would be the optimal number of 947C to produce?

    (f) Should the more expensive part 947C be produced instead of the two existing parts 947A and 947B. Why?

    Hint: compare the expected total costs. Also, suppose that $D \sim \text{Normal}(\mu, \sigma^2)$.

$$\begin{aligned}\int_0^q x\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx &= \int_0^q (x-\mu)\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \\ &\quad +\mu\int_0^q \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \\ &= \int_{\mu^2}^{(q-\mu)^2} \frac{1}{2\sqrt{2\pi}\sigma}e^{-\frac{t}{2\sigma^2}}dt + \mu\mathbf{Pr}\{0 \le D \le q\}\end{aligned}$$

    where, in the 2nd step, we changed variable by letting $t = (x-\mu)^2$.

13. A warranty department manages the after-sale service for a critical part of a product. The department has an obligation to replace any damaged parts in the next 6 months. The number of damaged parts X in the next 6 months is assumed to be a random variable that follows the following distribution:

x	100	200	300	400
$\mathbf{Pr}\{X = x\}$	.1	.2	.5	.2

The department currently has 200 parts in stock. The department needs to decide if it should make one last production run for the part to be used for the next 6 months. To start the production run, the fixed cost is \$2000. The unit cost to produce a part is \$50. During the warranty period of next 6 months, if a replacement request comes and the department does not have a part available in house, it has to buy a part from the spot-market at the cost of \$100 per part. Any part left at the end of 6 month sells at \$10. (There is no holding cost.) Should the department make the production run? If so, how many items should it produce?

14. A store sells a particular brand of fresh juice. By the end of the day, any unsold juice is sold at a discounted price of \$2 per gallon. The store gets the juice daily from a local producer at the cost of \$5 per gallon, and it sells the juice at \$10 per gallon. Assume that the daily demand for the juice is uniformly distributed between 50 gallons to 150 gallons.

    (a) What is the optimal number of gallons that the store should order from the distribution each day in order to maximize the expected profit each day?

    (b) If 100 gallons are ordered, what is the expected profit per day?

15. An auto company is to make one final purchase of a rare engine oil to fulfill its warranty services for certain car models. The current price for the engine oil is \$1 per gallon. If the company runs out the oil during the warranty period, it will purchase the oil from a supply at the market price of \$4 per gallon. Any leftover engine oil after the warranty period is useless, and costs \$1 per gallon to get rid of. Assume the engine oil demand during the warranty is uniformly distributed (continuous distribution) between 1 million gallons to 2 million gallons, and that the company currently has half million gallons of engine oil in stock (free of charge).

    (a) What is the optimal amount of engine oil the company should purchase now in order to minimize the total expected cost?

    (b) If 1 million gallons are purchased now, what is the total expected cost?

16. A company is obligated to provide warranty service for Product A to its customers next year. The warranty demand for the product follows the following distribution.

d	100	200	300	400
$\mathbf{Pr}\{D = d\}$	.2	.4	.3	.1

The company needs to make one production run to satisfy the warranty demand for entire next year. Each unit costs \$100 to produce; the penalty cost of a unit is \$500. By the end of the year, the savage value of each unit is \$50.

(a) Suppose that the company has currently 0 units. What is the optimal quantity to produce in order to minimize the expected total cost? Find the optimal expected total cost.

(b) Suppose that the company has currently 100 units at no cost and there is \$20000 fixed cost to start the production run. What is the optimal quantity to produce in order to minimize the expected total cost? Find the optimal expected total cost.

17. Suppose you are running a restaurant having only one menu, lettuce salad, in the Tech Square. You should order lettuce every day 10pm after closing. Then, your supplier delivers the ordered amount of lettuce 5am next morning. Store hours is from 11am to 9pm every day. The demand for the lettuce salad for a day (11am-9pm) has the following distribution.

d	20	25	30	35
$\mathbf{Pr}\{D = d\}$	1/6	1/3	1/3	1/6

One lettuce salad requires two units of lettuce. The selling price of lettuce salad is \$6, the buying price of one unit of lettuce is \$1. Of course, leftover lettuce of a day cannot be used for future salad and you have to pay 50 cents per unit of lettuce for disposal.

(a) What is the optimal order-up-to quantity of lettuce for a day?

(b) If you ordered 50 units of lettuce today, what is the expected profit of tomorrow? Include the purchasing cost of 50 units of lettuce in your calculation.

Chapter 2

Queueing Theory

Before getting into Discrete-time Markov Chains, we will learn about general issues in the queueing theory. Queueing theory deals with a set of systems having waiting space. It is a very powerful tool that can model a broad range of issues. Starting from analyzing a simple queue, a set of queues connected with each other will be covered as well in the end. This chapter will give you the background knowledge when you read the required book, *The Goal*. We will revisit the queueing theory once we have more advanced modeling techniques and knowledge.

2.1 Introduction

Think about a service system. All of you must have experienced waiting in a service system. One example would be the Student Center or some restaurants. This is a human system. A bit more automated service system that has a queue would be a call center and automated answering machines. We can imagine a manufacturing system instead of a service system.

These waiting systems can be generalized as a set of buffers and servers. There are key factors when you try to model such a system. What would you need to analyze your system?

- How frequently customers come to your system? → **Inter-arrival Times**
- How fast your servers can serve the customers? → **Service Times**
- How many servers do you have? → **Number of Servers**
- How large is your waiting space? → **Queue Size**

If you can collect data about these metrics, you can characterize your queueing system. In general, a queueing system can be denoted as follows.

$$\mathrm{G/G}/s/k$$

The first letter characterizes the distribution of inter-arrival times. The second letter characterizes the distribution of service times. The third number denotes the number of servers of your queueing system. The fourth number denotes the total capacity of your system. The fourth number can be omitted and in such case it means that your capacity is infinite, i.e. your system can contain any number of people in it up to infinity. The letter "G" represents a general distribution. Other candidate characters for this position is "M" and "D" and the meanings are as follows.

- G: General Distribution
- M: Exponential Distribution
- D: Deterministic Distribution (or constant)

The number of servers can vary from one to many to infinity. The size of buffer can also be either finite or infinite. To simplify the model, assume that there is only a single server and we have infinite buffer. By infinite buffer, it means that space is so spacious that it is as if the limit does not exist.

Now we set up the model for our queueing system. In terms of analysis, what are we interested in? What would be the performance measures of such systems that you as a manager should know?

- How long should your customer wait in line on average?
- How long is the waiting line on average?

There are two concepts of average. One is average over time. This applies to the average number of customers in the system or in the queue. The other is average over people. This applies to the average waiting time per customer. You should be able to distinguish these two.

Example 2.1. Assume that the system is empty at $t = 0$. Assume that

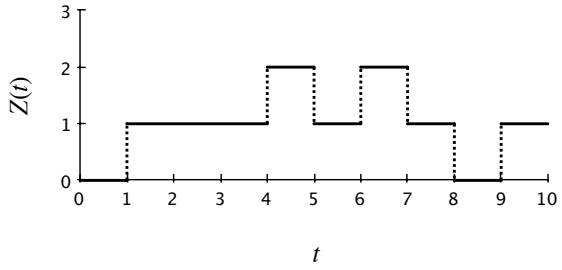
$$u_1 = 1, u_2 = 3, u_3 = 2, u_4 = 3,$$
$$v_1 = 4, v_2 = 2, v_3 = 1, v_4 = 2.$$

(u_i is ith customer's inter-arrival time and v_i is ith customer's service time.)

1. What is the average number of customers in the system during the first 10 minutes?
2. What is the average queue size during the first 10 minutes?
3. What is the average waiting time per customer for the first 4 customers?

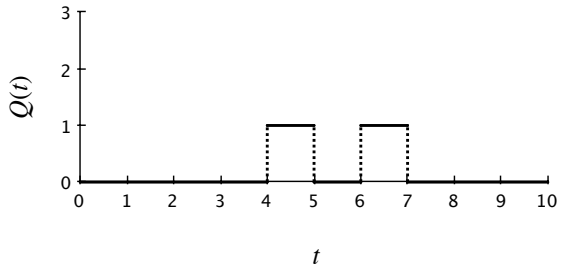
Answer:

1. If we draw the number of people in the system at time t with respect to t, it will be as follows.



$$\mathbf{E}[Z(t)]_{t\in[0,10]} = \frac{1}{10}\int_0^{10} Z(t)dt = \frac{1}{10}(10) = 1$$

2. If we draw the number of people in the queue at time t with respect to t, it will be as follows.



$$\mathbf{E}[Q(t)]_{t\in[0,10]} = \frac{1}{10}\int_0^{10} Q(t)dt = \frac{1}{10}(2) = 0.2$$

3. We first need to compute waiting times for each of 4 customers. Since the first customer does not wait, $w_1 = 0$. Since the second customer arrives at time 4, while the first customer's service ends at time 5. So, the second customer has to wait 1 minute, $w_2 = 1$. Using the similar logic, $w_3 = 1, w_4 = 0$.

$$\mathbf{E}[W] = \frac{0+1+1+0}{4} = 0.5 \text{ min}$$

2.2 Lindley Equation

From the previous example, we now should be able to compute each customer's waiting time given u_i, v_i. It requires too much effort if we have to draw graphs every time we need to compute w_i. Let us generalize the logic behind calculating waiting times for each customer. Let us determine $(i+1)$th customer's waiting

time. If $(i+1)$th customer arrives after all the time ith customer waited and got served, $(i+1)$th customer does not have to wait. Its waiting time is 0. Otherwise, it has to wait $w_i + v_i - u_{i+1}$. Figure 2.1, and Figure 2.2 explain the two cases.

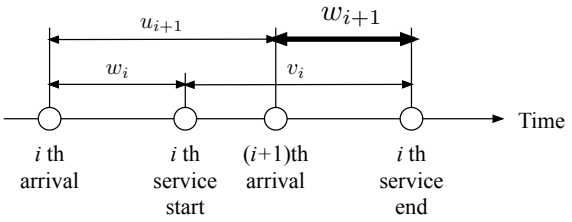


Figure 2.1: $(i+1)$th arrival before ith service completion. $(i+1)$th waiting time is $w_i + v_i - u_{i+1}$.

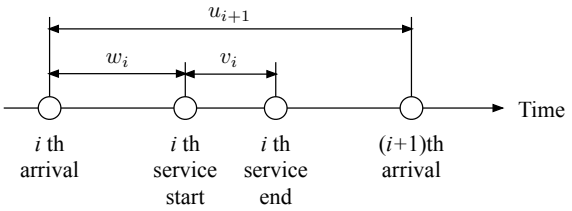


Figure 2.2: $(i+1)$th arrival after ith service completion. $(i+1)$th waiting time is 0.

Simply put,

$$w_{i+1} = (w_i + v_i - u_{i+1})^+.$$

This is called the Lindley Equation.

Example 2.2. Given the following inter-arrival times and service times of first 10 customers, compute waiting times and system times (time spent in the system including waiting time and service time) for each customer.

$$u_i = 3, 2, 5, 1, 2, 4, 1, 5, 3, 2$$
$$v_i = 4, 3, 2, 5, 2, 2, 1, 4, 2, 3$$

Answer: Note that system time can be obtained by adding waiting time and service time. Denote the system time of ith customer by z_i.

u_i	3	2	5	1	2	4	1	5	3	2
v_i	4	3	2	5	2	2	1	4	2	3
w_i	0	2	0	1	4	2	3	0	1	1
z_i	4	5	2	6	6	4	4	4	3	4

2.3 Traffic Intensity

Suppose

$$\begin{aligned}\mathbf{E}[u_i] &= \text{mean inter-arrival time} = 2 \text{ min} \\ \mathbf{E}[v_i] &= \text{mean service time} = 4 \text{ min.}\end{aligned}$$

Is this queueing system stable? By stable, it means that the queue size should not go to the infinity. Intuitively, this queueing system will not last because average service time is greater than average inter-arrival time so your system will soon explode. What was the logic behind this judgement? It was basically comparing the average inter-arrival time and the average service time. To simplify the judgement, we come up with a new quantity called the traffic intensity.

Definition 2.1 (Traffic Intensity)**.** Traffic intensity ρ is defined to be

$$\rho = \frac{\lambda}{\mu} = \frac{1/\mathbf{E}[u_i]}{1/\mathbf{E}[v_i]}$$

where λ is the arrival rate and μ is the service rate.

Given a traffic intensity, it will fall into one of the following three categories.

- If $\rho < 1$, the system is stable.
- If $\rho = 1$, the system is unstable unless both inter-arrival times and service times are deterministic (constant).
- If $\rho > 1$, the system is unstable.

Then, why don't we call ρ utilization instead of traffic intensity? Utilization seems to be more intuitive and user-friendly name. In fact, utilization just happens to be same as ρ if $\rho < 1$. However, the problem arises if $\rho > 1$ because utilization cannot go over 100%. Utilization is bounded above by 1 and that is why traffic intensity is regarded more general notation to compare arrival and service rates.

Definition 2.2 (Utilization)**.** Utilization is defined as follows.

$$\text{Utilization} = \begin{cases} \rho, & \text{if } \rho < 1 \\ 1, & \text{if } \rho \geq 1 \end{cases}$$

Utilization can also be interpreted as the long-run fraction of time the server is utilized.

2.4 Kingman Approximation Formula

Theorem 2.1 (Kingman's High-traffic Approximation Formula)**.** *Assume the traffic intensity $\rho < 1$ and ρ is close to 1. The long-run average waiting time in*

a queue

$$\mathbf{E}[W] \approx \mathbf{E}[v_i]\left(\frac{\rho}{1-\rho}\right)\left(\frac{c_a^2+c_s^2}{2}\right)$$

where c_a^2, c_s^2 are squared coefficient of variation of inter-arrival times and service times defined as follows.

$$c_a^2 = \frac{\mathrm{Var}[u_1]}{(\mathbf{E}[u_1])^2}, \quad c_s^2 = \frac{\mathrm{Var}[v_1]}{(\mathbf{E}[v_1])^2}$$

Example 2.3. 1. Suppose inter-arrival time follows an exponential distribution with mean time 3 minutes and service time follows an exponential distribution with mean time 2 minutes. What is the expected waiting time per customer?

2. Suppose inter-arrival time is constant 3 minutes and service time is also constant 2 minutes. What is the expected waiting time per customer?

Answer:

1. Traffic intensity is

$$\rho = \frac{\lambda}{\mu} = \frac{1/\mathbf{E}[u_i]}{1/\mathbf{E}[v_i]} = \frac{1/3}{1/2} = \frac{2}{3}.$$

Since both inter-arrival times and service times are exponentially distributed,

$$\mathbf{E}[u_i] = 3, \mathrm{Var}[u_i] = 3^2 = 9, \quad \mathbf{E}[v_i] = 2, \mathrm{Var}[v_i] = 2^2 = 4.$$

Therefore, $c_a^2 = \mathrm{Var}[u_i]/(\mathbf{E}[u_i])^2 = 1, c_s^2 = 1$. Hence,

$$\begin{aligned}\mathbf{E}[W] =& \mathbf{E}[v_i]\left(\frac{\rho}{1-\rho}\right)\left(\frac{c_a^2+c_s^2}{2}\right)\\ =& 2\left(\frac{2/3}{1/3}\right)\left(\frac{1+1}{2}\right) = 4 \text{ minutes.}\end{aligned}$$

2. Traffic intensity remains same, 2/3. However, since both inter-arrival times and service times are constant, their variances are 0. Thus, $c_a^2 = c_s^2 = 0$.

$$\mathbf{E}[W] = 2\left(\frac{2/3}{1/3}\right)\left(\frac{0+0}{2}\right) = 0 \text{ minutes}$$

It means that none of the customers will wait upon their arrival.

As shown in the previous example, when the distributions for both inter-arrival times and service times are exponential, the squared coefficient of variation term becomes 1 from the Kingman's approximation formula and the formula

becomes exact to compute the average waiting time per customer for M/M/1 queue.

$$\mathbf{E}[W] = \mathbf{E}[v_i] \left(\frac{\rho}{1-\rho} \right)$$

Also note that if inter-arrival time or service time distribution is deterministic, c_a^2 or c_s^2 becomes 0.

Example 2.4. You are running a highway collecting money at the entering toll gate. You reduced the utilization level of the highway from 90% to 80% by adopting car pool lane. How much does the average waiting time in front of the toll gate decrease?
Answer:

$$\frac{0.9}{1-0.9} = 9, \quad \frac{0.8}{1-0.8} = 4$$

The average waiting time in in front of the toll gate is reduced by more than a half.

The Goal is about identifying bottlenecks in a plant. When you become a manager of a company and are running a expensive machine, you usually want to run it all the time with full utilization. However, the implication of Kingman formula tells you that as your utilization approaches to 100%, the waiting time will be skyrocketing. It means that if there is any uncertainty or random fluctuation input to your system, your system will greatly suffer. In lower ρ region, increasing ρ is not that bad. If ρ near 1, increasing utilization a little bit can lead to a disaster.

Atlanta, 10 years ago, did not suffer that much of traffic problem. As its traffic infrastructure capacity is getting closer to the demand, it is getting more and more fragile to uncertainty.

A lot of strategies presented in *The Goal* is in fact to decrease ρ. You can do various things to reduce ρ of your system by outsourcing some process, etc. You can also strategically manage or balance the load on different parts of your system. You may want to utilize customer service organization 95% of time, while utilization of sales people is 10%.

2.5 Little's Law

$$L = \lambda W$$

The Little's Law is much more general than G/G/1 queue. It can be applied to any black box with definite boundary. The Georgia Tech campus can be one black box. ISyE building itself can be another. In G/G/1 queue, we can easily get average size of queue or service time or time in system as we differently draw box onto the queueing system.

The following example shows that Little's law can be applied in broader context than the queueing theory.

Example 2.5 (Merge of I-75 and I-85)**.** Atlanta is the place where two interstate highways, I-75 and I-85, merge and cross each other. As a traffic manager of Atlanta, you would like to estimate the average time it takes to drive from the north confluence point to the south confluence point. On average, 100 cars per minute enter the merged area from I-75 and 200 cars per minute enter the same area from I-85. You also dispatched a chopper to take a aerial snapshot of the merged area and counted how many cars are in the area. It turned out that on average 3000 cars are within the merged area. What is the average time between entering and exiting the area per vehicle?
Answer:

$$\begin{aligned} L =& 3000 \text{ cars} \\ \lambda =& 100 + 200 = 300 \text{ cars/min} \\ \therefore W =& \frac{L}{\lambda} = \frac{3000}{300} = 10 \text{ minutes} \end{aligned}$$

2.6 Throughput

Another focus of *The Goal* is set on the throughput of a system. Throughput is defined as follows.

Definition 2.3 (Throughput)**.** Throughput is the rate of output flow from a system. If $\rho \leq 1$, throughput$= \lambda$. If $\rho > 1$, throughput$= \mu$.

The bounding constraint of throughput is either arrival rate or service rate depending on the traffic intensity.

Example 2.6 (Tandem queue with two stations)**.** Suppose your factory production line has two stations linked in series. Every raw material coming into your line should be processed by Station A first. Once it is processed by Station A, it goes to Station B for finishing. Suppose raw material is coming into your line at 15 units per minute. Station A can process 20 units per minute and Station B can process 25 units per minute.

1. What is the throughput of the entire system?
2. If we double the arrival rate of raw material from 15 to 30 units per minute, what is the throughput of the whole system?

Answer:

1. First, obtain the traffic intensity for Station A.

$$\rho_A = \frac{\lambda}{\mu_A} = \frac{15}{20} = 0.75$$

Since $\rho_A < 1$, the throughput of Station A is $\lambda = 15$ units per minute. Since Station A and Station B is linked in series, the throughput of Station

A becomes the arrival rate for Station B.

$$\rho_B = \frac{\lambda}{\mu_B} = \frac{15}{25} = 0.6$$

Also, $\rho_B < 1$, the throughput of Station B is $\lambda = 15$ units per minute. Since Station B is the final stage of the entire system, the throughput of the entire system is also $\lambda = 15$ units per minute.

2. Repeat the same steps.

$$\rho_A = \frac{\lambda}{\mu_A} = \frac{30}{20} = 1.5$$

Since $\rho_A > 1$, the throughput of Station A is $\mu_A = 20$ units per minute, which in turn becomes the arrival rate for Station B.

$$\rho_B = \frac{\mu_A}{\mu_B} = \frac{20}{25} = 0.8$$

$\rho_B < 1$, so the throughput of Station B is $\mu_A = 20$ units per minute, which in turn is the throughput of the whole system.

2.7 Simulation

Listing 2.1: Simulation of a Simple Queue and Lindley Equation

```
N = 100

# Function for Lindley Equation
lindley = function(u,v){
  for (i in 1:length(u)) {
    if(i==1) w = 0
    else {
      w = append(w, max(w[i-1]+v[i-1]-u[i], 0))
    }
  }
  return(w)
}

# CASE 1: Discrete Distribution
# Generate N inter-arrival times and service times
u = sample(c(2,3,4),N,replace=TRUE,c(1/3,1/3,1/3))
v = sample(c(1,2,3),N,replace=TRUE,c(1/3,1/3,1/3))

# Compute waiting time for each customer
w = lindley(u,v)
w

# CASE 2: Deterministic Distribution
# All inter-arrival times are 3 minutes and all service times are 2
    minutes
# Observe that nobody waits in this case.
```

```
u = rep(3, 100)
v = rep(2, 100)
w = lindley(u,v)
w
```

The Kingman's approximation formula is exact when inter-arrival times and service times follow iid exponential distribution.

$$\mathbf{E}[W] = \frac{1}{\mu}\left(\frac{\rho}{1-\rho}\right)$$

We can confirm this equation by simulating an M/M/1 queue.

Listing 2.2: Kingman Approximation

```
# lambda = arrival rate, mu = service rate
N = 10000; lambda = 1/10; mu = 1/7

# Generate N inter-arrival times and service times from exponential
    distribution
u = rexp(N,rate=lambda)
v = rexp(N,rate=mu)

# Compute the average waiting time of each customer
w = lindley(u,v)
mean(w)
> 16.20720

# Compare with Kingman approximation
rho = lambda/mu
(1/mu)*(rho/(1-rho))
> 16.33333
```

The Kingman's approximation formula becomes more and more accurate as N grows.

2.8 Exercise

1. Let Y be a random variable with p.d.f. ce^{-3s} for $s \geq 0$, where c is a constant.

   (a) Determine c.

   (b) What is the mean, variance, and squared coefficient of variation of Y where the squared coefficient of variation of Y is defined to be $\mathrm{Var}[Y]/(\mathbf{E}[Y]^2)$?

2. Consider a single server queue. Initially, there is no customer in the system. Suppose that the inter-arrival times of the first 15 customers are:

$$2, 5, 7, 3, 1, 4, 9, 3, 10, 8, 3, 2, 16, 1, 8$$

In other words, the first customer will arrive at $t = 2$ minutes, and the second will arrive at $t = 2 + 5$ minutes, and so on. Also, suppose that the service time of the first 15 customers are

$$1, 4, 2, 8, 3, 7, 5, 2, 6, 11, 9, 2, 1, 7, 6$$

(a) Compute the average waiting time (the time customer spend in buffer) of the first 10 departed customers.

(b) Compute the average system time (waiting time plus service time) of the first 10 departed customers.

(c) Compute the average queue size during the first 20 minutes.

(d) Compute the average server utilization during the first 20 minutes.

(e) Does the Little's law of hold for the average queue size in the first 20 minutes?

3. We want to decide whether to employ a human operator or buy a machine to paint steel beams with a rust inhibitor. Steel beams are produced at a constant rate of one every 14 minutes. A skilled human operator takes an average time of 700 seconds to paint a steel beam, with a standard deviation of 300 seconds. An automatic painter takes on average 40 seconds more than the human painter to paint a beam, but with a standard deviation of only 150 seconds. Estimate the expected waiting time in queue of a steel beam for each of the operators, as well as the expected number of steel beams waiting in queue in each of the two cases. Comment on the effect of variability in service time.

4. The arrival rate of customers to an ATM machine is 30 per hour with exponentially distirbuted in- terarrival times. The transaction times of two customers are independent and identically distributed. Each transaction time (in minutes) is distributed according to the following pdf:

$$f(s) = \begin{cases} 4\lambda^2 s e^{-2\lambda s}, & \text{if } s \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

where $\lambda = 2/3$.

(a) What is the average waiting for each customer?

(b) What is the average number of customers waiting in line?

(c) What is the average number of customers at the site?

5. A production line has two machines, Machine A and Machine B, that are arranged in series. Each job needs to processed by Machine A first. Once it finishes the processing by Machine A, it moves to the next station, to be processed by Machine B. Once it finishes the processing by Machine B, it leaves the production line. Each machine can process one job at a time. An arriving job that finds the machine busy waits in a buffer.

(The buffer sizes are assumed to be infinite.) The processing times for Machine A are iid having exponential distribution with mean 4 minutes. The processing times for Machine B are iid with mean 2 minutes. Assume that the inter-arrival times of jobs arriving at the production line are iid, having exponential distribution with mean of 5 minutes.

