Note about the theorem that I cited twice already:

**Theorem:**  \( \mu_k > 0 \) for all \( k \in \{1, 2, 3, \ldots\} \)

1. \( \lim_{t \to \infty} q(t) = \pi = [\pi_0, \pi_1, \pi_2, \ldots] \), with \( \pi_i > 0 \),
   independently of \( q(0) \),
   and \( \pi \) may be or not be a proper prob. distribution.

2. \( \pi \) satisfies the **Balance Equations**: \( \mu_k \pi_{k+1} = \lambda_k \pi_k \), \( k = 0, 1, 2, \ldots \)

**Remarks:**

- When the state space is finite, \( S = \{1, 2, \ldots, K\} \), then
  \( \pi \) (solution to the balance equations) is always a proper
  probability distribution.

- Also, if \( \lambda_k = 0 \) for \( k > M \), \( \pi \) is proper

The theorem holds under weaker conditions as well (intuitive).

* e.g., if \( \mu_k = 0 \) for \( k < M \),

Problems arise when there is more than one closed
communication class (this was also the case for
discrete-time Markov Chains).
for example, the following B&D process:

\[
\begin{array}{c}
0 \rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow \ldots
\end{array}
\]

is troublesome in the sense that \( q(t) \) may converge to something, but this will depend on \( q(0) \).

---

Condition for \( \Pi \) to be a proper probability distribution (assuming \( \mu_k \neq 0, \forall k \in \{1, 2, 3, \ldots\} \)).

\[
\Pi_k = \frac{\prod_{i=1}^{k} \lambda_i}{\prod_{i=1}^{k} \mu_i} \Pi_0,
\]

from balance equations.

\( \Pi \) is proper

\( \uparrow \downarrow \) (if and only if) actually by definition

\[
\sum_{k=0}^{\infty} C_k < \infty \quad (C_0 = 1)
\]

In which case:

\[
\Pi_0 = \frac{1}{\sum_{k=0}^{\infty} C_k}, \quad \Pi_k = C_k \Pi_0.
\]

If opposite is true: \( \Pi_k = 0, \forall k \in \{0, 1, 2, \ldots\} \)
Queuing models.

To justify the notation that follows, we will note that the counting process $N(t)$ of a Poisson process (with arbitrary rate) has the Markov property (for continuous-time stochastic processes):

$$
\text{if } t_1 < t_2 < t_3, \quad k_1 \leq k_2 \leq k_3,
$$

$$
P[N(t_3) = k_3 \mid N(t_2) = k_2, N(t_1) = k_1] = P[N(t_3) = k_3 \mid N(t_2) = k_2].
$$

You can prove it fairly easily. It follows from independence of inter-arrival times and memorylessness of exponential r.v.

(Birth and death stochastic processes have the same property.)

A queuing model is characterized by:

- Input source (typically Poisson)
- Queuing mechanism:
  - service time (typically exponential)
  - number of servers
- Capacity of queue (infinite or finite)

{four items we will mostly care about}
Remark: on the memorylessness of service time. If I am first in line, there are three servers (each with the same service rate \( \mu \)), then, when it’s my turn, my service time (r.v.) is the same as the other two customers, even though they started being served before me.

Typically, a queuing model is specified indicating four (or more) parameters.

E.g.: \( M/M/1/\infty \) or simply \( M/M/1 \).

\[ \begin{align*}
\text{Poisson input source (} M = \text{Markov)} \\
\text{Exponential service time} \\
\text{One server} \\
\text{Infinite capacity}
\end{align*} \]

In general:

1. **Input source**: M: Poisson r.v.
   D: degenerate distribution (constant inter-arrival)
   \( E_m \): Erlang of order \( m \)
   G: general distribution

2. **Service time**: M
   D
   \( E_n \)
   G
   makes sense when service is a sequence of stages
(3) Number of servers: \( s = 1, 2, 3, \ldots \)

(4) Capacity of queue: \( \infty \) (easy analysis)

\[ K < \infty \text{ (there is probability of loss)} \]

- limited buffer capacity
- limited number of seats in waiting room of

Probability of loss = probability of having at least one arrival while the queue is full.

We (and your book) only care about the above parameters, but you may find more sophisticated queue models that also take into account:

(5) Queue discipline:

- FCFS (first-come-first-served)
- LCFS (last-come-first-served)
- Random
- According to some priority procedure (special ticket)
- Other

Affects average waiting time.

(6) Population size (finite or \( \infty \)).
As we saw last time, the M/M/1/∞ queue can be described with a B&D process \((X(t) = \text{# of customers in service})\); for most queuing models this is the case.

It makes sense to consider the steady-state for such processes. The following notation assumes that the system is in steady-state condition.

- \(\Pi = [\pi_0, \pi_1, \pi_2 \ldots] \): steady-state distribution,
  
  \(\text{i.e.: } \pi_k = \text{probability of exactly } k \text{ customers in queuing system;}

- \(L = \text{expected } \# \text{ of customers in queuing systems}
  
  \left(= \mathbb{E}[X(t)] \text{ for large } t \right)
  
  = \sum_{k=0}^{\infty} k \pi_k

- \(L_q = \text{expected queue length}
  
  \text{expected } \# \text{ of customers in line (excludes those being served)}
  
  = 0 \cdot \pi_s + 1 \cdot \pi_{s+1} + 2
  
  = \sum_{k=s}^{\infty} (k-s) \pi_k

- \(W = \text{waiting time in system (r.v.) for individual customer}
  
  W = \mathbb{E}[W]

- \(W_q = \text{waiting time in queue (r.v.), excludes service time}
  
  W_q = \mathbb{E}[W_q]

  \text{[in hrs, will compute above quantities for simple finite-state processes]}\)
We will now study queue model $\text{M}/\text{M}/1/\infty$:

- simplest model, but once you know it it is fairly easy to learn more complicated models.
- The number of customers in system can be modeled as B&D process.
- assuming for now that $\lambda_k = \lambda$ and $\mu_k = \mu$.

**Flow Diagram:**

![Flow Diagram](image)

We saw last time that if $\frac{\lambda}{\mu} < 1$ then $q(t) \to \pi$, steady state distribution, with

$$\pi_k = \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^k, \quad k = 0, 1, 2, \ldots$$

i.e. $X(t) \sim G\left(1 - \frac{\lambda}{\mu}\right)$.

**Remark:** If $Y \sim G(p)$, i.e. $p_Y(k) = p(1-p)^k$, we have that [IMPORTANT PROCEDURE!]

$$E[Y] = \sum_{k=0}^{\infty} k \cdot p_Y(k) = p \sum_{k=0}^{\infty} k (1-p)^k = p \sum_{k=1}^{