(a) What is the utilization of Machine A? What is the utilization of Machine B?

(b) What is the throughput of the production system? (Throughput is defined to be the rate of final output flow, i.e. how many items will exit the system in a unit time.)

(c) What is the average waiting time at Machine A, excluding the service time?

(d) It is known the average time in the entire production line is 30 minutes per job. What is the long-run average number of jobs in the entire production line?

(e) Suppose that the mean inter-arrival time is changed to 1 minute. What are the utilizations for Machine A and Machine B, respectively? What is the throughput of the production system?

6. An auto collision shop has roughly 10 cars arriving per week for repairs. A car waits outside until it is brought inside for bumping. After bumping, the car is painted. On the average, there are 15 cars waiting outside in the yard to be repaired, 10 cars inside in the bump area, and 5 cars inside in the painting area. What is the average length of time a car is in the yard, in the bump area, and in the painting area? What is the average length of time from when a car arrives until it leaves?

7. A small bank is staffed by a single server. It has been observed that, during a normal business day, the inter-arrival times of customers to the bank are iid having exponential distribution with mean 3 minutes. Also, the the processing times of customers are iid having the following distribution (in minutes):

x	1	2	3
$\mathbf{Pr}\{X = x\}$	1/4	1/2	1/4

An arrival finding the server busy joins the queue. The waiting space is infinite.

(a) What is the long-run fraction of time that the server is busy?

(b) What the the long-run average waiting time of each customer in the queue, excluding the processing time?

(c) What is average number of customers in the bank, those in queue plus those in service?

(d) What is the throughput of the bank?

(e) If the inter-arrival times have mean 1 minute. What is the throughput of the bank?

8. You are the manager at the Student Center in charge of running the food court. The food court is composed of two parts: cooking station and cashier's desk. Every person should go to the cooking station, place an order, wait there and pick up first. Then, the person goes to the cashier's desk to check out. After checking out, the person leaves the food court. The cook and the cashier can handle one person at a time. We have only one cook and only one cashier. An arriving person who finds the cook or the cashier busy waits in line. The waiting space is assumed to be infinite. The processing times for the cook are iid having deterministic distribution with mean 2 minutes, i.e. it takes exactly 2 minutes for the cook to process one person. The processing times for the cashier are iid having exponential distribution with mean 1 minute. Assume the inter-arrival times of persons arriving at the food court are iid having exponential distribution with mean 3 minutes.

   (a) What is the utilization of the cook? What is the utilization of the cashier? What is the throughput of the system?

   (b) What is the long-run average system time at the cooking station, including the service time?

   (c) It is known that the average time spent in the food court is 15 minutes per person. What is the long-run average number of people in the food court?

   Assume the mean inter-arrival time is changed to 30 seconds for the next subquestions.

   (d) What is the utilization of the cook? What is the utilization of the cashier? What is the throughput of the system?

   (e) What is the long-run average system time at the cooking station, including the service time?

9. Suppose you are modeling the Georgia Tech (GT)'s undergraduate population. Assume that the undergraduate population is in steady-state.

   (a) Every year 2500 students come to GT as first-year students. It takes 4.5 years on average for each student to graduate. How many undergraduate students does GT have on average at a certain point of time?

   (b) In addition to the first-year students, a unknown number of students come to GT as transferred students every year. It takes 2.5 years on average for a transferred student to graduate. Suppose you know that the average size of GT undergraduate population at a certain

point of time is 13250 students. How many students come to GT as transferred students every year?

Chapter 3

Discrete Time Markov Chain

In the first chapter, we learned how to manage uncertainty when you are selling perishable items such as newspapers, milk, etc. How should we model differently if we are running a business dealing with durable or non-perishable goods? Discrete-time Markov Chain is the right tool to model such a situation.

3.1 Introduction

Suppose that your company is selling home appliances to customers. You are an inventory manager in charge of operating company inventory efficiently. Suppose that the demand for the appliances a week is as follows.

d	0	1	2	3
$\mathbf{Pr}\{D_n = d\}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{8}$

First of all, we need to define the general notion of inventory policy.

Definition 3.1 ((s, S) Policy)**.** If the inventory policy is (s, S), then the ordering principle is as follows given that the current inventory level is X_n.

- If we have less than or equal to s items ($X_n \leq s$), order up to S items, i.e. order $S - X_n$.
- If we have more than s items ($X_n > S$), do not order for the next period.

It is simple but widely adopted inventory policy in real world. In our example, assume our inventory policy to be $(s, S) = (1, 4)$. Let X_n be the inventory level at the end of nth week.

3.1.1 State Space

State space is a set containing all possible values that X_n can take. In this case, X_n can take either one of $0, 1, 2, 3, 4$, so the state space is $\{0, 1, 2, 3, 4\}$.

$$S = \{0, 1, 2, 3, 4\}$$

3.1.2 Transition Probability Matrix

Now, we would like to compute the probability of a future event given the information about our current state. First, think about $\mathbf{Pr}\{X_{n+1} = 3 \mid X_n = 1\}$. In plain English, it means that if we have 1 item at the end of the nth week, what is the probability that we finish the $(n+1)$th week with 3 items. Due to $(s, S) = (1, 4)$, over the weekend between week n and $n+1$, we will order 3 items to fill our inventory up to 4. Consequently, we will start the $(n+1)$th week with 4 items. The question boils down to the probability of us ending up with 3 items given that we had 4 items at the beginning. It is equivalent to the probability we sell only 1 item during the $(n+1)$th week which is $1/4$.

$$\begin{aligned}
&\mathbf{Pr}\{X_{n+1} = 3 \mid X_n = 1\} \\
=&\mathbf{Pr}\{\text{Next week ends with 3 items} \mid \text{This week ended with 1 item}\} \\
=&\mathbf{Pr}\{\text{Next week ends with 3 items} \mid \text{Next week begins with 4 items}\} \\
=&\mathbf{Pr}\{\text{Selling 1 item during the next week}\} \\
=&\mathbf{Pr}\{D = 1\} = \frac{1}{4}
\end{aligned}$$

How about $\mathbf{Pr}\{X_{n+1} = 3 \mid X_n = 3\}$? In this case, we will not order because our current inventory at the end of nth week is 3 which is greater than $s = 2$. Therefore, we will start the next week with 3 items. It boils down to the probability of us ending up with 3 items at the end of the next week given that we have 3 items at the beginning of the next week, meaning that we sell nothing for the next week. The probability of selling nothing during a certain week is $1/8$.

$$\begin{aligned}
&\mathbf{Pr}\{X_{n+1} = 3 \mid X_n = 3\} \\
=&\mathbf{Pr}\{\text{Next week ends with 3 items} \mid \text{This week ended with 3 items}\} \\
=&\mathbf{Pr}\{\text{Next week ends with 3 items} \mid \text{Next week begins with 3 items}\} \\
=&\mathbf{Pr}\{\text{Selling 0 items during the next week}\} \\
=&\mathbf{Pr}\{D = 0\} = \frac{1}{8}
\end{aligned}$$

You should now be able to compute $\mathbf{Pr}\{X_{n+1} = j \mid X_n = i\}$ for all possible (i, j) pairs. There will be in this case $5 \times 5 = 25$ (i, j) pairs since we have 5

states in the state space. We use matrix notation to denote these probabilities.

$$P = \begin{array}{r} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{pmatrix} 0 & 1/8 & 1/2 & 1/4 & 1/8 \\ 0 & 1/8 & 1/2 & 1/4 & 1/8 \\ 5/8 & 1/4 & 1/8 & 0 & 0 \\ 1/8 & 1/2 & 1/4 & 1/8 & 0 \\ 0 & 1/8 & 1/2 & 1/4 & 1/8 \end{pmatrix}$$

Note that row numbers represent the current state and column numbers represent the next state. So, (i, j) element of this matrix is

$$P_{i,j} = \mathbf{Pr}\{X_{n+1} = j \mid X_n = i\}.$$

Another important feature of a transition probability matrix is that each row sums to 1 because you will jump from a state to another in S for sure.

Now we know all possible probabilities between one-period transition. How can we compute the transition probability between multiple period? What is the probability that we finish with 3 items two weeks later given that we ended this week with 1 item? It will involve a property called the Markov property that will be mentioned shortly. Let us proceed for now. In mathematical notation, what is $\mathbf{Pr}\{X_{n+2} = 3 \mid X_n = 1\}$? At this point, let us recall the definition of conditional probability.

$$\mathbf{Pr}\{A \mid B\} = \frac{\mathbf{Pr}\{A \cap B\}}{\mathbf{Pr}\{B\}} = \frac{\mathbf{Pr}\{A, B\}}{\mathbf{Pr}\{B\}} \text{ or } \mathbf{Pr}\{A, B\} = \mathbf{Pr}\{A \mid B\}\mathbf{Pr}\{B\}$$

Using this definition,

$$\begin{aligned}
&\mathbf{Pr}\{X_{n+2} = 3 \mid X_n = 1\} \\
=&\sum_{k=0}^{4} \mathbf{Pr}\{X_{n+2} = 3, X_{n+1} = k \mid X_n = 1\} \\
=&\sum_{k=0}^{4} \mathbf{Pr}\{X_{n+2} = 3 \mid X_{n+1} = k, X_n = 1\}\mathbf{Pr}\{X_{n+1} = k \mid X_n = 1\} \\
=&\sum_{k=0}^{4} \mathbf{Pr}\{X_{n+2} = 3 \mid X_{n+1} = k\}\mathbf{Pr}\{X_{n+1} = k \mid X_n = 1\} \\
&\because \text{Due to the Markov property} \\
=&\mathbf{Pr}\{X_{n+2} = 3 \mid X_{n+1} = 0\}\mathbf{Pr}\{X_{n+1} = 0 \mid X_n = 1\} \\
&+ \mathbf{Pr}\{X_{n+2} = 3 \mid X_{n+1} = 1\}\mathbf{Pr}\{X_{n+1} = 1 \mid X_n = 1\} \\
&+ \mathbf{Pr}\{X_{n+2} = 3 \mid X_{n+1} = 2\}\mathbf{Pr}\{X_{n+1} = 2 \mid X_n = 1\} \\
&+ \mathbf{Pr}\{X_{n+2} = 3 \mid X_{n+1} = 3\}\mathbf{Pr}\{X_{n+1} = 3 \mid X_n = 1\} \\
&+ \mathbf{Pr}\{X_{n+2} = 3 \mid X_{n+1} = 4\}\mathbf{Pr}\{X_{n+1} = 4 \mid X_n = 1\}.
\end{aligned}$$

This formula looks very complicated and it will become more and more complicated and error-prone as the state space grows. Here comes the power of matrix

notation. You can just square the transition probability matrix and look up the corresponding number. In this case, $P^2_{1,3}$ is the probability we are looking for. In general, the transition probability between k periods

$$\mathbf{Pr}\{X_{n+k} = j \mid X_n = i\} = P^k_{i,j}.$$

3.1.3 Initial Distribution

With transition probability matrix, we know what the probabilities are for jumping from one state to another. However, we do not know from which state the chain starts off at the beginning unless we are given the specific information. Initial distribution is the information that determines where the chain starts at time 0. Initial distribution plays a key role to compute unconditioned probability such as $\mathbf{Pr}\{X_n = j\}$ not $\mathbf{Pr}\{X_n = j \mid X_0 = i\}$. In order to compute $\mathbf{Pr}\{X_n = j\}$, we need information about $\mathbf{Pr}\{X_0 = i\}$ for all i. Then,

$$\mathbf{Pr}\{X_n = j\} = \sum_{i \in S} \mathbf{Pr}\{X_n = j \mid X_0 = i\}\mathbf{Pr}\{X_0 = i\}.$$

Example 3.1 (Computing Unconditioned Probability)**.** Suppose that your daily mood only depends on your mood on the previous day. Your mood has three distinctive states: Happy, So-so, Gloomy. The transition probability matrix is given as

$$P = \begin{matrix} \text{Happy} \\ \text{So-so} \\ \text{Gloomy} \end{matrix} \begin{pmatrix} .7 & .3 & 0 \\ .3 & .5 & .2 \\ 0 & .9 & .1 \end{pmatrix}.$$

In addition, your mood on day 0 is determined by a coin toss. You are happy with probability 0.5 and gloomy with probability 0.5. Then, what is the probability that you are gloomy on day 2?

Answer: Denote the states happy, so-so, gloomy by 0, 1, 2, respectively. Since we need two-period transition probabilities from day 0 to day 2, we need to compute P^2.

$$P^2 = \begin{pmatrix} .58 & .36 & .06 \\ .36 & .52 & .12 \\ .27 & .54 & .19 \end{pmatrix}$$

The probability that you are gloomy on day 2 is $\mathbf{Pr}\{X_2 = 2\}$.

$$\begin{aligned} \mathbf{Pr}\{X_2 = 2\} =& \sum_{i=0}^{2} \mathbf{Pr}\{X_2 = 2 \mid X_0 = i\}\mathbf{Pr}\{X_0 = i\} \\ =& \mathbf{Pr}\{X_2 = 2 \mid X_0 = 0\}\mathbf{Pr}\{X_0 = 0\} \\ &+ \mathbf{Pr}\{X_2 = 2 \mid X_0 = 1\}\mathbf{Pr}\{X_0 = 1\} \\ &+ \mathbf{Pr}\{X_2 = 2 \mid X_0 = 2\}\mathbf{Pr}\{X_0 = 2\} \\ =& (.27)(.5) + (.54)(0) + (.19)(.5) = .23 \end{aligned}$$

Hence, you are gloomy on day 2 with 23%.

Handling initial distribution becomes easy when we use matrix notation. An initial distribution can be denoted by a vector a_0, of which each element represents the probability that the chain is in a state at time 0. For example, if $a_0 = (.2, .5, .3)$, it is interpreted as the chain starts from state 1 with probability .2, from state 2 with probability .5, and from state 3 with probability .3. This notation can be extended to any time n. For instance, a_n represents the probabilities that the chain is at a certain state at time n. Then, the following relation exists with this notation.

$$a_{n+1} = a_n P, \quad a_n = a_0 P^n$$

where $a_n = (\mathbf{Pr}\{X_n = 1\}, \mathbf{Pr}\{X_n = 2\}, \ldots, \mathbf{Pr}\{X_n = k\})$ if the state space $S = \{1, 2, \ldots, k\}$.

Now that we learn all necessary elements of DTMC, let us try to model our inventory system using DTMC.

Example 3.2 (Inventory Model for Non-perishable Items)**.** A store sells a particular product which is classified as durable goods, i.e. left-over items can be sold next period. The weekly demands for the product are iid random variables with the following distribution:

d	0	1	2	3
$\mathbf{Pr}\{D = d\}$	.3	.4	.2	.1

The store is closed during the weekends. Unsold items at the end of a week can be sold again the following week. Each Friday evening if the number of remaining items is less than or equal to 2, the store orders enough items to bring the total number of items up to 4 on Monday morning. Otherwise, do not order. (The ordered items reach the store before the store opens on the Monday morning.) Assume that any demand is lost when the product is out of stock. Let X_n be the number of items in stock at the beginning of the nth week, while Y_n be the number of items at the end of the nth week. X_n, Y_n are DTMC. Then, what is the state space and transition probability matrix of X_n, Y_n respectively?

Answer: Our inventory policy, first of all, is $(s, S) = (2, 4)$. Think about X_n first. Since X_n is the inventory level at the beginning of each week, the state space $S = \{3, 4\}$ because if we had less than or equal to 2 items at the end of the previous week, we would have ordered up to 4. Hence, the beginning inventory is always greater than 2 which is our small s in the (s, S) policy. The transition probability matrix is

$$P = \begin{matrix} 3 \\ 4 \end{matrix} \begin{pmatrix} .3 & .7 \\ .4 & .6 \end{pmatrix}. \tag{3.1}$$

How about Y_n? Since Y_n is the end-of-week inventory level, the state space

$S = \{0, 1, 2, 3, 4\}$. The transition probability matrix is

$$P = \begin{array}{r} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{pmatrix} 0 & .1 & .2 & .4 & .3 \\ 0 & .1 & .2 & .4 & .3 \\ 0 & .1 & .2 & .4 & .3 \\ .1 & .2 & .4 & .3 & 0 \\ 0 & .1 & .2 & .4 & .3 \end{pmatrix}.$$

Note that row 0, 1, 2, 4 are identical because if we ended a week with 0, 1, 2, or 4 items, we will start the next week with $S = 4$ items anyway.

Now let us define DTMC in a formal way. As shown above, a DTMC has the following three elements.

- **State space** S: For example, $S = \{0, 1, 2\}$ or $S = \{0, 1, 2, 3, \dots\}$. S does not have to be finite.
- **Transition probability matrix** P: The sum of each row of P should be equal to 1. The sum of each column does not have to be equal to 1.
- **Initial distribution** a: It gives you the information about where your DTMC starts at time $n = 0$.

Definition 3.2 (Discrete Time Markov Chain (DTMC))**.** A discrete time stochastic process $X = \{X_n : n = 0, 1, 2, \dots\}$ is said to be a DTMC on state space S with transition matrix P if for each $n \geq 1$ for $i_0, i_1, i_2, \dots, i, j \in S$

$$\begin{aligned} &\mathbf{Pr}\{X_{n+1} = j \mid X_0 = i_0, X_1 = i_1, X_2 = i_2, \dots, X_{n-1} = i_{n-1}, X_n = i\} \\ =&\mathbf{Pr}\{X_{n+1} = j \mid X_n = i\} = P_{i,j} = (i, j)\text{th entry of } P. \end{aligned} \tag{3.2}$$

3.1.4 Markov Property

The most important part of Definition 3.2 is (3.2). (3.2) is called the Markov property. In plain English, it says that once you know todays state, tomorrows state has nothing to do with past information. No matter how you reached the current state, your tomorrow will only depend on the current state. In (3.2),

- Past states: $X_0 = i_0, X_1 = i_1, X_2 = i_2, \dots, X_{n-1} = i_{n-1}$
- Current state: $X_n = i$
- Future state: $X_{n+1} = j$.

Definition 3.3 (Markov Property)**.** Given the current information (state), future and past are independent.

From information gathering perspective, it is very appealing because you just need to remember the current. For an opposite example, Wikipedia keeps track of all the histories of each article. It requires tremendous effort. This is the beauty of the Markov property.

What if I think my situation depends not only the current but one week ago? Then, you can properly define state space so that each state contains two weeks instead of one week. I have to stress that you are the one who decides how your model should be: what the state space is, etc. You can add a bit more assumption to fit your situation to Markov model.

Then, is our inventory model DTMC? Yes, because we know that we do not have to know the past stock level to decide whether to order or not. Let us look at the following example solely about the Markov property.

Example 3.3 (Weird Frog Jumping Around)**.** Imagine a situation that a frog is jumping around lily pads.

1. Suppose the frog is jumping around 3 pads. Label each pad as 1, 2, and 3. The frog's jumping has been observed extensively and behavioral scientists concluded this frog's jumping pattern to be as follows.

$$1 \to 2 \to 3 \to 2 \to 1 \to 2 \to 3 \to 2 \to \dots$$

   Let X_n denote the pad label that the frog is sitting on at time n. Is X_n a DTMC? If so, what are the state space and the transition probability matrix? If not, why?

2. A friend of the frog is also jumping around lily pads in the near pond. This friend frog jumps around 4 pads instead of 3. Label each pad as 1, 2A, 2B, and 3. The friend frog's jumping pattern is characterized as follows.

$$1 \to 2A \to 3 \to 2B \to 1 \to 2A \to 3 \to 2B \to 1 \to 2A \to \dots$$

   Let Y_n denote the pad on which the friend frog is sitting at time n. Is Y_n a DTMC? If so, what are the state space and the transition probability matrix? If not, why?

Answer:

1. It is not a DTMC because X_n does not have the Markov property. To prove the point, let us try to complete the transition probability matrix given that the state space $S = \{1, 2, 3\}$.

$$P = \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} \begin{pmatrix} 0 & 1 & 0 \\ ? & ? & ? \\ 0 & 1 & 0 \end{pmatrix}$$

   It is impossible to predict the next step of the frog by just looking at the current location of it if the frog is sitting on pad 2. If we know additionally where it came from to reach pad 2, it becomes possible to predict its next step. That is, it requires past information to determine probability of future. Hence the Markov property does not hold. Thus, X_n cannot be a DTMC.

2. It is DTMC. State space $S = \{1, 2A, 2B, 3\}$ and the transition probability matrix is

$$P = \begin{matrix} 1 \\ 2A \\ 2B \\ 3 \end{matrix} \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

3.1.5 DTMC Models

DTMC is a very versatile tool which can simplify and model various situations in the real world. Only inventory situation has been modeled using DTMC so far. Let us explore several other possibilities.

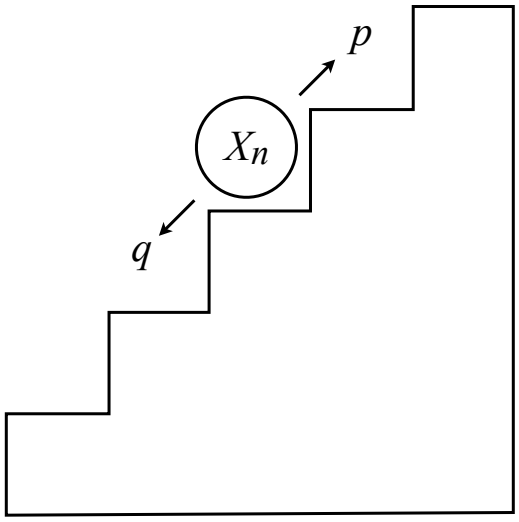
Example 3.4 (Two State Model)**.** $S = \{0, 1\}$ and

$$P = \begin{matrix} 0 \\ 1 \end{matrix} \begin{pmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{pmatrix} = \begin{pmatrix} 3/4 & 1/4 \\ 1/2 & 1/2 \end{pmatrix}$$

What can be modeled using this type of DTMC? Give one example.
Answer: There can be many examples with two-state DTMC. The followings are some examples.

- Consider we are modeling the weather of a day. State 0 means hot and state 1 means cold. Then, the probability of hot after a hot day is 3/4.
- Another example would be machine repairing process. State 0 is the machine is up and running and state 1 means machine is under repair.

Example 3.5 (Simple Random Walk)**.** Suppose you toss a coin at each time n and you go up if you get a head, down if you get a tail. Then, the state space $S = \{\dots, -2, -1, 0, 1, 2, \dots\}$ and X_n is the position after nth toss of the coin. What is the transition probability matrix given that the probability of getting a head is p and that of getting a tail is $q = 1 - p$?



Answer:

$$P_{i,j} = \begin{cases} p, & \text{if } j = i+1 \\ q, & \text{if } j = i-1 \\ 0, & \text{otherwise} \end{cases}$$

Example 3.6 (Google PageRank Algorithm)**.** Suppose you are analyzing browsing behavior of users. There are three web pages of your interest (A, B, C) linked to each other. Since transition from one page to another solely depends on the current page you are viewing not on your browsing history, this case exhibits the Markov property. Hence, we can model this browsing behavior using DTMC. After collecting statistics regarding which link is clicked, you may come up with the following transition probability matrix. It is just one possible example.

$$P = \begin{matrix} \text{A} \\ \text{B} \\ \text{C} \end{matrix} \begin{pmatrix} 0 & .7 & .3 \\ .2 & 0 & .8 \\ .5 & .5 & 0 \end{pmatrix}$$

In practice, there are a great number of web sites and pages, so the corresponding state space and transition probability matrix will be extremely large. However, in this way, you can predict users' browsing behavior and estimate the probability a user will visit a certain web site.

Theorem 3.1. *Suppose a function*

$$f : S \times (-\infty, \infty) \mapsto S, \quad f(i,u) \in S$$

such that $X_{n+1} = f(X_n, U_n)$ *and* $\{U_i : i = 1, 2, 3, \dots\}$ *is an iid sequence. Then,* $\{X_n : n = 1, 2, 3, \dots\}$ *is a DTMC.*

As in Example 3.3, it is usually easy to show that something does not possess the Markov property. It is, however, hard to prove that something does possess the Markov property. Theorem 3.1 is a useful tool in such cases. In plain English, Theorem 3.1 means that if the next state is determined by the current state and independent and identical random trials, the chain has the Markov property. Let us see how Theorem 3.1 works in our examples. Recall the simple random walk case in Example 3.5. Where we will stand is determined by our current position and a coin toss. The coin toss is independent from where we are standing now. Hence, the simple random walk described in Example 3.5 does have the Markov property.

3.2 Stationary Distribution

Now that we have the basic elements of DTMC, the natural question will be what we can do with this new tool. As in the newsvendor model, we are interested in what will happen if we run this system for a long period of time. To understand what will happen in the long run, we introduce a new concept called stationary distribution.

Definition 3.4 (Stationary Distribution)**.** Suppose $\pi = (\pi_i, i \in S)$ satisfies

- $\pi_i \geq 0$ for all $i \in S$ and $\sum_{i \in S} \pi_i = 1$,
- $\pi = \pi P$.

Then, π is said to be a stationary distribution of the DTMC.

Example 3.7. Assume the demand distribution as follows.

d	0	1	2	3
$\mathbf{Pr}(D = d)$	.1	.4	.3	.2

If our inventory policy is $(s, S) = (1, 3)$ and X_n is the inventory at the end of each week, then

$$P = \begin{pmatrix} .2 & .3 & .4 & .1 \\ .2 & .3 & .4 & .1 \\ .5 & .4 & .1 & 0 \\ .2 & .3 & .4 & .1 \end{pmatrix}.$$

What is the stationary distribution of this DTMC?
Answer: Let us apply the $\pi = \pi P$ formula. Then, we have the following system of linear equations.

$$\begin{aligned} \pi_0 &= .2\pi_0 + .2\pi_1 + .5\pi_2 + .2\pi_3 \\ \pi_1 &= .3\pi_0 + .3\pi_1 + .4\pi_2 + .3\pi_3 \\ \pi_2 &= .4\pi_0 + .4\pi_1 + .1\pi_2 + .4\pi_3 \\ \pi_3 &= .1\pi_0 + .1\pi_1 + 0\pi_2 + .1\pi_3 \\ 1 &= \sum_{i=0}^{3} \pi_i \end{aligned}$$

To solve this a bit more easily, you can fix one variable, in this case, $\pi_3 = 1$. After solving with this setting, you will obtain the relative ratio of $\pi_0, \pi_1, \pi_2, \pi_3$. In the end, you need to scale it down or up to make its sum 1. As a result, you will get the following result.

$$\pi_0 = \frac{38}{9}\frac{9}{130}, \quad \pi_1 = \frac{43}{9}\frac{9}{130}, \quad \pi_2 = \frac{40}{9}\frac{9}{130}, \quad \pi_3 = \frac{9}{130} \approx .0692.$$

In the test, you could be asked to give the stationary distribution for a simpler transition probability matrix. Hence, you should practice to compute the stationary distribution from the given transition matrix.

3.2.1 Interpretation of Stationary Distribution

There are two ways to interpret the stationary distribution.

1. Stationary distribution represents the long-run fraction of time that the DTMC stays in each state.
2. Stationary distribution represents the probability that the DTMC will be in a certain state if you run it for very long time.

Using these two interpretations, you can claim the following statements based on the stationary distribution computed in the previous example.

- **Example of Interpretation 1.** If we run this inventory system for a long time, $\pi_0 = 38/130 \approx 29\%$ of time we end a week with zero inventory left.
- **Example of Interpretation 2.** If we run this inventory system for a long time, the chance that we will end a week with three items is $\pi_3 = 9/130 \approx 7\%$.

3.2.2 Function of Stationary Distribution

We want to compute long run average cost or profit of a DTMC. So far, we have not considered any cost associated with inventory. Let us assume the following cost structure in addition to the previous example.

1. Holding cost for each item left by the end of a Friday is \$100 per item.
2. Variable cost for purchasing one item is \$1000.
3. Fixed cost is \$1500.
4. Each item sells \$2000.
5. Unfulfilled demand is lost.

If you are a manager of a company, you will be interested in something like this: What is the long-run average profit per week? You can lose one week, earn another week. But, you should be aware of profitability of your business in general.

Let $f(i)$ be the expected profit of the following week, given that this week's inventory ends with i items. Let us first think about the case $i = 0$ first. You need to order three items, and cost of doing it will be associated with both variable cost and fixed cost. We should also count in the revenue you will earn next week. Fortunately, we do not have to pay any holding cost because we do not have any inventory at the end of this week.

$$\begin{aligned} f(0) &= -\text{ Cost} + \text{Revenue} \\ &= [-3(\$1000) - \$1500] + [3(\$2000)(.2) + 2(\$2000)(.3) + 1(\$2000)(.4) + 0(.1)] \\ &= -\$1300 \end{aligned}$$

This is not the best week you want. How about the case you are left with 2 items at the end of this week? First of all, you should pay the holding cost \$100

per item. When calculating the expected revenue, you should add probabilities for D is 2 or 3. This is because even if the demand is 3, you can only sell 2 items. Since you do not order, there will be no cost associated with ordering.

$$\begin{aligned} f(2) =& - \text{Cost} + \text{Revenue} \\ =& [-2(\$100)] + [(\$0)(.1) + (\$2000)(.4) + (\$4000)(.3+.2)] \\ =& \$2600 \end{aligned}$$

This seems to be quite a good week. The rest can be computed as follows.

$$\begin{aligned} f(1) =& [-1(\$100) - 2(\$1000) - \$1500] \\ & + [3(\$2000)(.2) + 2(\$2000)(.3) + 1(\$2000)(.4) + 0(.1)] = -\$400 \\ f(3) =& [-3(\$100)] + [3(\$2000)(.2) + 2(\$2000)(.3) + 1(\$2000)(.4) + 0(.1)] = \$2900 \end{aligned}$$

Based on this, how would you compute the long-run average profit? This is where the stationary distribution comes into play.

$$\begin{aligned} \text{Long-run avg profit} =& \sum_{i=0}^{3} f(i)\pi_i \\ =& f(0)\pi_0 + f(1)\pi_1 + f(2)\pi_2 + f(3)\pi_3 \\ \approx& \$488.39 \end{aligned}$$

It means that under the $(s, S) = (1, 3)$ policy, we are earning about 500 dollars every week using this policy. So, probably you can now decide whether or not to keep running the business. It does not guarantee that you earn exactly \$488.39 each week. It is the long-run average profit. In terms of getting the optimal policy maximizing the long-run average profit, it involves advanced knowledge on stochastic processes. Especially, this kind of problem cannot be solved using the simplex method because it is highly nonlinear problem. You will be better off by trying to solve it numerically using Matlab, R or other computer aids.

Example 3.8. Suppose your employment status for each week can be modeled as a DTMC. You are either employed and unemployed for a certain week and whether your employment status for the following week solely depends on the this week's status. The transition probability matrix is as follows.

$$P = \begin{matrix} \\ \text{Employed} \\ \text{Unemployed} \end{matrix} \begin{matrix} \begin{matrix} \text{Employed} & \text{Unemployed} \end{matrix} \\ \begin{pmatrix} .9 & .1 \\ .3 & .7 \end{pmatrix} \end{matrix}$$

When employed, your net income is \$300 that week. Otherwise, you spend \$100 for the week, i.e. net income is $-\$100$ for ths week. What is your long-run average net income per week?

Answer: First of all, you need to compute the stationary distribution of the DTMC. π_0 represents the unemployed part, while π_1 represents the employed part.

$$\begin{aligned} .9\pi_0 + .3\pi_1 =& \pi_0 \\ .1\pi_0 + .7\pi_1 =& \pi_1 \\ \pi_0 + \pi_1 =& 1 \end{aligned}$$

Solving this, you can obtain $\pi = (\pi_0, \pi_1) = (.75, .25)$. Hence, the long-run average net income per week $= (-\$100)\pi_0 + (\$300)\pi_1 = \$200$.

3.3 Irreducibility

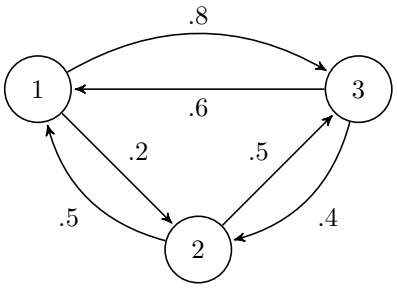
In the following three sections, we will learn a few important characteristics of a DTMC. We will start with irreducibility of a chain. Before defining an irreducible DTMC, it is necessary to understand accessibility of a state.

3.3.1 Transition Diagram

Suppose we have the following transition matrix.

$$P_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0 & .2 & .8 \\ .5 & 0 & .5 \\ .6 & .4 & 0 \end{pmatrix} \end{array}$$

State space $S = \{1, 2, 3\}$. Is the following diagram equivalent to the matrix?



The diagram contains exactly the same amount of information as the transition matrix has.

3.3.2 Accessibility of States

Let $X = \{X_n : n = 0, 1, 2, \dots\}$ be a DTMC on S with transition matrix P.

Definition 3.5 (Accessibility and Irreducibility)**.**

1. State i can reach state j if there exists an n such that $\mathbf{Pr}(X_n = j | X_0 = i) > 0$, i.e. $P_{ij}^n > 0$. This is mathematically noted as $i \to j$.

2. States i and j are said to communicate if $i \to j$ and $j \to i$.

3. X is said to be irreducible if all states communicate. Otherwise, it is said to be reducible.

Why is irreducibility important? If a chain is reducible we may have more than one stationary distribution.

Theorem 3.2. *1. If X is irreducible, there exists, at most one stationary distribution.*

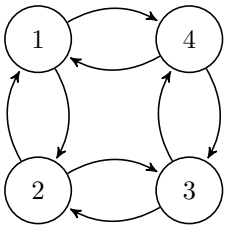
2. If X has a finite state space, it has at least one stationary distribution.

Why is stationary distribution important? As seen in last section, when we compute long-run average profit of a company, we need the stationary distribution. Combining the two cases in Theorem 3.2, we have the following unified assertion.

Theorem 3.3. *For a finite state, irreducible DTMC, there exists a unique stationary distribution.*

3.4 Periodicity

Think of the following DTMC.



In this case, we have $P^{100} \neq P^{101}$. However, according to our corollary, there should be a unique stationary distribution for this chain, too. How do we obtain it given that $P^{100} \neq P^{101}$? How about getting the average of the two?

$$\frac{P^{100} + P^{101}}{2}$$

Even in this case, we cannot obtain the limiting distribution, because this chain oscillates as $n \to \infty$. How do we formally classify such cases?

Definition 3.6 (Periodicity)**.** For a state $i \in S$,

$$d(i) = \gcd\{x : P_{ii}^n > 0\}$$

where gcd is the greatest common divisor.

For example, in the first example,

$$\begin{aligned} d(1) &= \gcd\{2, 4, 3, \dots\} = 1 \\ d(2) &= 1. \end{aligned}$$

In fact, if $i \leftrightarrow j$, $d(i) = d(j)$. It is called the solidarity property. Since all states in a irreducible DTMC communicate, the periods of all states are the same. We will revisit it after covering recurrence and transience.

From the third example, $d(4) = \gcd\{2, 4, 6, \dots\} = 2$, so this DTMC is periodic with period $d = 2$.

Theorem 3.4. *1. If DTMC is aperiodic, then* $\lim_{n\to\infty} P^n$ *exists.*

2. If DTMC is periodic with period $d \geq 2$*,*

$$\lim_{n\to\infty} \frac{P^n + P^{n+1} + P^{n+2} + \cdots + P^{n+d-1}}{n}$$

exists.

Since we are getting too abstract, let us look at another example.

Example 3.9.

$$P = \begin{pmatrix} .2 & .8 \\ .5 & .5 \end{pmatrix}$$

Is this DTMC irreducible? Yes. Now check $P_{ii}^1 > 0$. Therefore, $d(i) = 1$. It is an easy way to check if a DTMC is aperiodic. Therefore, the limiting distribution exists, each row of which is equal to the stationary distribution. What would be the stationary distribution then?

$$\begin{aligned} \pi P &= \pi \\ (\pi_1, \pi_2) \begin{pmatrix} .2 & .8 \\ .5 & .5 \end{pmatrix} &= (\pi_1, \pi_2) \\ \therefore \quad \pi &= \left(\frac{5}{13}, \frac{8}{13} \right) \end{aligned}$$

Hence, even without the help from Matlab, we can say that we know

$$\lim_{n\to\infty} P^n = \begin{pmatrix} 5/13 & 8/13 \\ 5/13 & 8/13 \end{pmatrix}.$$

Example 3.10.

$$P = \begin{pmatrix} 0 & .5 & 0 & .5 \\ .5 & 0 & .5 & 0 \\ 0 & .5 & 0 & .5 \\ .5 & 0 & .5 & 0 \end{pmatrix}$$

According to our theorems, how many distributions do we have? Just one. Can you give the stationary distribution?

$$\pi = (25\%, 25\%, 25\%, 25\%)$$

It means that the long-run average fraction of time you are in each state is a quarter. Then, is this true?

$$P^n = \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}$$

It is not true because P^n alternates between two matrices as n grows. Rather, the following statement is true for this DTMC.

$$\lim_{n\to\infty}(P^n + P^{n+1})/2$$

3.5 Recurrence and Transience

Let X be a DTMC on state space S with transition matrix P. For each state $i \in S$, let τ_i denote the first $n \geq 1$ such that $X_n = i$.

Definition 3.7.
1. State i is said to be recurrent if $\mathbf{Pr}(\tau_i < \infty | X_0 = i) = 1$.
2. State i is said to be positive recurrent if $\mathbf{E}(\tau_i | X_0 = i) < \infty$.
3. State i is said to be transient if it is not recurrent.

Let us revisit the solidarity property to make it complete.

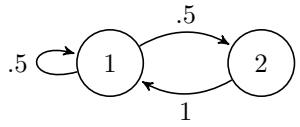
Definition 3.8 (Solidarity Property)**.** If $i \leftrightarrow j$, then $d(i) = d(j)$ where $d(k)$ is the period of state k. Also, i is recurrent or transient if and only if j is recurrent or transient

The solidarity property states that if two states communicate with each other, it is not possible that one state i is recurrent and the other state j is transient. Therefore, if we apply the solidarity property to a irreducible DTMC, only one of the following statements is true.

1. All states are transient.
2. All states are (positive) recurrent.

Remember that all states communicate with each other in a irreducible DTMC.

Example 3.11.

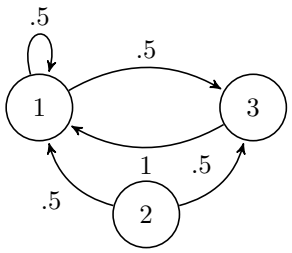


Given that $X_0 = 1$, is it possible $\tau_1 = 1$, meaning that the chain returns to state 1 at time 1? Yes.

$$\begin{aligned}
\mathbf{Pr}(\tau_1 = 1 | X_0 = 1) &= \frac{1}{2} \\
\mathbf{Pr}(\tau_1 = 2 | X_0 = 1) &= \mathbf{Pr}(X_1 \neq 1, X_2 = 1 | X_0 = 1) = (0.5)1 = 0.5 \\
\mathbf{Pr}(\tau_1 = 3 | X_0 = 1) &= 0 \\
\mathbf{E}(\tau_1 | X_0 = 1) &= 1\left(\frac{1}{2}\right) + 2(0.5) = \frac{3}{2} < \infty
\end{aligned}$$

Note that the second probability is not 0.25+0.5, because X_1 should not be equal to 1 for $\tau_1 = 2$. Since the last expectation is finite, this chain is positive recurrent.

Example 3.12.



Is state 1 positive recurrent? Yes. State 2 is transient and how about state 3?

$$\begin{aligned}
\mathbf{Pr}(\tau_3 = 1 | X_0 = 3) =& 0 \\
\mathbf{Pr}(\tau_3 = 2 | X_0 = 3) =& 0.5 \\
\mathbf{Pr}(\tau_3 = 3 | X_0 = 3) =& (0.5)^2 \\
\mathbf{Pr}(\tau_3 = 4 | X_0 = 3) =& (0.5)^3 \\
& \vdots
\end{aligned}$$

τ_3 is basically a geometric random variable and $\mathbf{E}(\tau_3 | X_0 = 3) = 1/p = 2 < \infty$.

3.5.1 Geometric Random Variable

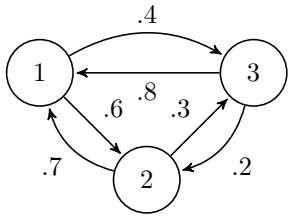
Let us review how to compute the geometric random variable. Let X be the number of tosses to get the 1st head, then X is a geometric random variable. It could take millions of tosses for you to get the 1st head. But, what is the probability X is finite?

$$\begin{aligned}
\mathbf{Pr}(X = 1) =& p \\
\mathbf{Pr}(X = 2) =& pq \\
\mathbf{Pr}(X = 3) =& pq^2 \\
\mathbf{Pr}(X = 4) =& pq^3 \\
& \vdots \\
\mathbf{Pr}(X < \infty) =& \sum_{n=1}^{\infty} \mathbf{Pr}(X = n) \\
=& p + pq + pq^2 + pq^3 + \dots \\
=& p(1 + q + q^2 + \dots) = \frac{p}{1-q} = 1
\end{aligned}$$

How about the expectation?

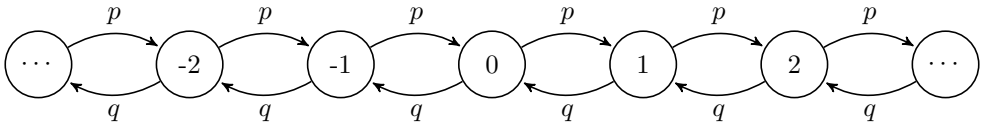
$$\begin{aligned}\mathbf{E}(X) =& 1p + 2pq + 3pq^2 + \dots \\ =& p(1 + 2q + 3q^2 + \dots) = p(q + q^2 + q^3 + \dots)' \\ =& p\left(\frac{1}{1-q}\right)' = \frac{p}{(1-q)^2} = \frac{1}{p}\end{aligned}$$

Example 3.13.



In this case, we know that state 1 is positive recurrent. We also know that $1 \leftrightarrow 2$ and $1 \leftrightarrow 3$. Thus, we can conclude that all states in this chain is positive recurrent.

Example 3.14. Consider a simple random walk.



1. $p = 1/3, q = 2/3$: State i is transient. Say you started from 100, there is positive probability that you never return to state 100. Hence, every state is transient.

2. $p = 2/3 > q$: By the strong law of large numbers, $\mathbf{Pr}(S_n/n \to -1/3) = 1$ because

$$S_n = \sum_{i=1}^{n} \xi_i, \quad \frac{S_n}{n} = \frac{\sum_{i=1}^{n} \xi_i}{n} \to \mathbf{E}(\xi_1) = (-1)q + (1)p = -\frac{2}{3} + \frac{1}{3} = -\frac{1}{3}.$$

3. $p = q = 1/2$: State i is recurrent, but not positive recurrent.

Note from the example above that if the chain is irreducible, every state is either recurrent or transient.

Theorem 3.5. *Assume X is irreducible. X is positive recurrent if and only if X has a (unique) stationary distribution $\pi = (\pi_i)$. Furthermore,*

$$\mathbf{E}(\tau_i | X_0 = i) = \frac{1}{\pi_i}.$$

Recall that one of the interpretation of stationary distribution is the long-run average time the chain spend in a state.

Theorem 3.6. *Assume that X is irreducible in a finite state space. Then, X is recurrent. Furthermore, X is positive recurrent.*

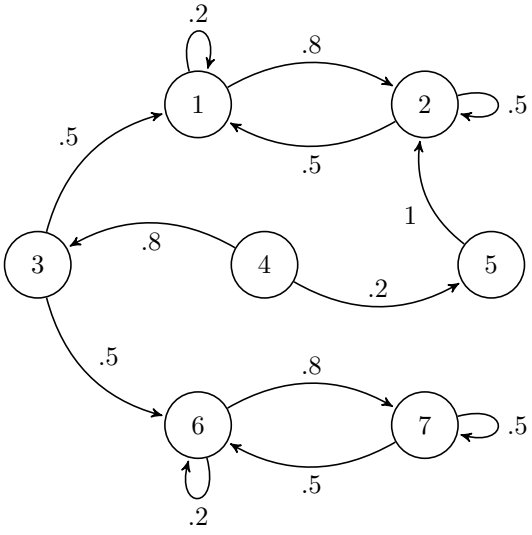
In a finite state space DTMC, there is no difference between positive recurrence and recurrence. These two theorems are important because it gives us a simple tool to check if stationary distribution exists. How about limit distribution? Depending on transition matrix, there may or may not be limiting distribution. When do we have limiting distribution? Aperiodic. If periodic, we used average. Let us summarize.

Theorem 3.7. *Let X be an irreducible finite state DTMC.*

1. *X is positive recurrent.*
2. *X has a unique stationary $\pi = (\pi_i)$.*
3. *$\lim_{n\to\infty} P_{ij}^n = \pi_j$, independent of what i is if the DTMC is aperiodic.*
4. $$\lim_{n\to\infty} \frac{(P_{ij}^n + P_{ij}^{n+1} + P_{ij}^{n+2} + \cdots + P_{ij}^{n+d-1})}{d} = \pi_j,$$
   if the DTMC is periodic with period d.

3.6 Absorption Probability

Think of the following DTMC. You are asked to compute $\lim_{n\to\infty} P^n$.



When computing rows 1, 2, you can just forget about states except for 1 and 2 because there is no arrow going out. Same for rows 6, 7.

$$P^{100} \approx \lim_{n\to\infty} P^n = \begin{matrix}1\\2\\3\\4\\5\\6\\7\end{matrix}\begin{pmatrix} 2/3 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 2/3 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ ? & ? & 0 & 0 & 0 & ? & ? \\ ? & ? & 0 & 0 & 0 & ? & ? \\ ? & ? & 0 & 0 & 0 & ? & ? \\ 0 & 0 & 0 & 0 & 0 & 2/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 2/3 & 1/3 \end{pmatrix}$$

What would be P_{31}? Before this, note that $\{1,2\}, \{6,7\}$ are closed sets, meaning no arrows are going out. Contrarily, $\{3,4,5\}$ are transient states. It means that the DTMC starts from state 3, it may be *absorbed* into either $\{1,2\}$ or $\{6,7\}$.

Let us define a new notation. Let $f_{3,\{1,2\}}$ denote the probability that, starting state 3, the DTMC ends in $\{1,2\}$. Thus, $f_{3,\{1,2\}} + f_{3,\{6,7\}} = 1$. Let us compute the numbers for these.

$$\begin{aligned} f_{3,\{1,2\}} &= (.25)(1) + (.5)(f_{4,\{1,2\}}) + (.25)(0) \\ f_{4,\{1,2\}} &= (.5)(1) + (.5)(f_{5,\{1,2\}}) \\ f_{5,\{1,2\}} &= (.5)(f_{3,\{1,2\}}) + (.25)(f_{4,\{1,2\}}) + (.25)(0) \end{aligned}$$

We have three unknowns and three equations, so we can solve this system of linear equations.

$$\begin{cases} x = & .25 + .5y \\ y = & .5 + .5z \\ z = & .5x + .25y \end{cases} \quad \Rightarrow \quad \begin{cases} x = & f_{3,\{1,2\}} = \frac{5}{8} \\ y = & f_{4,\{1,2\}} = \frac{3}{4} \\ z = & f_{5,\{1,2\}} = \frac{1}{2} \end{cases}$$

We also now know that $f_{3,\{6,7\}} = 1-5/8 = 3/8$, $f_{4,\{6,7\}} = 1-3/4 = 1/4$, $f_{5,\{6,7\}} = 1 - 1/2 = 1/2$.

However, to compute P_{31}, we consider not only $f_{3,\{1,2\}}$ but also the probability that the DTMC will be in state 1 not in state 2. Therefore,

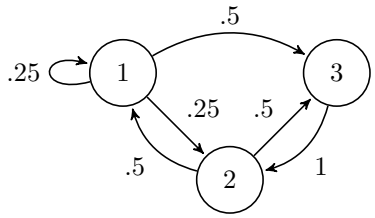
$$\begin{cases} P_{31} = & f_{3,\{1,2\}}\pi_1 = \frac{5}{8}\frac{2}{3} \\ P_{32} = & f_{3,\{1,2\}}\pi_2 = \frac{5}{8}\frac{1}{3} \\ P_{36} = & f_{3,\{6,7\}}\pi_6 = \frac{3}{8}\frac{2}{3} \\ P_{37} = & f_{3,\{6,7\}}\pi_7 = \frac{3}{8}\frac{1}{3} \end{cases}, \quad \begin{cases} P_{41} = & f_{4,\{1,2\}}\pi_1 = \frac{3}{4}\frac{2}{3} \\ P_{42} = & f_{4,\{1,2\}}\pi_2 = \frac{3}{4}\frac{1}{3} \\ P_{46} = & f_{4,\{6,7\}}\pi_6 = \frac{1}{4}\frac{2}{3} \\ P_{47} = & f_{4,\{6,7\}}\pi_7 = \frac{1}{4}\frac{1}{3} \end{cases}, \quad \begin{cases} P_{51} = & f_{5,\{1,2\}}\pi_1 = \frac{1}{2}\frac{2}{3} \\ P_{52} = & f_{5,\{1,2\}}\pi_2 = \frac{1}{2}\frac{1}{3} \\ P_{56} = & f_{5,\{6,7\}}\pi_6 = \frac{1}{2}\frac{2}{3} \\ P_{57} = & f_{5,\{6,7\}}\pi_7 = \frac{1}{2}\frac{1}{3} \end{cases}.$$

Finally, we have the following complete limiting transition matrix.

$$P^{100} \approx \lim_{n\to\infty} P^n = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{pmatrix} 2/3 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 2/3 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 5/12 & 5/24 & 0 & 0 & 0 & 1/4 & 1/8 \\ 1/2 & 1/4 & 0 & 0 & 0 & 1/6 & 1/12 \\ 1/3 & 1/6 & 0 & 0 & 0 & 1/3 & 1/6 \\ 0 & 0 & 0 & 0 & 0 & 2/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 2/3 & 1/3 \end{pmatrix}$$

3.7 Computing Stationary Distribution Using Cut Method

Example 3.15. Assume that X is an irreducible DTMC that is positive recurrent. It means X has a unique stationary distribution. Transition matrix can be described as in the following state diagram.



How do we find the stationary distribution? As far as we learned, we can solve the following system of linear equations.

$$\pi = \pi P, \quad \sum_i \pi_i = 1$$

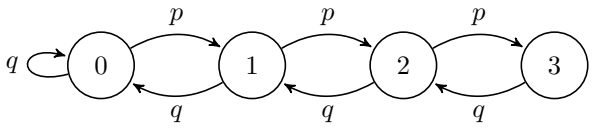
This is sometimes doable, but becomes easily tedious as the number of states increases.

We can instead use the *flow balance equations*. The idea is that for a state, rate into the state should be equal to rate out of the state. Considering flow in and out, we ignore self feedback loop. Utilizing this idea, we have the following equations.

$$\begin{aligned}
\text{State 1}: &\quad \text{rate in} = .5\pi_2 \\
&\quad \text{rate out} = \pi_1(.5 + .25) \\
\text{State 2}: &\quad \text{rate in} = .25\pi_1 + 1\pi_3 \\
&\quad \text{rate out} = .25\pi_2 \\
\text{State 3}: &\quad \text{rate in} = .5\pi_1 + .5\pi_2 \\
&\quad \text{rate out} = 1\pi_3
\end{aligned}$$

Equating each pair of rate in and out, now we have three equations. These three equations are equivalent to the equations we can get from $\pi = \pi P$.

Example 3.16 (Reflected Random Walk)**.** Suppose X has the following transition diagram.



If the probability of the feedback loop at state 0 is 1, it means that you will be stuck there once you get in there. In business analogy or gambler analogy, it means the game is over. Somebody forcefully pick the DTMC out from there to keep the game going. The Wall Street in 2008 resembles this model. They expected they would be bailed out. It is an incentive problem, but we will not cover that issue now.

Suppose $p = 1/3, q = 2/3$. Then, the chain is irreducible. I can boldly say that there exists a unique stationary distribution. However, solving $\pi = \pi P$ gives us infinite number of equations. Here the flow balance equations come into play. Let us generalize the approach we used in the previous example.

For any subset of states $A \subset S$,

$$\text{rate into } A = \text{rate out of } A.$$

If $A = \{0, 1, 2\}$, we essentially look at the flow between state 2 and 3 because this state diagram is linked like a simple thin chain. We have the following equations.

$$\begin{aligned}
\pi_1 &= \left(\frac{p}{q}\right)\pi_0 \\
\pi_2 &= \left(\frac{p}{q}\right)\pi_1 = \left(\frac{p}{q}\right)^2\pi_0 \\
\pi_3 &= \left(\frac{p}{q}\right)\pi_2 = \left(\frac{p}{q}\right)^3\pi_0 \\
\pi_4 &= \left(\frac{p}{q}\right)\pi_3 = \left(\frac{p}{q}\right)^4\pi_0
\end{aligned}$$

Every element of the stationary distribution boils down to the value of π_0. Recall

we always have one more condition: $\sum_i \pi_i = 1$.

$$\begin{aligned}\sum_{i=0}^{\infty} \pi_i =& \pi_0 \left(1 + \left(\frac{p}{q}\right) + \left(\frac{p}{q}\right)^2 + \left(\frac{p}{q}\right)^3 + \dots\right) = 1 \\ \pi_0 =& \frac{1}{1 + \left(\frac{p}{q}\right) + \left(\frac{p}{q}\right)^2 + \left(\frac{p}{q}\right)^3 + \dots} = \frac{1}{\frac{1}{1-\frac{p}{q}}} = 1 - \frac{p}{q} = \frac{1}{2} \\ \pi_n =& \left(\frac{p}{q}\right)^n \pi_0\end{aligned}$$

So far, we assumed that $p < q$. What if $p = q$? What would π_0 be? $\pi_0 = \pi_1 = \pi_2 = \cdots = 0$. Then, is π a probability distribution? No. Since we do not have the stationary distribution, this chain is not positive recurrent. What if $p > q$? It gets worse. The chain cannot be positive recurrent because you have nonzero probability of not coming back. In fact, In this case, every state is transient.

I call this method *cut method*. You probably realized that it would be very useful. Remember that positive recurrence means that the chain will eventually come back to a state and the chain will have some types of cycles. You will see the second example often as you study further. It is usually called *birth-death process*.

3.8 Introduction to Binomial Stock Price Model

Suppose you invest in stocks. Let X_n denote the stock price at the end of period n. This period can be month, quarter, or year as you want. Investing in stocks is risky. How would you define the return?

$$R_n = \frac{X_n - X_{n-1}}{X_{n-1}}, \quad n \geq 1$$

R_n is the return in period n. We can think that X_0 is the initial money you invest in. Say $X_0 = 100, X_1 = 110$.

$$R_1 = \frac{X_1 - X_0}{X_0} = \frac{110 - 100}{100} = \frac{1}{10} = 10\%$$

Suppose you are working in a financial firm. You should have a model for stock prices. No model is perfect, but each model has its own strength. One way to model the stock prices is using iid random variables. Assume that $\{R_n\}$ is a series of iid random variables. Then, X_n can be represented as follows.

$$X_n = X_0(1 + R_1)(1 + R_2)(1 + R_3)\dots(1 + R_n) = X_0 \prod_{i=1}^{n}(1 + R_i)$$

$\prod$ is similar to $\sum$ except that it represents multiplication instead of summation. Assuming R_n iid, is $X = \{X_n : n = 0, 1, \ldots, n\}$ DTMC? Let us have more concrete distribution of R_n.

$$\begin{aligned}\mathbf{Pr}(R_n = 0.1) =& p = 20\% \\ \mathbf{Pr}(R_n = 0.05) =& q = (1-p) = 80\%\end{aligned}$$

If you look at the formula for X_n again,

$$X_n = X_{n-1}(1 + R_n) = f(X_{n-1}, R_n)$$

so you know that X_n is a DTMC because X_n can be expressed as a function of X_{n-1} and iid R_n.

This type of model is called *binomial model*. In this model, only two possibilities for R_n exist: 0.1 or 0.05. Hence, we can rewrite X_n as follows.

$$X_n = X_0(1 + .1)^{Z_n}(1 + .05)^{n - Z_n}$$

where $Z_n \sim \text{Binomial}(n, 0.2)$.

Many Wall Street firm uses computational implementation based on these models. Practical cases are so complicated that they may not have analytic solutions. Health care domain also uses this kind of Markov models, e.g. gene mutation. DTMC has a huge application area. Our school has a master program called Quantitative Computational Finance. It models stock prices not only in a discrete manner but in a continuous manner. They have a concept called geometric Brownian motion. Brownian motion has the term like $e^{B(t)}$. In binomial model, we also had the term with powers. If you look at the stock price from newspapers, it looks continuous depending on the resolution. However, it is in fact discrete. There are two mainstreams in academia and practice regarding stock price modeling: discrete and continuous. Famous Black-Scholes formula for options pricing is a continuous model. This is a huge area, so let us keep it this level.

3.9 Simulation

Let us start from the basics of DTMC. Suppose that a DTMC on $S = \{1, 2, 3\}$ has the following transition probability matrix.

$$P = \begin{pmatrix} .3 & .2 & .5 \\ .6 & 0 & .4 \\ 0 & .8 & .2 \end{pmatrix}$$

Suppose we know the initial distribution is $a_0 = (.5, .2, .3)$. How can we compute $\mathbf{Pr}\{X_{10} = 1 \mid X_0 = 3\}$ and $\mathbf{Pr}\{X_{10} = 1\}$ using computer?

Listing 3.1: Basics of DTMC

```
# Surprisingly, R does not have native function for raising a matrix
     to the power. You have to install and import a library called
    expm
library(expm)

# Set up P and a_0
P = matrix(c(.3,.2,.5, .6,0,.4, 0,.8,.2), 3,3, byrow=TRUE)
a0 = c(.5,.2,.3)

# Compute P^10 and read (3,1) element
P %^% 10
(P %^% 10)[3,1]

# Compute a_10 = a_0 * P^10 and read the first element
a0 %*% (P %^% 10)
(a0 %*% (P %^% 10))[1]
```

The different limiting behavior between periodic and aperiodic DTMC can be also easily observed using R. Suppose that we have two slightly different transition probability matrices.

$$P1 = \begin{pmatrix} 0, .6, 0, .4 \\ .2, 0, .8, 0 \\ 0, .1, 0, .9 \\ .7, 0, .3, 0 \end{pmatrix}, \quad P2 = \begin{pmatrix} 0, .2, .4, .4 \\ .2, 0, .8, 0 \\ 0, .1, 0, .9 \\ .7, 0, .3, 0 \end{pmatrix}$$

$P1$ is periodic, while $P2$ is aperiodic. Observe $P1^{100}$ differs from $P1^{101}$, while $P2^{100} = P2^{101}$.

Listing 3.2: Limiting Distribution P^{100} of periodic and aperiodic DTMC

```
library(expm)

# Periodic DTMC
P1 = matrix(c(0,.6,0,.4, .2,0,.8,0, 0,.1,0,.9, .7,0,.3,0), 4,4,
    byrow=TRUE)
P1%^%100
P1%^%101
# Observe that P1^100 differs from P1^101.

# Aperiodic DTMC
P2 = matrix(c(0,.2,.4,.4, .2,0,.8,0, 0,.1,0,.9, .7,0,.3,0), 4,4,
    byrow=TRUE)
P2%^%100
P2%^%101
# Observe that P1^100 is equal to P1^101.
```

3.10 Exercise

1. A store stocks a particular item. The demand for the product each day is 1 item with probability 1/6, 2 items with probability 3/6, and 3 items with probability 2/6. Assume that the daily demands are independent

and identically distributed. Each evening if the remaining stock is less than 3 items, the store orders enough to bring the total stock up to 6 items. These items reach the store before the beginning of the following day. Assume that any demand is lost when the item is out of stock.

(a) Let X_n be the amount in stock at the beginning of day n; assume that $X_0 = 5$. If the process is a Markov chain, give the state space, initial distribution, and transition matrix. If the process is not, explain why it's not.

(b) Let Y_n be the amount in stock at the end of day n; assume that $Y_0 = 2$. If the process is a Markov chain, give the state space, initial distribution, and transition matrix. If the process is not, explain why it's not.

Also compute the following.

(c) $\mathbf{Pr}\{X_2 = 6 \mid X_0 = 5\}$.

(d) $\mathbf{Pr}\{X_2 = 5, X_3 = 4, X_5 = 6 \mid X_0 = 3\}$.

(e) $\mathbf{E}[X_2 \mid X_0 = 6]$.

(f) Assume the initial distribution is $(0, 0, .5, .5)$. Find $\mathbf{Pr}\{X_2 = 6\}$.

(g) With the same initial distribution, find $\mathbf{Pr}\{X_4 = 3, X_1 = 5 \mid X_2 = 6\}$

2. A six-sided die is rolled repeatedly. After each roll $n = 1, 2, \ldots$, let X_n be the largest number rolled in the first n rolls. Is $\{X_n, n \geq 1\}$ a discrete-time Markov chain? If it's not, show that it is not. If it is, answer the following questions:

(a) What is the state space and the transition probabilities of the Markov chain?

(b) What is the distribution of X_1?

3. Redo the previous problem except replace X_n with Y_n where Y_n is the number of sixes among the first n rolls. (So the first question will be, is $\{Y_n, n \geq 1\}$ a discrete-time Markov chain?)

4. Consider two stocks. Stock 1 always sells for \$10 or \$20. If stock 1 is selling for \$10 today, there is a .80 chance that it will sell for \$10 for tomorrow. If it is selling for \$20 today, there is a .90 chance that it will sell for \$20 tomorrow.

   Stock 2 always sells for \$10 or \$25. If stock 2 sells today for \$10 there is a .90 chance that it will sell tomorrow for \$10. If it sells today for \$25, there is a .85 chance that it will sell tomorrow for \$25. Let X_n denote the price of the 1st stock and Y_n denote the price of the 2nd stock during the nth day. Assume that $\{X_n : n \geq 0\}$ and $\{Y_n : n \geq 0\}$ are discrete time Markov chains.

(a) What is the transition matrix for $\{X_n : n \geq 0\}$? Is $\{X_n : n \geq 0\}$ irreducible?

(b) What is the transition matrix for $\{Y_n : n \geq 0\}$? Is $\{Y_n : n \geq 0\}$ irreducible?

(c) What is the stationary distribution of $\{X_n : n \geq 0\}$?

(d) What is the stationary distribution of $\{Y_n : n \geq 0\}$?

(e) On January 1st, your grand parents decide to give you a gift of 300 shares of either Stock 1 or Stock 2. You are to pick one stock. Once you pick the stock you cannot change your mind. To take advantage of a certain tax law, your grand parents dictate that one share of the chosen stock is sold on each trading day. Which stock should you pick to maximize your gift account by the end of the year? (Explain your answer.)

5. Suppose each morning a factory posts the number of days worked in a row without any injuries. Assume that each day is injury free with probability 98/100. Furthermore, assume that whether tomorrow is injury free or not is independent of which of the preceding days were injury free. Let $X_0 = 0$ be the morning the factory first opened. Let X_n be the number posted on the morning after n full days of work.

(a) Is $\{X_n, n \geq 0\}$ a Markov chain? If so, give its state space, initial distribution, and transition matrix P. If not, show that it is not a Markov chain.

(b) Is the Markov chain irreducible? Explain.

(c) Is the Markov chain periodic or aperiodic? Explain and if it is periodic, also give the period.

(d) Find the stationary distribution.

(e) Is the Markov chain positive recurrent? If so, why? If not, why not?

6. Consider the following transition matrix:

$$P = \begin{pmatrix} 0 & 0.5 & 0 & 0.5 \\ 0.6 & 0 & 0.4 & 0 \\ 0 & 0.7 & 0 & 0.3 \\ 0.8 & 0 & 0.2 & 0 \end{pmatrix} \tag{3.3}$$

(a) Is the Markov chain periodic? Give the period of each state.

(b) Is $(\pi_1, \pi_2, \pi_3, \pi_4) = (33/96, 27/96, 15/96, 21/96)$ the stationary distribution of the Markov Chain?

(c) Is $P_{11}^{100} = \pi_1$? Is $P_{11}^{101} = \pi_1$? Give an expression for π_1 in terms of P_{11}^{100} and P_{11}^{101}.

7. For each of the following transition matrices, do the following: (1) Determine whether the Markov chain is irreducible; (2) Find a stationary distribution π; is the stationary distribution unique; (3) Determine whether the Markov chain is periodic; (4) Give the period of each state. (5) Without using any software package, find P^{100} approximately.

   (a)
   $$P = \begin{pmatrix} 0 & 1/3 & 2/3 \\ 2/3 & 0 & 1/3 \\ 1/3 & 2/3 & 0 \end{pmatrix}$$

   (b)
   $$P = \begin{pmatrix} .2 & .8 & 0 & 0 \\ .5 & .5 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & .5 & .5 \end{pmatrix}$$

8. A store sells a particular product. The weekly demands for the product are iid random variables with the following distribution:

d	0	1	2	3
$\Pr\{D = d\}$	.3	.4	.2	.1

   The store is closed during the weekends. Unsold items at the end of a week can be sold again the following week. Each Friday evening if the number of remaining items is at most 2, the store orders enough items to bring the total number of items up to 4 on Monday morning. Otherwise, do not order. (The ordered items reach the store before the store opens on the Monday morning.) Assume that any demand is lost when the product is out of stock. Let X_n be the number of items in stock at the *beginning* of the nth week.

   (a) If week 1 starts with 3 items, what is the probability that week 2 starts with 3 items?

   (b) Is $X = \{X_n : n = 0, 1, \ldots\}$ a discrete time Markov chain? If so, give its state space and the transition matrix, otherwise explain why it is not a Markov chain.

9. Let $X = \{X_n : n = 0, 1, 2, \ldots\}$ be a discrete time Markov chain on state space $S = \{1, 2, 3, 4\}$ with transition matrix
   $$P = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & 0 & \frac{3}{4} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & 0 & \frac{3}{4} & 0 \end{pmatrix}.$$

   (a) Draw a transition diagram

   (b) Find $\mathbf{Pr}\{X_2 = 4 | X_0 = 2\}$.

(c) Find $\mathbf{Pr}\{X_2 = 2, X_4 = 4, X_5 = 1 | X_0 = 2\}$.

(d) What is the period of each state?

(e) Let $\pi = (\frac{1}{8}, \frac{1}{4}, \frac{3}{8}, \frac{1}{4})$. Is π the unique stationary distribution of X? Explain your answer.

(f) Let P^n be the nth power of P. Does $\lim_{n\to\infty} P^n_{1,4} = \frac{1}{4}$ hold? Explain your answer.

(g) Let τ_1 be the first $n \geq 1$ such that $X_n = 1$. Compute $\mathbf{E}(\tau_1 | X_0 = 1)$. (If it takes you a long time to compute it, you are likely on a wrong track.)

10. Let

$$P = \begin{pmatrix} .2 & .8 & 0 & 0 & 0 \\ .5 & .5 & 0 & 0 & 0 \\ 0 & .25 & 0 & .75 & 0 \\ 0 & 0 & .5 & 0 & .5 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Find approximately P^{100}.

11. Suppose that the weekly demand D of a non-perishable product is iid with the following distribution.

d	10	20	30
$\mathbf{Pr}\{D = d\}$	.2	.5	.3

Unused items from one week can be used in the following week. The management decides to use the following inventory policy: whenever the inventory level in Friday evening is less than or equal to 10, an order is made and it will arrive by Monday morning. The order-up-to quantity is set to be 30.

(a) If this week (week 0) starts with inventory 20, what is the probability that week 2 has starting inventory level 10?

(b) If this week (week 0) starts with inventory 20, what is the probability that week 100 has starting inventory level 10?

(c) Suppose that each item sells at \$200. Each item costs \$100 to order, and each leftover item by Friday evening has a holding cost of \$20. Suppose that each order has a fixed cost \$500. Find the long-run average profit per week.

12. Let $X = \{X_n : n = 0, 1, 2, 3, \ldots\}$ be a discrete time Markov chain on the state space $\{1, 2, 3, 4\}$ and transition matrix

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ .2 & 0 & .8 & 0 \\ 0 & .4 & 0 & .6 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

(a) Draw the transition diagram of the Markov chain. Is the Markov chain irreducible? What is the period of state 2?

(b) Find $\mathbf{Pr}\{X_1 = 2, X_3 = 4 | X_0 = 3\}$.

(c) Find $\mathbf{E}(X_2 | X_0 = 3)$.

(d) Find the stationary distribution $\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$ of the Markov chain.

(e) Is the expression $\mathbf{Pr}\{X_{100} = 2 | X_0 = 3\} \approx \pi_2$ correct? If not, correct the expression.

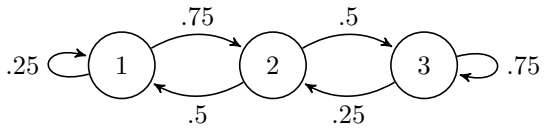
13. Let

$$P = \begin{pmatrix} .4 & .6 & 0 & 0 & 0 & 0 \\ .7 & .3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & .5 & .5 \\ .6 & 0 & .4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

(a) Compute P^{100} approximately. (Leaving fractional forms as they are is okay.)

(b) Approximately, is P^{101} different from P^{100}?

14. Suppose you are working as an independent consultant for a company. Your performance is evaluated at the end of each month and the evaluation falls into one of three categories: 1, 2, and 3. Your evaluation can be modeled as a discrete-time Markov chain and its transition diagram is as follows.



Denote your evaluation at the end of nth month by X_n and assume that $X_0 = 2$.

(a) What are state space, transition probability matrix and initial distribution of X_n?

(b) What is the stationary distribution?

(c) What is the long-run fraction of time when your evaluation is either 2 or 3?

Your monthly salary is determined by the evaluation of each month in the following way.

$$\text{Salary when your evaluation is } n = \$5000 + n^2 \times \$5000, \quad n = 1, 2, 3$$

(d) What is the long-run average monthly salary?

15. Suppose you are modeling your economic status using DTMC. There are five possible economic status: Trillionaire (T), Billionaire (B), Millionaire (M), Thousandaire (Th), Bankrupt (Bk). Your economic status changes every month. Once you are bankrupt, there is no chance of coming back to other states. Also, once you become a billionaire or a trillionaire, you will never be millionaire, thousandaire, or bankrupt. Other chances of transitions are as follows. (Chances for transitions to itself are omitted.)

Transition	Probability
T $\to$ B	0.9
B $\to$ T	0.3
M $\to$ B	0.1
M $\to$ Th	0.5
Th $\to$ M	0.2
Th $\to$ Bk	0.3

(a) Is this chain irreducible? Explain your answer to get credit for this subquestion.

(b) What is the transition probability matrix of this DTMC model?

(c) Compute P^{359}.

(d) What is the probability that you will become a trillionaire (T) 30 years later given that you are a thousandaire (Th) now?

Chapter 4

Poisson Process

In Chapter 2, we learned general queueing theory and how the flow of incoming customers can be modeled. The way we modeled was to denote the inter-arrival times between customers by a random variable and assumed that the random variable has a specific probability distribution. In this chapter, we will learn a special case: the case where the inter-arrival times follow iid exponential distribution. We will learn how it is different from other distributions.

4.1 Exponential Distribution

First of all, let us start from the basics of an exponential distribution.

Definition 4.1. A random variable X is said to have exponential distribution with rate λ (with mean $1/\lambda$) if it has c.d.f. of

$$F(x) = 1 - e^{-\lambda x}$$

and p.d.f. of

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases}.$$

Be careful that rate λ implies that the mean of the exponential distribution is $1/\lambda$. Inter-arrival times are often modeled by a sequence of iid exponential random variables. What do we know about the exponential distribution?

1. $\mathbf{E}[X] = 1/\lambda$
2. $c^2(X) = 1$
3. $\text{Var}[X] = 1/\lambda^2$
4. Memoryless property

What is the memoryless property?

4.1.1 Memoryless Property

$$\mathbf{Pr}\{X > t + s \mid X > s\} = \mathbf{Pr}\{X > t\}$$

Let me paraphrase this concept. Look at the light bulb on the ceiling. Let X denote the lifetime of the bulb and assume that this light bulb's lifetime follows the exponential distribution. If it is on now, the length of its lifetime is as long as a new bulb. s is the time until now and t is additional lifetime. Given that the light has been on for s units of time, the probability that it will be on for t more units of time is same as that of a new light bulb. How is it so?

$$\mathbf{Pr}\{X > t + s \mid X > s\} = \frac{\mathbf{Pr}\{X > t + s\}}{\mathbf{Pr}\{X > s\}} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathbf{Pr}\{X > t\}$$

In fact, it is the defining property of exponential property. It means that if a non-negative continuous random variable has the memoryless property, it must be exponential distribution.

Theorem 4.1.

$$\mathbf{Pr}\{X > t + S \mid X > S\} = \mathbf{Pr}\{X > t\}$$

for any positive random variable S that is independent of X.

Mathematically, we just replaced s with S. Let me explain the change in plain English. If we are saying, "if the light is on at noon today, its remaining lifetime distribution is as if it is new," we are referring to s. If we are saying, "if the light is on when IBM stock price hits 100 dollars, its remaining lifetime is as if it is new," we are referring to S which is random. The important thing to note is that S should be independent of X. Suppose $S = X/2$, i.e. S depends on X.

$$\begin{aligned}\mathbf{Pr}\left\{X > t + \frac{X}{2} \mid X > \frac{X}{2}\right\} =& \frac{\mathbf{Pr}\left\{X > t + \frac{X}{2}\right\}}{\mathbf{Pr}\{X > X/2\}} \\ =& \mathbf{Pr}\left\{X > t + \frac{X}{2}\right\} = \mathbf{Pr}\{X > 2t\} \\ \neq& \mathbf{Pr}\{X > t\}\end{aligned}$$

The memory property does not hold if X and S are not independent.

Come back to other derivations from exponential distribution. Let X_1 and X_2 be independent exponential random variables with rates λ_1 and λ_2 respectively, i.e. $X_1 \sim \text{Exp}(\lambda_1), X_2 \sim \text{Exp}(\lambda_2)$. Let $X = \min(X_1, X_2)$, then X denote the time when any one of two bulbs fails.

$$\begin{aligned}\mathbf{Pr}\{X > t\} =& \mathbf{Pr}\{X_1 > t, X_2 > t\} = \mathbf{Pr}\{X_1 > t\}\mathbf{Pr}\{X_2 > t\} \quad \because X_1 \perp\!\!\!\perp X_2 \\ =& e^{-\lambda_1 x} e^{-\lambda_2 x} = e^{-(\lambda_1+\lambda_2)x}\end{aligned}$$

Therefore, we can say that $X = \min(X_1, X_2) \sim \text{Exp}(\lambda_1 + \lambda_2)$.

Example 4.1. Let X_1 and X_2 follow exponential distribution with means 2 hours and 6 hours respectively. Then, what would be the mean of $\min(X_1, X_2)$?

$$\mathbf{E}[\min(X_1, X_2)] = \frac{1}{\frac{1}{2} + \frac{1}{6}} = \frac{1}{4/6} = 1.5 \text{ hours.}$$

How about the expectation of $\max(X_1, X_2)$? We can use the fact that $X_1 + X_2 = \min(X_1, X_2) + \max(X_1, X_2)$.

$$\mathbf{E}[\max(X_1, X_2)] = \mathbf{E}[X_1 + X_2] - \mathbf{E}[\min(X_1, X_2)] = 8 - 1.5 = 6.5 \text{ hours.}$$

We do not have a convenient way to compute the expectation of $\max(X_1, X_2, \ldots, X_n)$ more than two exponential random variables.

Theorem 4.2.

$$\mathbf{Pr}\{X_1 < X_2\} = \frac{\lambda_1}{\lambda_1 + \lambda_2}, \quad \mathbf{Pr}\{X_1 > X_2\} = \frac{\lambda_2}{\lambda_1 + \lambda_2}, \quad \mathbf{Pr}\{X_1 = X_2\} = 0$$

How to compute this? One way is to guess the answer. As λ_1 becomes large, X_1 gets shorter. It implies that $\mathbf{Pr}\{X_1 < X_2\}$ goes close to 1. Another way is to double integrate.

4.1.2 Comparing Two Exponentials

Let X_1 and X_2 be two independent random variables having exponential distribution with rates λ_1 and λ_2.

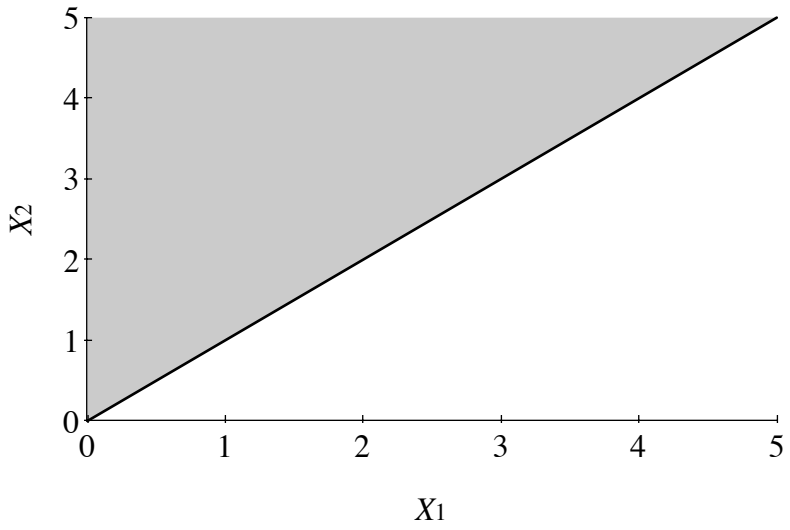
Theorem 4.3.

$$\mathbf{Pr}\{X_1 < X_2\} = \frac{\lambda_1}{\lambda_1 + \lambda_2}, \quad f(x_1, x_2) = \lambda_1 e^{-\lambda_1 x_1} e^{-\lambda_2 x_2}$$

How would you remember the correct formula was $\lambda_1/(\lambda_1+\lambda_2)$ not $\lambda_2/(\lambda_1+\lambda_2)$? With λ_1 increasing, the expected lifetime of the first lightbulb X_1 becomes shorter, hence the probability X_1 breaks down before X_2 gets close to 1. We can also compute the probability using the joint pdf.

$$\mathbf{Pr}\{X_1 < X_2\} = \iint_D f(x_1, x_2) dx_1 dx_2 = \int_0^\infty \left(\int_0^{x_2} f(x_1, x_2) dx_1 \right) dx_2$$

When computing double integral, it is helpful to draw the region where integration applies. The shaded region in the following figure is the integration range in 2 dimensions.



Example 4.2. 1. A bank has two teller, John and Mary. John's processing times are iid exponential distributions X_1 with mean 6 minutes. Mary's processing times are iid exponential distributions X_2 with mean 4 minutes. A car with three customers A, B, C shows up at 12:00 noon and two tellers are both free. What is expected time the car leaves the bank?

Using intuition, we can see that it would be between 8 and 12 minutes. Suppose A, B first start to get service. Once one server completes, C will occupy either A or B's position depending on which server finishes first. If C occupies A's server after A is completed, B has already been in service while A was getting served. Let Y_1, Y_2 denote the remaining processing time for John and Mary respectively. The expected time when the car leaves is

$$\begin{aligned}\mathbf{E}[W] =& \mathbf{E}[\min(X_1, X_2)] + \mathbf{E}[\max(Y_1, Y_2)] \\ =& \mathbf{E}[\min(X_1, X_2)] + \mathbf{E}[\max(X_1, X_2)]\end{aligned}$$

because of the memoryless property of exponential distribution. Even if B has gone through service for a while, due to memoryless property of exponential distribution, B's service time is as if B just started the service. Thus,

$$\begin{aligned}\mathbf{E}[W] =& \mathbf{E}[\min(X_1, X_2) + \max(X_1, X_2)] \\ =& \mathbf{E}[X_1 + X_2] = 10 \text{ minutes}.\end{aligned}$$

2. What is the probability that C finishes service before A?

First compute the probability that B finishes service before A.

$$\begin{aligned}\mathbf{Pr}\{B \text{ before } A\} =& \frac{1/6}{1/4+1/6} = \frac{4}{4+6} = \frac{4}{10}\\ \therefore \mathbf{Pr}\{C \text{ before } A\} =& \left(\frac{4}{10}\right)^2\end{aligned}$$

This is because we can think of A just starting to get service when C started service.

3. What is the probability that C finishes last?

$$\begin{aligned}\mathbf{Pr}\{C \text{ finishes last}\} =& 1 - \mathbf{Pr}\{C \text{ not last}\}\\ =& 1 - (\mathbf{Pr}\{C \text{ finishes before } A\} + \mathbf{Pr}\{C \text{ finishes before } B\})\\ =& 1 - \left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2 = \frac{12}{25}\end{aligned}$$

4.2 Homogeneous Poisson Process

A sequence of customers having iid exponential inter-arrival times is specially called a Poisson process. We will first learn about *homogeneous* Poisson process. Homogeneous Poisson process differs from non-homogeneous Poisson process in that the exponential distribution associated with it has a constant rate λ.

Definition 4.2 (Poisson Process)**.** $N = \{N(t), t \geq 0\}$ is said to be a Poisson process with rate λ if

1. $N(t) - N(s) \sim \text{Poisson}(\lambda \times (t-s))$ for any $0 \leq s < t$
2. N has independent increments, i.e., $N(t) - N(s)$ is independent of $N(u)$ for $0 \leq u \leq s$
3. $N(0) = 0$

N is said to model a counting process, e.g. $N(t)$ is the number of arrivals in $[0, t]$. First of all, whichever time interval you look at closely, say $(s, t]$, the number of arrivals in the interval follows the distribution of Poisson random variable. Second, how many people arrive during a time interval in the past does not affect the number of arrivals in a future interval.

Where do we see this thing in the real world? We have very large population base, say 1 million or 100 thousands. Each person makes independent decision, e.g. placing a call, with small probability, then you will see the Poisson process.

Example 4.3. Assume N is a Poisson process with rate $\lambda = 2/\text{minutes}$.

1. Find the probability that there are exactly 4 arrivals in first 3 minutes.

$$\mathbf{Pr}(N(3) - N(0) = 4) = \frac{(2(3-0))^4}{4!} e^{-2(3-0)} = \frac{6^4}{4!} e^{-6} = 0.1339$$

2. What is the probability that exactly two arrivals in $[0, 2]$ and at least 3 arrivals in $[1, 3]$?

$$\begin{aligned}
\mathbf{Pr}(\{N(2) =& 2\} \cap \{N(3) - N(1) \geq 3\}) \\
=& \mathbf{Pr}(N(1) = 0, N(2) = 2, N(3) - N(1) \geq 3) \\
&+ \mathbf{Pr}(N(1) = 1, N(2) = 2, N(3) - N(1) \geq 3) \\
&+ \mathbf{Pr}(N(1) = 2, N(2) = 2, N(3) - N(1) \geq 3) \\
=& \mathbf{Pr}(N(1) = 0, N(2) - N(1) = 2, N(3) - N(2) \geq 1) \\
&+ \mathbf{Pr}(N(1) = 1, N(2) - N(1) = 1, N(3) - N(2) \geq 2) \\
&+ \mathbf{Pr}(N(1) = 2, N(2) - N(1) = 0, N(3) - N(2) \geq 3) \\
=& \mathbf{Pr}(N(1) = 0)\mathbf{Pr}(N(2) - N(1) = 2)\mathbf{Pr}(N(3) - N(2) \geq 1) \\
&+ \mathbf{Pr}(N(1) = 1)\mathbf{Pr}(N(2) - N(1) = 1)\mathbf{Pr}(N(3) - N(2) \geq 2) \\
&+ \mathbf{Pr}(N(1) = 2)\mathbf{Pr}(N(2) - N(1) = 0)\mathbf{Pr}(N(3) - N(2) \geq 3)
\end{aligned}$$

What are we doing here? Basically, we are decomposing the intervals into non-overlapping ones. Then, we will be able to use the independent increment property. We have learned how to compute $\mathbf{Pr}(N(1) = 0), \mathbf{Pr}(N(2) - N(1) = 2)$. How about $\mathbf{Pr}(N(3) - N(2) \geq 1)$?

$$\begin{aligned}
\mathbf{Pr}\{N(3) - N(2) \geq 1\} =& 1 - \mathbf{Pr}\{N(3) - N(2) < 1\} = 1 - \mathbf{Pr}\{N(3) - N(2) = 0\} \\
=& 1 - \frac{2^0}{0!} e^{-2} = 1 - e^{-2}
\end{aligned}$$

3. What is the probability that there is no arrival in $[0, 4]$?

$$\mathbf{Pr}\{N(4) - N(0) = 0\} = e^{-8}$$

4. What is the probability that the first arrival will take at least 4 minutes? Let T_1 be the arrival time of the first customer. Is T_1 a continuous or discrete random variable? Continuous.

$$\mathbf{Pr}\{T_1 > 4\} = \mathbf{Pr}\{N(4) = 0\} = e^{-8}$$

Can you understand the equality above? In plain English, "the first arrival takes at least 4 minutes" is equivalent to "there is no arrival for the first 4 minutes." It is very important duality. What if we change "4" minutes to t minutes?

$$\mathbf{Pr}\{T_1 > t\} = \mathbf{Pr}\{N(t) = 0\} = e^{-2t}$$

Surprisingly, T_1 is an exponential random variable. In fact, the times between arrivals also follows the same iid exponential distribution. We will cover this topic further after the Spring break.

Example 4.4. Let $N(t)$ be the number of arrivals in $[0, t]$. Assume that $N = \{N(t), t \geq 0\}$ is a Poisson process with rate $\lambda = 2/\text{min}$.

1. $$\mathbf{Pr}\{N(3, 7] = 5\} = \frac{(2(7-3))^5}{5!} e^{-2(7-3)} = \frac{8^5}{5!} e^{-8}$$

2. Let T_1 be the arrival time of the 1st customer.

$$\mathbf{Pr}\{T_1 > t\} = \mathbf{Pr}\{N(0, t] = 0\} = \frac{(2t)^0}{0!} e^{-2t} = e^{-2t}, \quad t \geq 0$$

What is this distribution? It is exponential distribution, meaning that T_1 follows exponential distribution. T_1 $\sim$exp with mean 0.5 minutes. Therefore,

$$\mathbf{Pr}\{T_1 \leq t\} = 1 - e^{-2t}.$$

3. Let T_3 be the arrival time of the 3rd customer. Which of the following is correct?

$$\begin{aligned}\mathbf{Pr}\{T_3 > t\} =& \mathbf{Pr}\{N(0, t] \leq 2\} = \mathbf{Pr}\{N(0, t] = 0\} + \mathbf{Pr}\{N(0, t] = 1\} + \mathbf{Pr}\{N(0, t] = 2\} \\ =& e^{-2t} + \frac{2t}{1!} e^{-2t} + \frac{(2t)^2}{2!} e^{-2t} \\ \mathbf{Pr}\{T_3 > t\} =& \mathbf{Pr}\{N(0, t] = 2\}\end{aligned}$$

The first equation is correct. Now we can compute the cdf of T_3.

$$\mathbf{Pr}\{T_3 \leq t\} = 1 - \mathbf{Pr}\{T_3 > t\} = 1 - \left(e^{-2t} + \frac{2t}{1!} e^{-2t} + \frac{(2t)^2}{2!} e^{-2t} \right)$$

Can we compute the pdf of T_3? We can take derivative of the cdf to obtain the pdf.

$$f_{T_3}(t) = \frac{2(2t)^2}{2!} e^{-2t}$$

What random variable is this? Poisson? No. Poisson is a discrete r.v. T_3 can obviously take non-integral values. It is gamma distribution. To help you distinguish different 2s here, let me use λ.

$$\frac{\lambda(\lambda t)^2}{2!} e^{-\lambda t} \sim \text{Gamma}(3, \lambda)$$

$\alpha = 3$ is the shape parameter and λ is the scale parameter. For gamma distribution, it is easy to understand when α is an integer. Even if α is not an integer, gamma distribution is still defined. However, it is tricky to compute something like (2.3)!. That's where $\Gamma_{(\alpha)}$, gamma function, comes

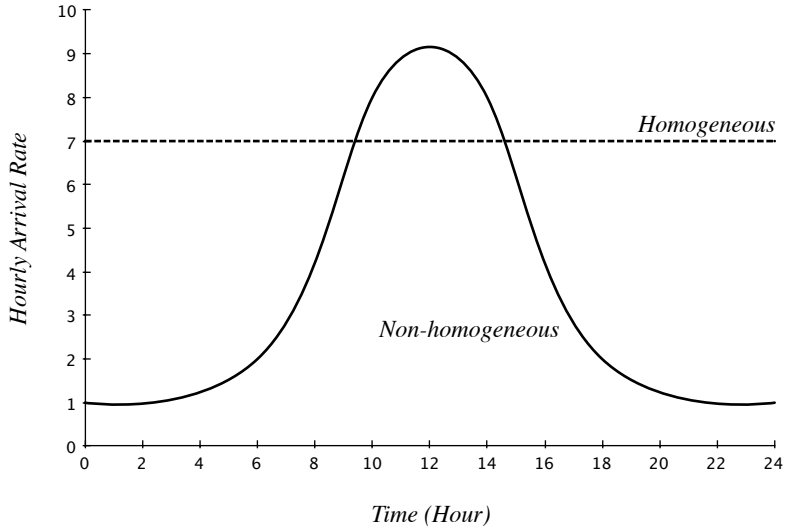
into play. You can look up the table to get the specific value for a certain α.

Now think why we get gamma distribution for T_3. Each inter-arrival time is exponential r.v. Also, they are iid. Therefore, we can compute T_3 by summing three of them. That's why we get gamma distribution.

By the way, Erlang distribution, if you have heard of them, is gamma distribution with integer α.

4.3 Non-homogeneous Poisson Process

We now move on to non-homogeneous Poisson process. λ associated with the exponential distribution of non-homogeneous Poisson process varies over time. That is, λ is a function of time t, $\lambda(t)$. Then, why do we ever need non-homogeneity which makes computation more complicated? Suppose that you are observing a hospital's hourly arrival rate. Arrival rate will not be constant over an entire day.



Therefore, to model this type of real world phenomena, we need more sophisticated model. The figure above clearly suggests that non-homogeneous Poisson process can capture and model the reality better than just homogeneous Poisson process. In other words, non-homogeneous Poisson process allows us to model a rush-hour phenomenon.

Definition 4.3 (Non-homogeneous Poisson Process)**.** $N = \{N(t), t \geq 0\}$ is said to be a time-inhomogeneous Poisson process with rate function $\lambda = \{\lambda(t), t \geq 0\}$ if

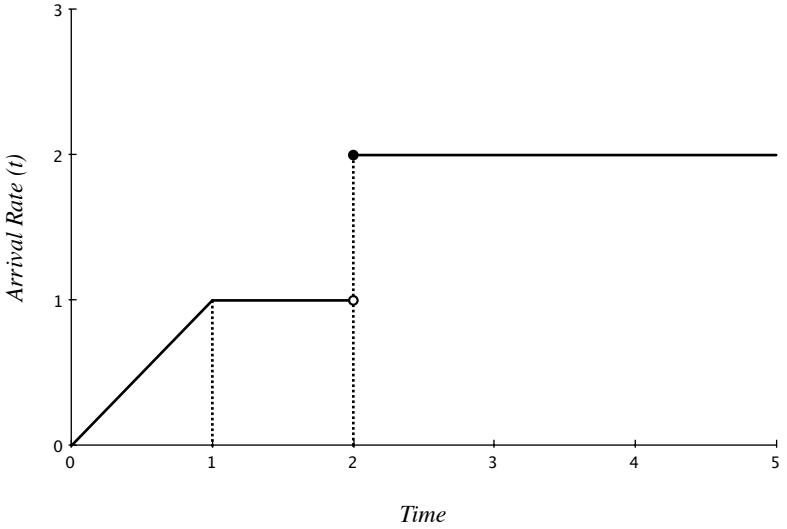
1. N has independent increments,

2.

$$\mathbf{Pr}\{N(s,t] = k\} = \frac{\left(\int_s^t \lambda(u)du\right)^k}{k!} e^{-\int_s^t \lambda(u)du},$$

where $N(s,t] = N(t) - N(s)$ is the number of arrivals in $(s,t]$.

Example 4.5. Suppose we modeled the arrival rate of a store as follows. One month after launching, the arrival rate settles down and it jumps up at the end of the second month due to discount coupon.



Assume N is a time-non-homogeneous Poisson process with rate function $\lambda = \{\lambda(t), t \geq 0\}$ given in the figure above. Then, average number of arrivals in $(0, .5]$ is $\int_0^{.5} \lambda(t)dt = (1/2)(.5)^2 = .125$.

1.

$$\begin{aligned}\mathbf{Pr}\{N(.5) \geq 2\} =& 1 - \mathbf{Pr}\{N(.5) < 2\} = 1 - \mathbf{Pr}\{N(.5) = 0\} - \mathbf{Pr}\{N(.5) = 1\}\\ \mathbf{Pr}\{N(.5) = 1\} =& \frac{.125^1}{1!} e^{-.125}\end{aligned}$$

2.

$$\begin{aligned}&\mathbf{Pr}\{\text{there are at least 1 arrival in } (0,1] \text{ and 4 arrivals in } (1,3]\}\\ =&\mathbf{Pr}\{N(0,1] \geq 1\}\mathbf{Pr}\{N(1,3] = 4\}\end{aligned}$$

by the independent increment property. Each of these can be computed as follows.

$$\begin{aligned}\mathbf{Pr}\{N(0,1] \geq 1\} =& 1 - \mathbf{Pr}\{N(0,1] = 0\} = 1 - e^{-1/2}\\ \mathbf{Pr}\{N(1,3] = 4\} =& \frac{3^4}{4!} e^{-3}\end{aligned}$$

It is important to note that you should first decompose the time intervals so that they do not overlap. Then, we can convert the probability into the product form in which we can compute each element of the product form.

4.4 Thinning and Merging

We covered most of Poisson process related topics. Some topics remaining are merging and splitting of Poisson arrival processes. Let us further develop our model for incoming customers stream. What if we have two independent streams of Poisson processes merged together and goes into a single queue? Will the merged process be also Poisson? Conversely, if there are two types of customers in a single flow and we separate the two from one stream, will the two split processes be Poisson? Understanding merging and thinning Poisson process will greatly enhance the applicability of Poisson process to reality.

4.4.1 Merging Poisson Process

Suppose that there are only two entrances in Georgia Tech. Denote the numbers of arrivals through gate A and B by N_A, N_B.

$$N_A = \{N_A(t), t \geq 0\}, \quad N_B = \{N_B(t), t \geq 0\}$$

Theorem 4.4. *Assume N_A is a Poisson process with rate λ_A. Assume N_B is a Poisson process with rate λ_B. Assume further that N_A and N_B are independent. Then, $N = \{N(t), t \geq 0\}$ is a Poisson processes with rate $\lambda = \lambda_A + \lambda_B$ where $N(t) = N_A(t) + N_B(t)$.*

The independence assumption may or may not be true. Thinking of the Georgia Tech case here, more people through gate A may mean that less people through gate B. Think about Apple's products. Would sales of iPad be correlated with sales of iPhone? Or, would it be independent? We need to be careful about independence condition when you model the real world.

4.4.2 Thinning Poisson Process

Now think about splitting a Poisson process. $N = \{N(t), t \geq 0\}$ is a Poisson process describing the arrival process. Each customer has to make a choice which of two stores they shop, A or B. To make the decision, they flip a biased coin with probability p of getting a head. Let $N_A(t)$ be the number of arrivals to store A in $(0, t]$. Let $N_B(t)$ be the number of arrivals to store B in $(0, t]$. Compose $N(t) = N_A(t) + N_B(t)$.

Theorem 4.5. *Suppose N is Poisson with rate λ. Then, N_A is a Poisson process with rate $p\lambda$ and N_B is a Poisson process with rate $(1-p)\lambda$.*

It may be silly that you shop and choose by flipping a coin. However, from a company's perspective, they view people choose based on flipping a coin using

statistical inference. Suppose a company has two products. They will model people's choice using a lot of tracking data and make a conclusion based on statistical inference.

The following is the sketch of the proof.

Proof.

$$N_A(t) \sim \text{Poisson}(\lambda p)$$
$$N_B(t) \sim \text{Poisson}(\lambda(1-p))$$

Define

$$N_A(t) = \sum_{i=1}^{N(t)} Y_i, \quad Y_i = \begin{cases} 1, & \text{if the } i\text{th toss is a head} \\ 0, & \text{if the } i\text{th toss is a tail} \end{cases}$$

where $\{Y_i\}$ iid and is independent of $N(t)$. Then,

$$\begin{aligned}
\mathbf{Pr}(N_A(t)=k) =& \frac{(\lambda pt)^k}{k!}e^{-\lambda pt} = \mathbf{Pr}(\sum_{i=1}^{N(t)} Y_i = k) \\
&= \sum_{n=k}^{\infty} \mathbf{Pr}\left(\sum_{i=1}^{N(t)} Y_i = k | N(t)=n\right) \mathbf{Pr}(N(t)=n) \\
&= \sum_{n=k}^{\infty} \mathbf{Pr}\left(\sum_{i=1}^{n} Y_i = k | N(t)=n\right) \frac{(\lambda t)^n}{n!}e^{-\lambda t} \\
&= \sum_{n=k}^{\infty} \mathbf{Pr}\left(\sum_{i=1}^{n} Y_i = k\right) \frac{(\lambda t)^n}{n!}e^{-\lambda t} \\
&= \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} \frac{(\lambda t)^n}{n!}e^{-\lambda t} \\
&= \sum_{n=k}^{\infty} \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} \frac{(\lambda t)^n}{n!}e^{-\lambda t} \\
&= e^{-\lambda t}\frac{p^k}{k!} \sum_{n=k}^{\infty} \frac{1}{(n-k)!}(1-p)^{n-k}(\lambda t)^k(\lambda t)^{n-k} \\
&= e^{-\lambda t}\frac{p^k(\lambda t)^k}{k!} \sum_{n=k}^{\infty} \frac{1}{(n-k)!}(1-p)^{n-k}(\lambda t)^{n-k} \\
&= e^{-\lambda t}\frac{p^k(\lambda t)^k}{k!} \left(1 + \frac{((1-p)\lambda t)^1}{1!} + \frac{((1-p)\lambda t)^2}{2!} + \frac{((1-p)\lambda t)^3}{3!} + \cdots\right) \\
&= e^{-\lambda t}\frac{p^k(\lambda t)^k}{k!} e^{-(1-p)\lambda t} \quad \text{using Taylor expansion} \\
&= \frac{(\lambda pt)^k}{k!}e^{-\lambda pt}
\end{aligned}$$

The difficulty arising here is that, in the summation of Y_i, Y_i and N are entangled. But, we have conditioning! In conditioning, we used the following property of conditional probability, the law of total probability.

$$\mathbf{Pr}(A) = \sum_n \mathbf{Pr}(A|B_n)\mathbf{Pr}(B_n)$$

The Taylor expansion is

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \dots.$$

Although we obtained what we wanted at first, we still need to prove the independence between N_A and N_B. To do that, we need to show

$$\mathbf{Pr}(N_A(t) = k, N_B(t) = l) = \frac{(\lambda pt)^k e^{-\lambda pt}}{k!}\frac{(\lambda pt)^l e^{-\lambda pt}}{l!}$$

which is the product form. We do not go over the computation, but it will be very similar. □

In Poisson process, independent increment is another important concept, but we will not go over that very much. FYI, as another example, if each customer does not toss a coin and instead odd number customers go to store A and even number customers go to store B, the split processes will not be Poisson.

4.5 Simulation

First, let us simulate a homogeneous Poisson process with a fixed rate 2 people per minute.

Listing 4.1: Simulating Homogeneous Poisson Process

```
# Set up arrival rate
lambda = 2

# How many arrivals will be generated?
N=100

# Generate N random inter-arrival times from exponential
    distribution
interarrivals=rexp(N, rate=lambda)
interarrivals

# Arrival timestamps (cumulative sum of inter-arrival times)
cumsum(interarrivals)
```

Simulating non-homogeneous Poisson process is substantially more complicated than simulating homogeneous Poisson process. To simulate a non-homogeneous Poisson process with intensity function $\lambda(t)$, choose a sufficiently large λ so that $\lambda(t) = \lambda p(t)$ and simulate a Poisson process with rate parameter

λ. Accept an event from the Poisson simulation at time t with probability $p(t)$.[1] Let us simulate a non-homogeneous Poisson process with rate $\lambda(t) = \sqrt{t}$.

Listing 4.2: Simulating Non-homogeneous Poisson Process

```
# Set up arrival rate as a function of time and the large constant c
lambda = function(t) {sqrt(t)}
c = 10000

# To simulate N non-homogeneous arrivals, we will need much larger
    size (N_homo > N*c) of homogeneous arrivals
N=100
N_homo = N*c*3

# Generate N_homo homogeneous Poisson arrivals
h_arrivals = cumsum(rexp(N_homo, rate=c))
head(h_arrivals)

# Collect events with probability lambda(t)/c
arrivals = c()
for (i in 1:length(h_arrivals)) {
  if (runif(1)<lambda(h_arrivals[i])/c) {
    arrivals = append(arrivals, h_arrivals[i])
    if (length(arrivals)>=N) break
  }
}
arrivals
```

Finally, we can also simulate merging and thinning Poisson processes. Suppose we are modeling the inflow of people into the Georgia Tech Atlanta campus. There are two entrances to the campus: Gate A and B. Through Gate A, people come in following a homogeneous Poisson process with rate 3. Through Gate B, arrivals form a Poisson process with rate 5. If we want to model the arrivals to the campus as a whole, we know that the merged process is a homogeneous Poisson process with rate 8 from theory. Still, we can programmatically merge two independent processes.

Listing 4.3: Merging Poisson Processes

```
# Set up two arrival rates
l1 = 3
l2 = 5

# How many arrivals will be generated?
N=100

# Simulate two N homogeneous Poisson arrivals
arrivals1 = cumsum(rexp(N, rate=l1))
arrivals2 = cumsum(rexp(N, rate=l2))

# Merge the two processes and truncate first N arrivals
merged = sort(c(arrivals1,arrivals2))[1:N]
merged
```

[1] Ross, Sheldon M. (2006). Simulation. Academic Press. p. 32.

```
# Compare with another independent homogeneous Poisson process with
    rate 8
comparison = cumsum(rexp(N, rate=l1+l2))
comparison
```

4.6 Exercise

1. Suppose there are two tellers taking customers in a bank. Service times at a teller are independent, exponentially distributed random variables, but the first teller has a mean service time of 2 minutes while the second teller has a mean of 5 minutes. There is a single queue for customers awaiting service. Suppose at noon, 3 customers enter the system. Customer A goes to the first teller, B to the second teller, and C queues. To standardize the answers, let us assume that T_A is the length of time in minutes starting from noon until Customer A departs, and similarly define T_B and T_C.
   (a) What is the probability that Customer A will still be in service at time 12:05?
   (b) What is the expected length of time that A is in the system?
   (c) What is the expected length of time that A is in the system if A is still in the system at 12:05?
   (d) How likely is A to finish before B?
   (e) What is the mean time from noon until a customer leaves the bank?
   (f) What is the average time until C starts serv
   (g) What is the average time that C is in the system?
   (h) What is the average time until the system is empty?
   (i) What is the probability that C leaves before A given that B leaves before A?
   (j) What are the probabilities that A leaves last, B leaves last, and C leaves last?
   (k) Suppose D enters the system at 12:10 and A, B, and C are still there. Let W_D be the time that D spends in the system. What is the mean time that D is in the system?
2. Suppose we agree to deliver an order in one day. The contract states that if we deliver the order within one day we receive $1000. However, if the order is late, we lose money proportional to the tardiness until we receive nothing if the order is two days late. The length of time for us to complete the order is exponentially distributed with mean 0.7 days. For notation, let T be the length of time until delivery.
   (a) What are the probabilities that we will deliver the order within one day and within two days?

(b) What is the expected tardiness?

(c) Notice that this contract is the one in the Littlefield simulation game 1. We now add two more contracts (which similarly have the rule of losing money proportional to tardiness):

- price = \$750; quoted lead time = 7 days; maximum lead time = 14 days.
- price = \$1250; quoted lead time = 0.5 day; maximum lead time = 1 days.

Notice that in the \$1000 contract, quoted lead time = 1 day; maximum lead time = 2 days. What are the expected revenues for these three contracts? Which contract would be the most lucrative assuming the mean time to delivery is 0.7 days?

(d) Suppose the mean time to delivery is exponentially distributed with mean δ days. For what values of δ is the shortest \$1250 contract optimal, for what values is the medium length \$1000 contract optimal, and for what values is the longest \$750 contract optimal?

3. Suppose passengers arrive at a MARTA station between 10am -5pm following a Poisson process with rate $\lambda = 60$ per hour. For notation, let $N(t)$ be the number of passengers arrived in the first t hours, $S_0 = 0$, S_n be the arrival time of the nth passenger, X_n be the interarrival time between the $(n-1)$st and nth passenger.

(a) What is the expected number of passengers arrived in the first 2 hours?

(b) What is the probability that no passengers arrive in the first one hour?

(c) What is the mean number of arrivals between noon and 1pm?

(d) What is the probability that exactly two passengers arrive between 2pm and 4pm?

(e) What is the probability that ten passengers arrive between 2pm and 4pm given that no customer arrive in the first half hour?

(f) What is the average time of the first arrival?

(g) What is the expected time of the thirtieth arrival?

(h) What is the expected time of the first arrival given that no passenger arrives in the first hour?

(i) What is the expected arrival time of the first passenger who arrives after 2pm?

(j) Suppose that 40% of passengers are female. What is the expected number of female passengers arrived in the first 2 hours? For notation, let $W(t)$ be the number of female passengers and $M(t)$ the number of male passengers arrive in the first t hours. Let S_n^W be

the arrival time of the nth female passenger, and X_n^W be the interarrival time between the nth and $(n-1)$st female passenger. Similarly define, S_n^M and X_n^M for the male passengers.

(k) What is the probability of at least one female passengers arrive in the first half hour given that no male passengers arrive?

(l) What is the probability that the first three passengers are all female?

4. Nortel in Canada operates a call center for customer service. Assume that each caller speaks either English or French, but not both. Suppose that the arrival for each type of calls follows a Poisson process. The arrival rates for English and French calls are 2 and 1 calls per minute, respectively. Assume that call arrivals for different types are independent.

   (a) What is the probability that the 2nd English call will arrive after minute 5?

   (b) Find the probability that, in first 2 minutes, there are at least 2 English calls and exactly 1 French call?

   (c) What is the expected time for the 1st call that can be answered by a bilingual operator (that speaks both English and French) to arrive?

   (d) Suppose that the call center is staffed by bilingual operators that speak both English and French. Let $N(t)$ be the total number of calls of both types that arrive during $(0, t]$. What kind of a process is $N(t)$? What is the mean and variance of $N(t)$?

   (e) What is the probability that no calls arrive in a 10 minute interval?

   (f) What is the probability that at least two calls arrive in a 10 minute interval?

   (g) What is the probability that two calls arrive in the interval 0 to 10 minutes and three calls arrive in the interval from 5 to 15 minutes?

   (h) Given that 4 calls arrived in 10 minutes, what is the probability that all of these calls arrived in the first 5 minutes?

   (i) Find the probability that, in the first 2 minutes, there are at most 1 call, and, in the first 4 minutes, there are exactly 3 calls?

5. Suppose customer arrive to a bank according to a Poisson process but the arrival rate fluctuates over time. From the opening time at 9 a.m. until 11, customers arrive at a rate of 10 customers per hour. From 11 to noon, the arrival rate increases linearly until it reaches 20 customers per hour. From noon to 1pm, it decreases linearly to 15 customers per hour, and remains at 15 customers per hour until the bank closes at 5 p.m. For notation, let $N(t)$ be the number of arrivals in the t hours since the bank opened and $\lambda(t)$ the arrival rate at t hours after opening.

   (a) What is the arrival rate at 12:30pm?

   (b) What is the average number of customers per day?

(c) What is the probability of k arrivals between 11:30 and 11:45?

(d) What is the probability of k arrivals between 11:30 and 11:45 given that there were 7 arrivals between 11:00 and 11:30?

(e) Consider the first customer that arrives after noon and let T be the length of time since noon to when that customer arrives. What is the probability that this customer arrives after 12:10, after 12:20? Is T exponentially distributed?

6. Calls to a call center follow a Poisson process with rate $\lambda = 2$ calls per minute. The call center has 10 phone lines. There is a single agent whose processing times are exponentially distributed mean 4 minutes. Assume that at 12 noon (time 0), there are two calls in the system; the first call is being processed, the second call is waiting, and the third call is on its way to arrive.

(a) What is the expected time that the second call leaves the system?

(b) What is the probability that the second call leaves an empty system?

(c) What is the expected time until the 3rd call leaves the system?

(d) At 12:05pm, what is the probability that the second call is the only call in the system?

7. Calls to a center follow a Poisson process with rate 120 calls per hour. Each call has probability 1/4 from a male customer. The call center opens at 9am each morning.

(a) What is the probability that there are no incoming calls in the first 2 minutes?

(b) What is the probability that there are exactly one calls in the first 2 minutes and exactly three calls from minute 1 to minute 4?

(c) What is the probability that during the first 2 minutes there are 1 call from male customer and 2 calls from female customers?

(d) What is the expected elapse of time from 9am until the second female customer arrives?

(e) What is the probability that the second call arrives after 9:06am?

8. Assume that passengers arrive at a MARTA station between 10am -12noon following a non-homogeneous Poisson process with rate function $\lambda(t)$ given as follows. From 10am-11am, the arrival rate is constant, 60 passengers per hour. From 11am to noon, it increases linearly to 120 passengers per hour.

(a) What is the probability that there is at least 1 passenger arrival in the first 2 minutes (after 10am)?

(b) What is the probability that there are 2 arrivals between 10:00am to 10:10am *and* 1 arrival between 10:05am to 10:15am?

(c) What is the expected number of arrivals between 10:50am and 11:10am?

(d) What is the probability that there are exactly 10 arrivals between 10:50am and 11:10am?

(e) What is the probability that the 1st passenger after 10am will take at least 2 minutes to arrival?

(f) What is the probability that the 1st passenger after 11am will take at least 2 minutes to arrival?

9. Assume that call arrival to a call center follows a non-homogeneous Poisson process. The call center opens from 9am to 5pm. During the first hour, the arrival rate increases linearly from 0 at 9am to 60 calls per hour at 10am. After 10am, the arrival rate is constant at 60 calls per hour.

   (a) Plot the arrival rate function $\lambda(t)$ as a function of time t; indicate clearly the time unit used.

   (b) Find the probability that exactly 5 calls have arrived by 9:10am.

   (c) What is the probability that the 1st call arrives after 9:20am?

   (d) What is the probability that there are exactly one call between 11:00am and 11:05am and at least two calls between 11:03am and 11:06am?

10. A call center is staffed by two agents, Mary and John. Mary's processing times are iid, exponentially distributed with mean 2 minutes and John's processing times are iid, exponentially distributed with mean 4 minutes.

    (a) Suppose that when you arrive, both Mary and John are busy, and you are the only one waiting in the waiting line. What is the probability that John will serve you? What is your expected time in system (waiting plus service)?

    (b) What is the probability that you will be answered in 2 minutes?

    (c) Suppose that when you arrive, you are the 2nd one in the waiting line. What is the probability that you be answered in 2 minutes?

11. Suppose you are observing an one-way road in Georgia Tech. Also, suppose there are only two types of vehicles on the road: sedan and truck. The interval between two sedans follows an iid exponential distribution with mean time 2 minute. The arrivals of trucks form a Poisson process with rate 60 per hour. The arrival of two types are independent from each other.

    (a) What is the probability that the next vehicle passing you is a truck?

    (b) What is the probability that at least three sedans will pass you in next 10 minutes?

(c) What is the probability that you will not be hit by a vehicle if you cross the road blindly? Assume that it takes 1 minute for you to cross the road with your eyes closed.

(d) What is the expected time until the next vehicle arrives regardless of the type of the vehicle?

You realized that there is another type of vehicle on the road: motorcycle. The arrivals of motorcycles is also a Poisson process but its arrival rate of a day is as follows.

Time	Arrival Rate (per hour)
12am-8am	20
8am-4pm	40
4pm-12am	30

The time is currently 7:30am. The arrival of three types of vehicles are all independent from one another. (So, for part (e)-(f), there are three types of vehicles: sedan, truck, motorcycle.)

(e) What is the probability that 30 motorcycles will pass you by 8:30am?

(f) What is the probability that 2 motorcycles will pass you between 7:30am and 8:10am and 3 motorcycles will pass you between 7:50am and 8:30am?

Chapter 5

Continuous Time Markov Chain

In Chapter 3, we learned about discrete time Markov chain. In DTMC, the time period is discretized so that time is denoted by integers. DTMC itself has wide range of applications. However, in some cases in reality, discretizing time may be too much simplifying the situation. In such cases, continuous time Markov chain comes into play. An outstanding application domain where CTMC is heavily used is queueing theory. We will cover various ways to model different queueing systems.

5.1 Introduction

Let us start with some examples.

Example 5.1. A machine goes up and down over time when it is down, it takes d_i amount of time to be repaired, where d_i is exponentially distributed with mean 1 hour. $\{d_i, i = 1, 2, 3, \dots\}$ is an iid sequence. When it is up, it will stay up u_i amount of time that is exponentially distributed with mean 10 hours. $\{u_i, i = 1, 2, 3, \dots\}$ is an iid sequence. Assume that up times and repair times are independent. Can the problem be modeled by a CTMC?

The answer is yes. What are the components of a CTMC? First, there has to be a state space, in this case $S = \{\text{up}, \text{down}\}$. Second, each state has to have corresponding holding time, in this case $\lambda_{\text{up}} = 1/10, \lambda_{\text{down}} = 1$. Last, we need a roadmap matrix R from the underlying DTMC.

$$R = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Example 5.2. We have two machines each of which having the same characteristics as before and there is only one repairperson. What is $X(t)$? Let $X(t)$ be the number of machines that are up at time t. State space $S = \{0, 1, 2\}$.

Is $X = \{X(t), t \geq 0\}$ a CTMC? What is the holding time of state 2? In the state 2, two machines are competing to fail first. The holding time is $\min(X_1, X_2) \sim \text{Exp}(1/10 + 1/10) = \text{Exp}(1/5)$. How about the holding time for state 1? At this state, the repairperson and the other operating machine are competing to finish first. Due to the memoryless property, the other alive machine is as if new. The holding time follows $\text{Exp}(1/1 + 1/10) = \text{Exp}(11/10)$. At the end of state 2, what would be the next state? This is where the roadmap matrix comes into play.

$$R_{1,2} = \frac{1}{1 + 1/10}, \quad R_{1,0} = \frac{1/10}{1 + 1/10}$$

You can compute the entire R matrix based on similar logic.

Suppose we have two machines and one repairman. Up times and down times follow the following distributions.

$$\text{Up times} \sim \text{Exp}(1/10)$$
$$\text{Down times} \sim \text{Exp}(1)$$

Let $X(t)$ be the number of machines that are up at time t. Since $X(t)$ is a stochastic process, I am going to model it using CTMC. For CTMC, we need three ingredients. First, *state space* $S = \{0, 1, 2\}$. Second, *holding times* $\lambda_0, \lambda_1, \lambda_2$. λ_i can be interpreted as how frequently the chain jumps from state i. Let us compute $\lambda_i, i \in S = \{0, 1, 2\}$.

- When $X(t) = 0$, the repairman is working on one of two machines both of which are down at the moment, so $\lambda_0 = 1$.
- When $X(t) = 1$, the holding time follows $\min(\text{up time}, \text{down time}) = \text{Exp}((1/10) + 1)$, so $\lambda_1 = 11/10$.
- When $X(t) = 2$, the holding time in this case follows $\min(\text{up time}, \text{up time}) = \text{Exp}(0.1 + 0.1)$, so $\lambda_2 = 1/5$.

Now we know when the chain will jump, but we don't know to which state the chain will jump. *Roadmap matrix*, the last ingredient of a CTMC, tells us the probability.

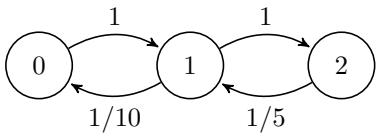
$$R = \begin{matrix} 0 \\ 1 \\ 2 \end{matrix} \begin{pmatrix} 0 & 1 & 0 \\ (1/10)/(1 + 1/10) & 0 & 1/(1 + (1/10)) \\ 0 & 1 & 0 \end{pmatrix}$$

We have specified the input data parameters we need to model a CTMC.

Let us introduce the concept of *generator matrix* which is somewhat more convenient than having holding times and roadmap matrix separately.

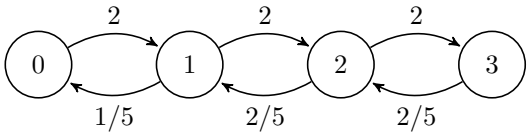
$$G = \begin{matrix} 0 \\ 1 \\ 2 \end{matrix} \begin{pmatrix} -\lambda_0 & \lambda_0 R_{01} & \lambda_0 R_{02} \\ \lambda_1 R_{10} & -\lambda_1 & \lambda_1 R_{12} \\ \lambda_2 R_{20} & \lambda_2 R_{21} & -\lambda_2 \end{pmatrix} = \begin{pmatrix} -1 & 1 & 0 \\ 1/10 & -11/10 & 1 \\ 0 & 1/5 & -1/5 \end{pmatrix}$$

We can also draw a rate diagram showing these data graphically. If you are given the following diagram, you should be able to construct G from it.



If you look at the diagram, you may be able to interpret the problem more intuitively.

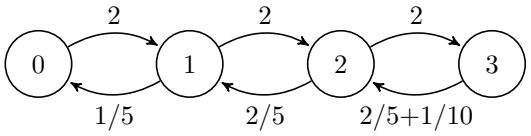
Example 5.3. We have two operators and three phone lines. Calls arrival follows a Poisson process with rate $\lambda = 2$ calls/minute. Each call processing time is exponentially distributed with mean 5 minutes. Let $X(t)$ be the number of calls in the system at time t. In this example, we assume that calls arriving when no phone line is available are lost. What would the state space be? $S = \{0, 1, 2, 3\}$.



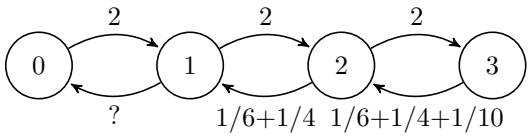
How do we compute the roadmap matrix based on this diagram? For example, you are computing R_{21}, R_{23}. Simply just add up all rates leaving from state 2 and divide the rate to the destination state by the sum.

$$R_{23} = \frac{2}{2 + (2/5)}, \quad R_{21} = \frac{2/5}{2 + (2/5)}$$

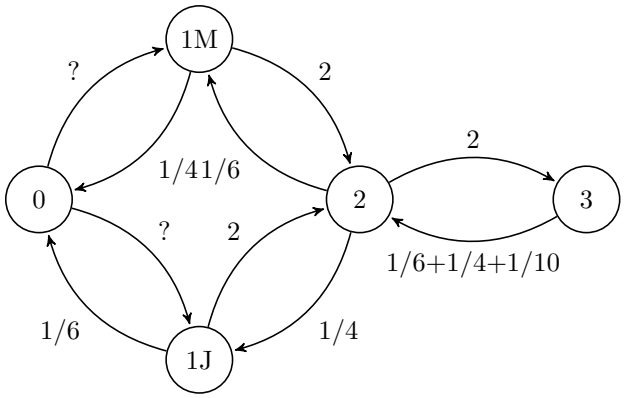
Now, let us add another condition. If each customer has a patience that is exponentially distributed with mean 10 minutes. When the waiting time in the queue of a customer exceeds his patience time, the customer abandons the system without service. Then, the only change we need to make is the rate on the arrow from state 3 to state 2.



Example 5.4. John and Mary are the two operators. John's processing times are exponentially distributed with mean 6 minutes. Mary's processing times are exponentially distributed with mean 4 minutes. Model this system by a CTMC. What would the state space be? $S = \{0, 1J, 1M, 2, 3\}$. Why can't we use $S = \{0, 1, 2, 3\}$? Let's see and let the question be open so far. Let us draw the diagram first assuming that $S = \{0, 1, 2, 3\}$.

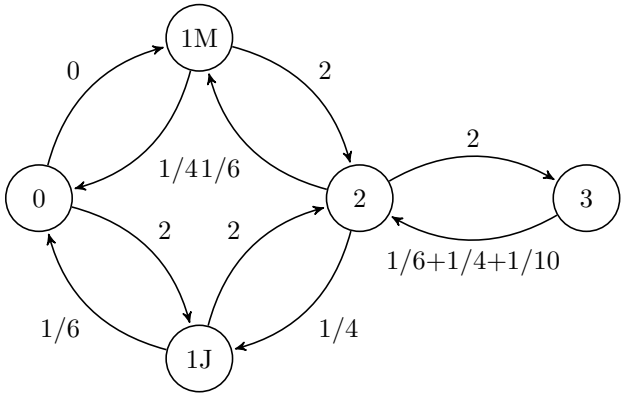


We cannot determine the rate from state 2 to state 1 because we don't know who is processing the call, John or Mary. So, we cannot live with $S = \{0, 1, 2, 3\}$. For Markov chain, it is really an important concept that we don't have to memorize all the history up to the present. It's like "Just tell me the state. I will tell you how it will evolve." Then, let's see if $S = \{0, 1J, 1M, 2, 3\}$ works.



Even in this case, we cannot determine who takes a call when the call arrives when both of them are free. It means that we do not have enough specification to model completely. In tests or exams, you will see more complete description. It is part of manager's policy. You may want John takes the call when both are free. You may toss a coin whenever such cases happen. What would the optimal policy be in this case? It depends on your objective. Probably, John and Mary are not paid same. You may want to reduce total labor costs or the average waiting time of customers.

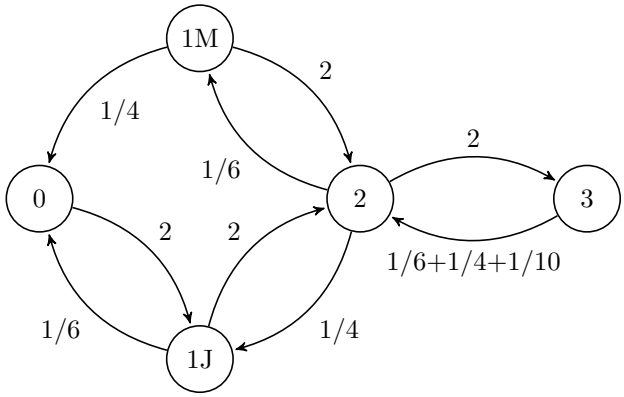
Suppose now that we direct every call to John when both are free.



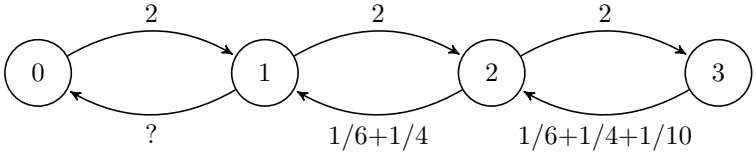
We have complete information for modeling a CTMC. These are inputs. Then, what would outputs we are interested in be? We want to know which fraction of time the chain stays in a certain state in the long run. We also want to know how many customers are lost. We may plan to install one more phone line and want to evaluate the improvement by new installation. We will cover these topics from next time.

Example 5.5. A small call center with 3 phone lines and two agents, Mary and John. Call arrivals follow a Poisson process with rate $\lambda = 2$ calls per minute. Mary's processing times are iid exponential with mean 4 minutes. John's processing times are iid exponential with mean 6 minutes. Customer's patience follows iid exponential with mean 10 minutes. An incoming call to an empty system always goes to John.

The rate diagram will be as follows.



It is very tempting to model this problem like the following. It is wrong. You won't get any credit for this if you model like this in test.



This model is not detail enough to capture all situations explained in the question. You cannot determine what number should go into the rate from state 1 to 0 because you did not take care of who is handling the phone call if there is only one.

Let us think about the formal definition of CTMC.

Definition 5.1 (Continuous-Time Markov Chain)**.** Let S be a discrete space. (It is called the state space.) For each state $i \in S$, let $\{u_i(j), j = 1, 2, \dots\}$ be a sequence of iid r.v.'s having exponential distribution with rate λ_i and $\{\phi_i(j), j = 1, 2, \dots\}$ be a sequence of iid random vectors.

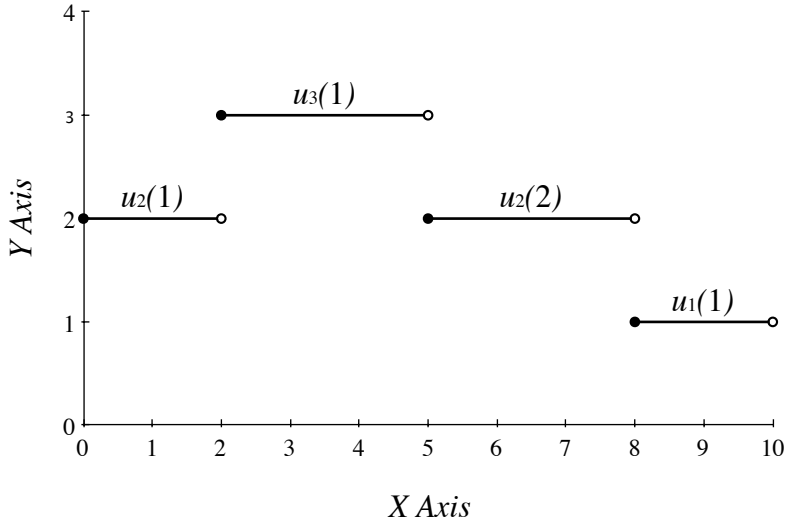
Also, as in DTMC, CTMC has the similar Markov property in a subtly different mathematical form.

$$\begin{aligned}&\mathbf{Pr}\{X(t+s)=j \mid X(t_0)=i_0, X(t_1)=i_1, \ldots, X(t_{n-1})=i_{n-1}, X(t)=i\}\\ =&\mathbf{Pr}\{X(t+s)=j \mid X(t)=i\} = \mathbf{Pr}\{X(s)=j \mid X(0)=i\} = P_{ij}(s)\end{aligned}$$

for any $t_0 < t_1 < \cdots < t_{n-1} < t$ and any $i_0, i_1, \ldots, i_{n-1} \in S$ and any $i, j \in S$. This is the defining property of a CTMC. As in DTMC, Kolmogorov-Chapman equation holds

$$P_{ij}(s) = \mathbf{Pr}\{X(s)=j \mid X(0)=i\}.$$

For example, $S = \{1, 2, 3\}$. $\phi_i(j)$ takes vector in $(1,0,0), (0,1,0), (0,0,1)$. Think you are throwing a three-sided die and choose one of three possible vectors for the next ϕ_i value.



Think of $\phi_i(j)$ as the outcome of the jth toss of the ith coin. If we have the roadmap matrix,

$$R = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/4 & 0 & 3/4 \\ 1/8 & 7/8 & 0 \end{pmatrix},$$

each row i represent ith coin. In this case, the first coin is a fair coin but the other two are biased. Now we can generate a series of vectors $\phi_i(j)$.

5.1.1 Holding Times

Then, how can we generate iid *exponentially* distributed random variables, $u_i(j)$? Computer is only able to drop a needle within an interval. Say $[0, 1]$. We can ask computer to drop a needle between 0 and 1. How can we generate an

exponentially distributed random variable from this basic building block? Say we are trying to generate exponential random variables with rate 4. Define

$$X = -\frac{1}{4}\ln(1-U).$$

Then,

$$\begin{aligned}
\mathbf{Pr}\{X > t\} =& \mathbf{Pr}\left\{-\frac{1}{4}(1-U) > t\right\} \\
=& \mathbf{Pr}\{\ln(1-U) < -4t\} = \mathbf{Pr}\left\{(1-U) < e^{-4t}\right\} \\
=& \mathbf{Pr}\left\{U > 1 - e^{-4t}\right\} \\
=& 1 - \left\{1 - e^{-4t}\right\} \\
=& e^{-4t}.
\end{aligned}$$

We got the exponential distribution we initially wanted by just using a uniform random variable. You will learn how to generate other types of random variables from a uniform distribution in simulation course. That will be the main topic there. Simulation in fact is an integral part of IE curriculum.

5.1.2 Generator Matrix

Let us examine the relationship between roadmap matrix at time t and generator matrix. Generator matrix is the derivative of transition matrix at time 0.

$$\begin{aligned}
P(t) =& (P_{ij}(t)) \\
P'_{1,1}(0) =& -\lambda_1 = G_{1,1} \\
P'_{i,j}(0) =& G_{ij}
\end{aligned}$$

Remember Kolmogorov-Chapman equation? Differentiate with respect to s.

$$\begin{aligned}
&P(t+s) = P(t)P(s) \quad \text{for any } t, s \geq 0 \\
&\frac{d}{ds}P(t+s)\bigg|_{s=0} = P'(t) = P(t)P'(0) = P(t)G, \quad t \geq 0 \\
&P(0) = I \quad P(t) = e^{tG} = \text{expm}(tG)
\end{aligned}$$

In other words, we can *generate* $P(t)$ for any t using the generator matrix G. That is why G is called *generator* matrix.

In reality, we can compute `expm` without Matlab only in a few cases. In such special case, you can first obtain eigenvalues of G matrix using the following formula.

$$GV_1 = a_1V_1, \quad GV_2 = a_2V_2, \quad GV_3 = a_3V_3$$

If all eigenvalues are distinct, we can exponentiate the matrix rather easily.

$$GV = V \begin{pmatrix} a_1 & & \\ & a_2 & \\ & & a_3 \end{pmatrix} \quad \Rightarrow \quad G = V \begin{pmatrix} a_1 & & \\ & a_2 & \\ & & a_3 \end{pmatrix} V^{-1}$$
$$\therefore G = V \begin{pmatrix} a_1 & & \\ & a_2 & \\ & & a_3 \end{pmatrix} V^{-1}, \quad \ldots, \quad G^n = V \begin{pmatrix} a_1^n & & \\ & a_2^n & \\ & & a_3^n \end{pmatrix} V^{-1}$$

Hence,

$$\begin{aligned} \text{expm}(G) &= \sum_{n=0}^{\infty} \frac{G^n}{n!} \\ &= V \begin{pmatrix} \sum_{n=0}^{\infty} \frac{a_1^n}{n!} & & \\ & \sum_{n=0}^{\infty} \frac{a_2^n}{n!} & \\ & & \sum_{n=0}^{\infty} \frac{a_3^n}{n!} \end{pmatrix} V^{-1} \\ &= V \begin{pmatrix} e^{a_1} & & \\ & e^{a_2} & \\ & & e^{a_3} \end{pmatrix} V^{-1}. \end{aligned}$$

Let us run Matlab experiment.

```
>> G=[-2 1 1; 2 -5 3; 2 2 -4]
>> expm(2*G)
>> expm(5*G)
>> exp(5*G)
```

When you compute e^{5G}, you will see all rows identical. It seems like the chain reaches steady-state. Also, if you look at the result from `exp(5*G)` command, it is just completely wrong for our purpose.

Now we can answer the questions raised at the beginning of the class.

$$\mathbf{Pr}\{X(2) = 3 \mid X(0) = 1\} = .2856$$
$$\mathbf{Pr}\{X(5) = 2, X(7) = 3 \mid X(0) = 1\} = (.2143)(.2858) = 0.06124694$$

Since you are not allowed to use Matlab in test, you will be given a matrix such as expm(G).

Example 5.6. Let X be a CTMC on state space $S = \{1, 2, 3\}$ with generator

$$G = \begin{pmatrix} -2 & 1 & 1 \\ 2 & -5 & 3 \\ 2 & 2 & -4 \end{pmatrix}.$$

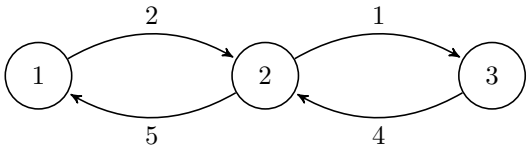
1. Find $\mathbf{Pr}(X(2) = 3 | X(0) = 1)$.

$$\mathbf{Pr}(X(2) = 3 | X(0) = 1) = e^{2G}_{1,3}$$

2. Find $\mathbf{Pr}(X(5) = 2, X(7) = 3|X(0) = 1)$.

$$\begin{aligned}&\mathbf{Pr}(X(5) = 2, X(7) = 3|X(0) = 1)\\ =&\mathbf{Pr}(X(5) = 2|X(0) = 1)\mathbf{Pr}(X(7) = 3|X(5) = 2)\\ =&\mathbf{Pr}(X(5) = 2|X(0) = 1)\mathbf{Pr}(X(2) = 3|X(0) = 2) = e^{5G}_{1,2}e^{2G}_{2,3}\end{aligned}$$

Example 5.7. Let $X = \{X(t), t \geq 0\}$ be a CTMC on state space $S = \{1, 2, 3\}$ with the following rate diagram.



The generator matrix is

$$G = \begin{pmatrix} -2 & 2 & 0 \\ 5 & -6 & 1 \\ 0 & 4 & -4 \end{pmatrix}.$$

Suppose that you are asked to compute $P_{1,3}(10)$ meaning the probability going from state 1 to state 3 after 10 minutes. Using Kolmogorov-Chapman equation,

$$P_{1,3}(10) = \sum_k P_{1,k}(5)P_{k,3}(5) = [P(5)]^2_{1,3}.$$

You can compute $P_{1,3}(10)$ given that you are given $P_{1,3}(5)$. How about $P_{1,3}(1)$ or $P_{1,3}(1/10)$? We can still compute $P_{1,3}(10)$ by exponentiating these to the 10th power or 100th power. In this way, we can have the following approximation formula.

$$P(t) = P'(0)t + P(0) = P'(0)t + I$$

What we are really talking about is the derivative of the matrix.

$$P(t) = e^{tG}$$

What are we talking about? Exponentiating a matrix? There are two commands in Matrax relevant to exponentiating a matrix.

```
>> exp(A)
[e1 e2; e3 e4]
>> expm(A)
This is what we want.
```

$$\text{expm}(A) = \sum_{n=0}^{\infty} \frac{A^n}{n!}$$

When you ask the calculator

This is called transient

5.2 Stationary Distribution

By definition, stationary distribution of CTMC means that if you start with this distribution you will not deviate from it forever. For example, if $(a, b, 1-a-b)$ is the stationary distribution of a CTMC $X(t)$ and $X(0) = (a, b, 1-a-b)$. Then, $X(5) = X(10.2) = (a, b, 1-a-b)$.

How can we compute the stationary distribution without using computer. As in DTMC, we might want to use $\pi P = \pi$. But, in this case, $P(t)$ can change over time. Which $P(t)$ should we use? In addition, it is usually hard to compute $P(t)$ from generator matrix without using a computer. We have to come up with the way to compute the stationary distribution only with the generator. Since $\pi = \pi P(t)$ should hold for all $t \geq 0$, if we take derivative of both sides with respect to t and evaluate it at $t = 0$, then

$$0 = \pi P'(0) \quad \Rightarrow \quad \pi G = 0.$$

In our case,

$$(\pi_1, \pi_2, \pi_3) \begin{pmatrix} -2 & 1 & 1 \\ 2 & -5 & 3 \\ 2 & 2 & -4 \end{pmatrix} = 0.$$

A theoretic parts are involved here.

1. Because X is irreducible, it has at most one stationary distribution.
2. Because S is finite, it has at least one stationary distribution.

Therefore, we have one unique stationary distribution in our case.

Example 5.8. Consider a call center with two homogeneous agents and 3 phone lines. Arrival process is Poisson with rate $\lambda = 2$ calls per minute. Processing times are iid exponentially distributed with mean 4 minutes.

1. What is the long-run fraction of time that there are no customers in the system? π_0
2. What is the long-run fraction of time that both agents are busy? $\pi_2 + \pi_3$
3. What is the long-run fraction of time that all three lines are used? π_3

Answer: $X(t)$ is the number of calls in the system at time t. $S = \{0, 1, 2, 3\}$. We can draw the rate diagram based on this information. In fact, having the rate diagram is equivalent to having the generator matrix. When we solve $\pi G = 0$, it is really just solving flow balancing equations, flow in = flow out in each state.

$$2\pi_0 = \frac{1}{4}\pi_1, \quad 2\pi_1 = \frac{1}{2}\pi_2, \quad 2\pi_2 = \frac{1}{2}\pi_3, \quad \pi_0 + \pi_1 + \pi_2 + \pi_3 = 1$$

Solving this by setting $\pi_0 = 1$ and normalizing the result, we obtain

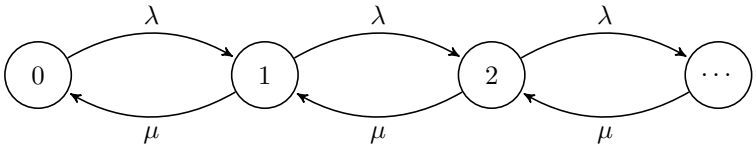
$$\pi = (1, 8, 32, 128) \quad \Rightarrow \quad \pi = \left(\frac{1}{169}, \frac{8}{169}, \frac{32}{169}, \frac{128}{169}\right).$$

Your manager may also be interested in other performance measures.

1. The number of calls lost per minute is $\lambda\pi_3 = 2(128/169)$ which seems to be quite high.
2. The throughput of the system is $\lambda(1 - \pi - 3)$.

5.3 M/M/1 Queue

Suppose we have an M/M/1 queue, meaning that we have Poisson arrival process with rate λ arrivals/minute and service times are iid exponentially distributed with rate μ. To illustrate the point, set $\lambda = 1/2, \mu = 1$. Assume that the buffer size is infinite. Let $X(t)$ be the number of customers in the system at time t. Then, $X = \{X(t), t \geq 0\}$ is a CTMC with state space $S = \{0, 1, 2, \dots\}$. What is the easiest way to model this as a CTMC? Draw the rate diagram.



Is this CTMC irreducible? Yes. Does it have a stationary distribution? Yes or no. It depends on the relationship between λ and μ. What if $\lambda > \mu$? The queue will eventually be fed up to infinity. In such case, we don't have a stationary distribution. If $\lambda < \mu$, we will have a unique stationary distribution. Even if $\lambda = \mu$, we won't have a stationary distribution. We will look into this later.

How can we determine the stationary distribution? We can get one by using $\pi G = 0$, but let us first try the cut method. If we cut states into two groups, in steady state, flow going out from and in to one group should equate. Therefore,

$$\begin{aligned}
\pi_0\lambda &= \pi_1\mu \quad \Rightarrow \pi_1 = \rho\pi_0 \\
\pi_1\lambda &= \pi_2\mu \quad \Rightarrow \pi_2 = \rho\pi_1 \\
\pi_2\lambda &= \pi_3\mu \quad \Rightarrow \pi_3 = \rho\pi_2 \\
&\vdots
\end{aligned}$$

where $\rho = \lambda/\mu = 0.5$ in this case. Solving the system of equations, you will get the following solution.

$$\begin{aligned}
\pi_1 &= \rho\pi_0 \\
\pi_2 &= \rho^2\pi_0 \\
\pi_3 &= \rho^3\pi_0 \\
&\vdots
\end{aligned}$$

The problem is that we don't know what π_0 is. Let us determine π_0 intuitively first. If server utilization is ρ, it means that in the long run the server is not

busy for $1-\rho$ fraction of time. Therefore, $\pi_0 = 1-\rho$. We can get $\pi_i, \forall i$ from $\rho^i \pi_0$. We can solve this analytically as well. Remember the sum of stationary distribution should be 1.

$$\begin{aligned}
\pi_0 + \pi_1 + \pi_2 + \cdots &= 1 \\
\pi_0 + \rho\pi_0 + \rho^2\pi_0 + \cdots &= 1 \\
\pi_0(1 + \rho + \rho^2 + \dots) &= 1 \\
\pi_0 \left(\frac{1}{1-\rho}\right) &= 1 \quad \Rightarrow \quad \pi_0 = 1-\rho
\end{aligned}$$

We expected this. More concretely,

$$\pi_2 = \left(\frac{1}{2}\right)^2 \left(1 - \frac{1}{2}\right) = \frac{1}{8} = 0.125.$$

What does this π_2 mean? It means that 12.5% of time the system has two customers.

In general, we can conclude that the CTMC has a stationary distribution if and only if $\rho < 1$. Let us assume $\rho < 1$ for further examples. As a manager, you will want to know more than long run fraction of time how many customers you have. You may want to know the long run average number of customers in the system.

$$\begin{aligned}
\sum_{n=0}^{\infty} n\pi_n &= 0\pi_0 + 1\pi_1 + 2\pi_2 + \dots \\
&= 1\rho(1-\rho) + 2\rho^2(1-\rho) + 3\rho^3(1-\rho) + \dots \\
&= \rho(1-\rho)(1 + 2\rho + 3\rho^2 + \dots) \\
&= \rho(1-\rho)(1 + \rho + \rho^2 + \rho^3 + \dots)' \\
&= \rho(1-\rho)\left(\frac{1}{1-\rho}\right)' \\
&= \rho(1-\rho)\frac{1}{(1-\rho)^2} = \frac{\rho}{1-\rho}
\end{aligned}$$

Next question you may wonder is what the average time in the system will be. Can we use the Little's Law, $L = \lambda W$? What do we know among these three variables?

$$\begin{aligned}
L &= \frac{\rho}{1-\rho} = \lambda W \\
W &= \frac{1}{\lambda}\frac{\rho}{1-\rho} = \frac{1}{\lambda}\frac{\lambda/\mu}{1-\rho} = \frac{1}{\mu}\frac{1}{1-\rho} = \frac{m}{1-\rho}
\end{aligned}$$

where m is the mean processing time. A little bit tweaked question will be what the average waiting time in the queue. We can again use the Little's Law as

long as we define the boundary of our system correctly. It should be $L_q = \lambda W_q$. How do we compute L_q, W_q?

$$\begin{aligned} L_q =& 0\pi_0 + 0\pi_1 + 1\pi_2 + 2\pi_3 + \dots \\ W_q =& W - m = \frac{m}{1-\rho} - m = m\frac{\rho}{1-\rho} \end{aligned}$$

Is this W_q formula familiar? Recall the Kingman's formula.

$$W = m\frac{\rho}{1-\rho}\left(\frac{c_a^2 + c_s^2}{2}\right)$$

In our case here, since both arrival and processing times are iid exponentially distributed, $c_a^2 = c_s^2 = 1$. That's why we did not have c_a, c_s terms in our original formula.

You should be able to see the connections among the topics covered during this semester.

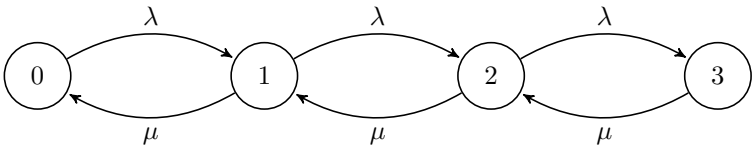
If you look at $W_q = m/(1-\rho)$, you will notice that $W_q \to \infty$ as $\rho \uparrow 1$. If $\rho = 1/2$, $W_q = m$, which means that the length of time you wait is equal to the length of time you get serviced. If $\rho = 0.95$, $W_q = 19m$. It means that you wait 19 times longer than you actually get service. It is extremely poor service. You, as a manager, should achieve both high utilization and short waiting time. Using a pool of servers, you can achieve both.

5.4 Variations of M/M/1 Queue

5.4.1 M/M/1/b Queue

Let us limit our queue size finite. Let me explain the notation here. First "M" means Poisson arrival process. In fact, Poisson arrival assumption is not very restrictive in reality. Many many empirical studies validate that a lot of human arrivals form a Poisson process. Second "M" means exponentially distributed service times. If we use "D" instead of "M", it means deterministic service times. Factory robots' service times are usually deterministic. Third "1" denotes the number of servers, in this case a single server. The last "b" means the number of spaces in the system. If you remember "G/G/1" queue taught earlier in the semester, "G" means a general distribution. If you think you have log-normal service times, we did not learn how to solve the queue analytically. One thing we can do is computer simulation. By the way, log-normal means that $\ln(X) \sim N(\mu, \sigma^2)$.

Take an example of limited size queueing system. For example, say $b = 2$. b is the maximum number of customers that can wait in the queue. The rate diagram will be as follows.



We still can use the detailed balance equations. Suppose $\lambda = 1/2, \mu = 1$ as in the previous example.

$$\pi_0 = \frac{1}{1+\rho+\rho^2+\rho^3} = \frac{1}{1+1/2+(1/2)^2+(1/2)^3} = \frac{8}{15}$$
$$\pi_1 = \frac{4}{15}, \quad \pi_2 = \frac{2}{15}, \quad \pi_3 = \frac{1}{15}$$

The questions you will be interested are

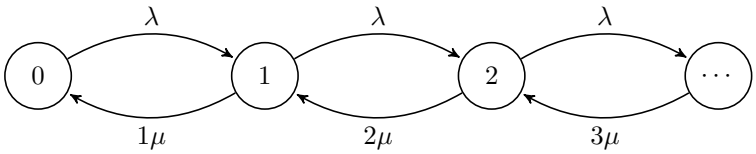
1. What is the average number of customers in the system?
$$0\frac{8}{15} + 1\frac{4}{15} + 2\frac{2}{15} + 3\frac{1}{15} = \frac{4+4+3}{15} = \frac{11}{15}$$

2. What is the average number of customers in the queue?
$$0\frac{8}{15} + 0\frac{4}{15} + 1\frac{2}{15} + 2\frac{1}{15} = \frac{4}{15}$$

3. What is average waiting time in queue? Again, we will use the Little's Law. This formula is the one you remember 10 years from now like "There was something called the Little's Law in IE curriculum."
$$L_q = \lambda W_q \quad \Rightarrow \quad \frac{4}{15} = \frac{1}{2} W_q \quad \Rightarrow \quad W_q = \frac{8}{15}$$
Is this correct? What is suspicious? We just used $\lambda = 1/2$, but in limited size queue case we lose some customers. We should use effective arrival rate, which excludes blocked customers.
$$L_q = \lambda_{\text{eff}} W_q = \lambda(1-\pi_3) W_q$$
This will generate correct W_q.

5.4.2 M/M/∞ Queue

Arrival process is Poisson with rate λ. Service times are iid exponentially distributed with rate μ. Let $X(t)$ be the number of customers in system at time t. $X = \{X(t), t \geq 0\}$ is a CTMC in $S = \{0, 1, 2, \dots\}$. Google's servers may be modeled using this model. In reality, you will never have infinite number of servers. However, Google should have so many servers that we can assume they have infinite number of servers. Model is just for answering a certain question. What would be the rate diagram?



How can we compute the stationary distribution? Use balance equations and the cut method.

$$\begin{aligned}
\lambda\pi_0 = \mu\pi_1 \quad &\Rightarrow \quad \pi_1 = \frac{\lambda}{\mu}\pi_0 \\
\lambda\pi_1 = 2\mu\pi_2 \quad &\Rightarrow \quad \pi_2 = \frac{\lambda^2}{2\mu^2}\pi_0 \\
\lambda\pi_2 = 3\mu\pi_3 \quad &\Rightarrow \quad \pi_3 = \frac{\lambda^3}{3!\mu^3}\pi_0 \\
&\vdots \\
&\pi_n = \frac{\lambda^n}{n!\mu^n}\pi_0
\end{aligned}$$

Since we have another condition, $\sum_{i=0}^{\infty} \pi_i = 1$,

$$\begin{aligned}
\sum_{i=0}^{\infty} \pi_i &= \pi_0 \left[1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{2\mu^2} + \frac{\lambda^3}{3!\mu^3} + \dots\right] = 1 \\
\pi_0 &= \frac{1}{1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{2\mu^2} + \frac{\lambda^3}{3!\mu^3} + \dots} = \frac{1}{e^{\lambda/\mu}} = e^{-\frac{\lambda}{\mu}} \\
\pi_n &= \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n e^{-\frac{\lambda}{\mu}}, \quad n = 0, 1, 2, \dots .
\end{aligned}$$

Thinking of $\pi = (\pi_0, \pi_1, \pi_2, \dots)$, what is the distribution? It is clearly not exponential because it is discrete. $\pi \sim$ a Poisson distribution. Compare it with M/M/1 queue. M/M/1 queue's stationary distribution is geometric.

You may be not fuzzy about Poisson process, Poisson random variable, etc. Poisson process tracks a series of incidents. At any given time interval, the number of incidents is a Poisson random variable following Poisson distribution.

Suppose $\lambda = 10, \mu = 1$. Then, for example,

$$\pi_2 = \frac{10^2 e^{-10}}{2!} = 50e^{-10}$$

which is quite small. The long run average number of customers in system is

$$0\pi_0 + 1\pi_1 + 2\pi_2 + \cdots = \frac{\lambda}{\mu}.$$

Capacity provisioning: You may want to achieve a certain level of service capacity. Suppose you have actually 100 servers. Then, you need to compute $\mathbf{Pr}\{X(\infty) > 100\}$. If $\mathbf{Pr}\{X(\infty) > 100\} = 10\%$, it may not be acceptable because I will be losing 10% of customers. How can we actually compute this number?

$$\begin{aligned}\mathbf{Pr}\{X(\infty) > 100\} =& 1 - \mathbf{Pr}\{X(\infty) \leq 100\} \\ =& 1 - [\mathbf{Pr}\{X(\infty) = 0\} + \mathbf{Pr}\{X(\infty) = 1\} + \mathbf{Pr}\{X(\infty) = 2\} + \dots]\end{aligned}$$

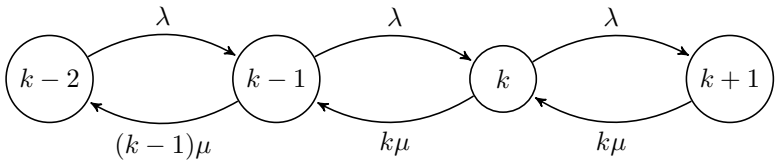
You can also compute how many servers you will need to limit the loss probability to a certain level. Say you want to limit the loss probability to 2%. Solve the following equation to obtain c.

$$\mathbf{Pr}\{X(\infty) > c\} = 0.02$$

This is an approximate way to model a many server situation. We will go further into the model with finite number of servers.

5.4.3 M/M/k Queue

In this case, we have only k servers instead of infinite number. Assume that the butter size is infinite. Then, the rate diagram will look like the following.



Solving the following system of linear equations,

$$\begin{aligned}\pi_1 =& \frac{\lambda}{\mu}\pi_0 \\ \pi_2 =& \left(\frac{\lambda}{\mu}\right)^2 \frac{1}{2!}\pi_0 \\ &\vdots \\ \pi_k =& \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}\pi_0,\end{aligned}$$

Let $\rho = \lambda/(k\mu)$. Then,

$$\pi_0 = \frac{1}{1 + \frac{\lambda}{\mu} + \frac{1}{2!}\left(\frac{\lambda}{\mu}\right)^2 + \frac{1}{3!}\left(\frac{\lambda}{\mu}\right)^3 + \dots + \frac{1}{(k-1)!}\left(\frac{\lambda}{\mu}\right)^{k-1} + \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k \frac{1}{1-\rho}}.$$

The probability that a customer will wait when they arrive is

$$\mathbf{Pr}(X(\infty) \geq k) = \frac{1}{1-\rho}\pi_k = \frac{1}{1-\rho}\left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}\pi_0.$$

As a manager managing this call center, you want to keep it reasonably low. Remember that the average waiting time in M/M/1 queue is

$$\mathbf{E}[W] = m\left[\frac{\rho}{1-\rho}\right] = mw.$$

Let us call w waiting time factor. This factor is highly nonlinear. Don't try to push up the utilization when you are already fed up. Conversely, a little more carpooling can dramatically improve traffic condition. What is it for this M/M/k case? Let us first compute the average queue size.

$$\begin{aligned}\mathbf{E}[Q] =& 1\pi_{k+1} + 2\pi_{k+2} + 3\pi_{k+3} + \dots \\ =& 1\rho\pi_k + 2\rho^2\pi_k + 3\rho^3\pi_k + \dots \\ =& \pi_k \frac{\rho}{(1-\rho)^2}\end{aligned}$$

Using Little's Law, we can compute average waiting time.

$$\mathbf{E}[W] = \frac{1}{\lambda}\frac{\rho}{(1-\rho)^2}\pi_k$$

If you use many servers parallel, I can achieve both quality and efficient service. If the world is so deterministic that arrival times and service times are both deterministic, you can achieve quality service of high utilization even with a single server. For example, every two minutes a customer arrives and service time is slightly less than 2 minutes. There will be no waiting even if you have only one server. In reality, it is full of uncertainty. In this case, having a large system will make your system more robust to uncertainty.

Sometimes, you will hear automated message when you call to a call center saying that there are 10 customers are ahead of you. This information is very misleading. Your actual waiting time will heavily depend on how many servers the call center hired.

You can analyze M/M/k/b queue in similar way. The only difference is to use $1 + \rho + \rho^2 + \cdots + \rho^b$ instead of $1 + \rho + \rho^2 + \dots$.

5.5 Open Jackson Network

5.5.1 M/M/1 Queue Review

Before going into Jackson Network, let us review M/M/1 queue first. In an M/M/1 queue, suppose the arrival rate $\alpha = 2$ jobs per minute. The mean processing time $m = 0.45$ minutes. The traffic intensity is computed as follows.

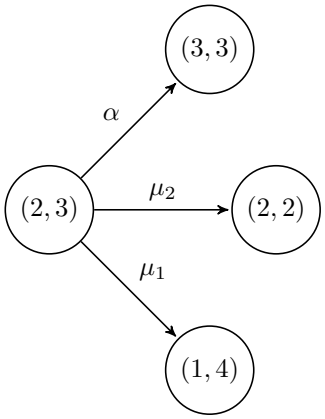
$$\rho = \alpha m = 2(0.45) = 0.9$$

Let $X(t)$ be the number of jobs in system at time t. The stationary distribution is

$$\begin{aligned}\pi_n =& \mathbf{Pr}\{X(\infty) = n\} = (1-\rho)\rho^n, \quad n = 0, 1, 2, 3, \ldots \\ \pi_2 =& (1-\rho)\rho^2 = (1-0.9)0.9^2 = 0.081 = 8.1\%.\end{aligned}$$

5.5.2 Tandem Queue

Now let us extend our model. Suppose two queues are connected in tandem meaning that the queues are in series. Jobs are still arriving at rate $\alpha = 2$ jobs per minute. The first queue's mean processing time is $m_1 = 0.45$ minutes and the second queue's mean processing time is $m_2 = 0.40$ minutes. Define $X(t) = (X_1(t), X_2(t))$ where $X_i(t)$ is the number of jobs at station i at time t. This model is much closer to the situation described in the book, "The Goal". If you draw the rate diagram of this CTMC, one part of the diagram will look like the following.



This CTMC may have a stationary distribution. How can you compute the following?

$$\begin{aligned}\mathbf{Pr}\{X(\infty) = (2,3)\} =& \mathbf{Pr}\{X_1(\infty) = 2, X_2(\infty) = 3\} \\ =& \mathbf{Pr}\{X_1(\infty) = 2\}\mathbf{Pr}\{X_2(\infty) = 3\} \\ & \because \text{We can just assume the independence} \\ & \text{because each queue does not seem to affect each other.} \\ =& (1-\rho_1)\rho_1^2(1-\rho_2)\rho_2^3\end{aligned}$$

where $\rho_1 = \alpha m_1, \rho_2 = \alpha m_2$. When determining ρ_2, you may be tempted to set it as m_2/m_1. However, if you think about it, the output rate of the first station is α not $1/m_1$ because $\rho_1 < 1$. If $\rho_1 \geq 1$, we don't have to care about the steady state of the system since the first station will explode in the long run.

Let us summarize. Let $\rho_i = \alpha m_i$ for $i = 1, 2$ be the traffic intensity at station i. Assume $\rho_1 < 1, \rho_2 < 1$. Then,

$$\begin{aligned}\mathbf{Pr}\{X(\infty) = (n_1, n_2)\} =& (1-\rho_1)\rho_1^{n_1}(1-\rho_2)\rho_2^{n_2}, \quad n_1, n_2 = 0, 1, 2, \dots \\ \pi_{(n_1,n_2)} =& (1-\rho_1)\rho_1^{n_1}(1-\rho_2)\rho_2^{n_2}.\end{aligned}$$

Suppose I want to claim $\pi = (\pi_{(n_1,n_2)}, n_1, n_2 = 0, 1, 2, \dots)$ is indeed the stationary distribution of the tandem queue. What should I do? I just need to verify that this distribution satisfies the following balance equation.

$$(\alpha + \mu_1 + \mu_2)\pi(2,3) = \alpha\pi(1,3) + \mu_2\pi(2,4) + \mu_1\pi(3,2)$$

Since this chain is irreducible, if one distribution satisfies the balance equation, it is the only one satisfying it, which is my unique stationary distribution. You may think this looks like cheating. There can be two ways to find the stationary distribution: one is to solve the equation $\pi P = \pi$ directly, the other is guess something first and just verify it is the one we are looking for. If possible, we can take the easy way. If you can remember M/M/1 queue's stationary distribution, you should be able to compute the stationary distribution of this tandem queue.

5.5.3 Failure Inspection

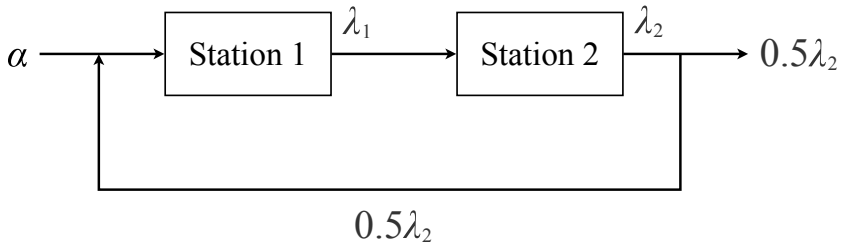
Returning to Open Jackson Network, "Open" means that arriving customers will eventually leave the system.

Let us extend our model once again. Every setting remains same except that there is inspection at the end of the second station. If a job is found to be defective, the job will go back to the queue of station 1 and get reprocessed. The chance of not being defective is 50%. Even in this case, the stationary distribution has the same form to the previous case.

$$\begin{aligned}\mathbf{Pr}\{X_1(\infty) = 2, X_2(\infty) = 3\} =& \mathbf{Pr}\{X_1(\infty) = 2\}\mathbf{Pr}\{X_2(\infty) = 3\} \\ =& (1-\rho_1)\rho_1^2(1-\rho_2)\rho_2^3\end{aligned}$$

The question is how to set ρ_1, ρ_2. Let us define a new set of variables. Denote the throughput from station i by λ_i. It is reasonable to assume that $\lambda_1 = \lambda_2$ for a stable system. Then, because of the feedback,

$$\lambda_1 = \alpha + 0.5\lambda_2 \quad \Rightarrow \quad \lambda_1 = \alpha + 0.5\lambda_1 \quad \Rightarrow \quad \lambda_1 = 2\alpha = 2.$$



Therefore,

$$\begin{aligned}\rho_1 =& \lambda_1 m_1 = 2(0.45) = 90\% \\ \rho_2 =& \lambda_2 m_2 = 2(0.40) = 80\%.\end{aligned}$$

This is called the traffic equation. How can we compute the average number of jobs in the system? Recall that the average number of jobs in M/M/1 queue is $\rho/(1-\rho)$. Then,

$$\begin{aligned}&\text{average number of jobs in the system} \\ =&\text{average \# of jobs in station 1} + \text{average \# of jobs in station 2} \\ =&\frac{\rho_1}{1-\rho_1} + \frac{\rho_2}{1-\rho_2} = \frac{0.9}{1-0.9} + \frac{0.8}{1-0.8} = 9 + 4 = 13 \text{ jobs.}\end{aligned}$$

Still, we did not prove that $(1-\rho_1)\rho_1^{n_1}(1-\rho_2)\rho_2^{n_2}$ is indeed the stationary distribution. The only thing we need to prove is that this distribution satisfies the balance equation. At first glance, the two stations seem to be quite interrelated, but in steady-state, the two are in fact independent. It was found by Jackson at the end of 50s and published in Operations Research. In theory, the independence holds for any number of stations and each station can have multiple servers. It is very general result.

What is the average time in system per job? Use Little's Law, $L = \lambda W$.

$$L = \lambda W \quad \Rightarrow \quad 13 = 1W \quad \Rightarrow \quad W = 13 \text{ minutes.}$$

It could be 13 hours, days, weeks. It is lead time that you can quote to your potential customer. How can we reduce the lead time? We can definitely hire or buy more servers or machines. Another thing you can do is to lower the arrival rate, in reality term, it means that you should reject some orders. It may be painful or even impossible. Suppose we reduce the failure rate from 50% to 40%. Then,

$$\begin{aligned}\rho_1 =& \frac{5}{3}(0.45) = (0.15)5 = 0.75 = \frac{3}{4}, \quad \rho_2 = \frac{5}{3}(0.4) = \frac{2}{3} \\ W =& \frac{\rho_1}{1-\rho_1} + \frac{\rho_2}{1-\rho_2} = \frac{3/4}{1/4} + \frac{2/3}{1/3} = 3 + 2 = 5 \text{ minutes.}\end{aligned}$$

Just with 10% point decrease, the lead time dropped more than a half. This is because of the nonlinearity of $\rho/(1-\rho)$. What if we have 55% failure than 50%? It will become much worse system. $W = 30$ minutes. You must now be convinced that you have to manage the bottleneck. You cannot load up your bottleneck which is already full.

Before finishing up, let us try out more complicated model. There are two inspections at the end of station 2 and 3. Failure rates are 30% and 20%

respectively. Then,

$$\begin{aligned}\lambda_2 =& \lambda_1 + 0.2\lambda_3\\ \lambda_1 =& \alpha + 0.3\lambda_2\\ \lambda_3 =& 0.7\lambda_2.\end{aligned}$$

Now we can compute $\rho_1 = \lambda_1 m_1, \rho_2 = \lambda_2 m_2, \rho_3 = \lambda_3 m_3$. The stationary distribution also has the same form.

$$\pi(n_1, n_2, n_3) = (1-\rho_1)\rho_1^{n_1}(1-\rho_2)\rho_2^{n_2}(1-\rho_3)\rho_3^{n_3}$$

Remember that it has been infinite queue size.

5.6 Simulation

The key difference between DTMC and CTMC comes from the use of generator matrix. Let us see how we can obtain transition probability matrix between time 0 and t from the generator matrix. Suppose we have to following generator matrix.

$$G = \begin{pmatrix} -2 & 1 & 1 \\ 2 & -5 & 3 \\ 2 & 2 & -4 \end{pmatrix}$$

Let us compute $P(0.5) = e^{0.5G}$ and $P(5) = e^{5G}$ from G where e^G is the matrix exponential of G.

Listing 5.1: Use of Generator Matrix

```
library(expm)

# Set up the generator matrix
G = matrix(c(-2,1,1, 2,-5,3, 2,2,-4), 3,3,byrow=TRUE)
G

# The following matrices are P(0.5) and P(5)
expm(0.5*G)
expm(5*G)
```

5.7 Exercise

1. Consider a CTMC $X = \{X(t), t \geq 0\}$ on $S = \{A, B, C\}$ with generator G given by

$$G = \begin{bmatrix} -12 & 4 & 8 \\ 5 & -6 & 1 \\ 2 & 0 & -2 \end{bmatrix}$$

   (a) Draw the rate diagram.

(b) Use a computer software like matlab or a good calculator to directly compute the transition probability matrix $P(t)$ at $t = 0.20$ minutes. (In matlab, the command to exponentiate a matrix A is expm(A).)

(c) Do the previous part for $t = 1.0$ minute.

(d) Using the results from parts (b) and (c), but without using a software package or calculator, find $P\{X(1.2) = C|X(0) = A\}$ and $P\{X(3) = A|X(1) = B\}$.

(e) Do part (b) for t = 5 minutes. What phenomenon have you observed?

2. Customers arrive at a two-server system according to a Poisson process with rate $\lambda = 10$ per hour. An arrival finding server 1 free will begin his service with him. An arrival finding server 1 busy, server 2 free will join server 2. An arrival finding both servers busy goes away. Once a customer is served by either server, he departs the system. The service times at both servers are exponential random variables. Assume that the service rate of the first server is 6 per hour and the service rate of the second sever is 4 per hour.

   (a) Describe a continuous time Markov chain to model the system and give the rate transition diagram.

   (b) Find the stationary distribution of the continuous time Markov chain.

   (c) What is the long-run fraction of time that server i is busy, $i = 1, 2$?

3. Suppose we have a small call center staffed by two operators A and B handling three telephone lines. A only handles line 1, and B only handles only line 2 and 3. Calls arrive according to a Poisson process with rate $\lambda = 100$ calls per hour. All arrivals prefer line 1. Arrivals find all lines are busy will go away. Service times are exponentially distributed with a mean of 4 minutes. Customers are willing to wait an exponentially distributed length of time with a mean of 8 minutes before reneging if service has not begun.

   Describe a continuous time Markov chain to model the system and give the rate transition diagram and the generator G.

4. Consider a call center that is staffed by K agents with three phone lines. Call arrivals follow a Poisson process with rate 2 per minute. An arrival call that finds all lines busy is lost. Call processing times are exponentially distributed with mean 1 minute. An arrival call that finds both agents busy will wait in the third phone line untill get service.

   (a) Find the throughput and average waiting time when $K = 1$.

   (b) Find the throughput and average waiting time when $K = 2$.

   (c) Find the throughput and average waiting time when $K = 3$.

5. Consider a production system consisting of three single-server stations in series. Customer orders arrive at the system according to a Poisson process with rate 1 per hour. Each customer order immediate triggers a job that is released to the production system to be processed at station 1 first, and then at station 2. After being processed at station 2, a job has $p_1 = 10\%$ probability going back to station 1 for rework and $1 - p_1$ probability continuing onto station 3. After being processed at station 3, a job has $p_2 = 5\%$ probability going back to station 1, $p_3 = 10\%$ probability going back to station 2, and $1 - p_2 - p_3$ probability leaving the production system as a finished product. Assume that the processing times of jobs at each station are iid, having exponential distribution, regardless of the history of the jobs. The average processing times at stations 1, 2 and 3 are $m_1 = 0.8$, $m_2 = 0.70$ and $m_3 = 0.8$ hours, respectively.

   (a) Find the long-run fraction of time that there are 2 jobs at station 1, 1 job at station 2 and 4 jobs at station 3.

   (b) Find the long-run average (system) size at station 3.

   (c) Find the long-run average time in system for each job.

   (d) Reduce p_1 to 5%. Answer 1(c) again. What story can you tell?

6. Consider the production system in the previous problem with following modification: $p_1 = 10\%$, $p_2 = p_3 = 0$ (station 3 makes 100% reliable operations.) Furthermore, the Management decides to adopt the make-to-stock policy, using the CONWIP job release policy. Recall that when the system is operating CONWIP policy only a job leaving station 3 triggers a new job to be released to station 1. Let N be the CONWIP level.

   (a) For $N = 2$, compute the throughput of the production system. What is the average time in system per job?

   (b) Is it possible to double the throughput? If so, what N is needed to achieve it? What is the corresponding average time in system per job?

7. Consider a production system that consists of three single-server stations. Each station has an infinite size waiting buffer. When a job to a station finds the server busy, the job waits in the buffer; when the server is free, the server picks the next job from the buffer. Assume that jobs arrive at station 1 following a Poisson process with rate $\alpha = 1$ job per minute. After being processed at station 1, a job is sent to station 2 without any delay. After being processed at station 2, it will be sent to station 3 with probability 50% or to station 1 with probability 50% to be reworked. After being processed at station 3, it is shipped to a customer with probability 50% or is sent to station 2 with probability 50% to be reworked. The job processing times at station i are iid exponentially distributed with mean m_i minutes, $i = 1, 2, 3$. Assume that

$$m_1 = .3 \text{ minutes}, \quad m_2 = .2 \text{ minutes}, \quad m_3 = .45 \text{ minutes}.$$

(a) What is long-run average utilization for station i, $i = 1, 2, 3$.

(b) What is the long-run fraction of time that station 1 has 2 jobs, station 2 has 3 jobs, and station 3 has one job? (Leaving your answer in a numerical expression is OK.)

(c) What is the long-run average number of jobs at station 2, including those in the buffer and possibly the one being served?

(d) What is the throughput of the system? Here, throughput is the rate at which the jobs are being shipped to customers.

(e) What is the long-run average time in system per job?

8. Let $X = \{X(t) : t \geq 0\}$ be a continuous time Markov chain with state space $\{1, 2, 3\}$ with generator

$$G = \begin{pmatrix} -4 & ? & 1 \\ 1 & ? & 2 \\ 1 & ? & -2 \end{pmatrix}.$$

Using `matlab`, one can compute e^G via command

$$\texttt{expm}(G) = \begin{pmatrix} 0.20539 & 0.36211 & 0.43250 \\ 0.19865 & 0.35727 & 0.44407 \\ 0.19865 & 0.33896 & 0.46239 \end{pmatrix}.$$

(a) Fill in entries in G.

(b) Find $\mathbf{Pr}\{X(2) = 3 | X(1) = 2\}$.

(c) Find $\mathbf{Pr}\{X(1) = 2, X(3) = 3 | X(0) = 2\}$ (Leaving answer in a numerical expression is OK.)

(d) Find the stationary distribution of the Markov chain.

(e) Find $\mathbf{Pr}\{X(200) = 3 | X(1) = 2\}$.

9. A call center has 4 phone lines and two agents. Call arrival to the call center follows a Poisson process with rate 2 calls per minute. Calls that receive a busy signal are lost. The processing times are iid exponentially distributed with mean 1 minute.

(a) Model the system by a continuous time Markov chain. Specify the state space. Clearly describe the meaning of each state. Draw the rate diagram.

(b) What is the long-run fraction of time that there are two calls in the system?

(c) What is the long-run average number of waiting calls, excluding those in service, in the system?

(d) What is the throughput (the rate at which completed calls leaves the call center) of the call center?

(e) What is the average waiting time of an accepted call?

10. A production line consists of two single-machine stations, working in serial. Each machine has an infinite buffer waiting area. Jobs arrival at the production line follows a Poisson process with rate 2 jobs per minute. Each job is first processed at station 1; afterwards, it is processed at station 2. After a job is processed at station 2, with 12% probability the job is sent back to station 1 for rework, and with 88% probability the job is shipped to customers, regardless of the job's history. When the machine at a station is busy, the arriving job at the station is waiting in its buffer. Assume that processing times at each machine are iid, exponentially distributed with mean .4 minutes.

    (a) Find the long-run average utilization of machine 1 and machine 2.

    (b) What is the long-run fraction of time that there 2 jobs at station 1 and 3 jobs at station 2? you may leave your answer in an expression.

    (c) What is the long-run average number of jobs at station 2?

    (d) What is the average time in system per job?

11. A production system has three machines working in parallel. The up times of a machine is assumed to be iid, exponentially distributed with mean 3 hours. When a machine is down, its repair times are iid, exponentially distributed with mean 1 hour. There are two repairmen, working at the same speed.

    (a) Using a CTMC to find the long-run fraction of time that all machines are up and running; you need to specify the meaning of each state; draw the rate diagram; spell out necessary equations that are used in your calculations.

    (b) When a machine is up and running, it processes customer orders at rate 4 orders per minute. Assume the customer orders are backlogged in the next month. What is the average production rate per minute of the production system?

12. Suppose you are running a factory having three stations. Parts come into Station 1 according to a Poisson process with rate 720 per hour. After being processed by Station 1, the parts move to Station 2 with probability 1/3. With 2/3 probability, the parts move to Station 3. After being processed by Station 2, the parts go back to Station 1. After being processed by Station 3, the parts immediately leave the system. The average processing times of three stations are exponential with mean $m_1 =$ 2 seconds, $m_2 = 3$ seconds, $m_3 = 4$ seconds. Each station has a single server and infinite waiting space.

    (a) What is the utilization of Station 1?

    (b) What is the long-run fraction of time that the entire factory is empty?

(c) What is the total average number of parts in the entire factory?

(d) What is the total average "waiting" time spent in the factory per part?

(e) What is the throughput of the entire factory?

13. Suppose you are running a laundry shop at the Tech Square. The shop is open 24 hours a day 7 days a week. You have virtually infinite number of laundry machines so that each student can start his/her laundry as soon as he/she arrives at your shop. Each machine can finish any amount of laundry in exponential time with mean 30 minutes. Students come to your shop according to a Poisson process with rate 6 per hour. Each student waits in your shop until his/her laundry is done. You are trying to model this shop using CTMC. Let $X(t)$ denote the number of students in your shop.

(a) Which of the following is most likely the right model for this problem? Choose one from "M/M/1", "M/M/1/∞", "M/M/∞", "M/M/∞/1".

(b) What is the state space of $X(t)$? Draw the rate diagram of $X(t)$.

(c) Compute the stationary distribution of $X(t)$.

(d) What is the long-run fraction of time that at least 1 student is in the shop?

(e) If there is at least 1 student is in the shop, you make \$100 per hour on average. On the other hand, when the shop is empty, you lose \$50 per hour on average. What is long-run average profit per day?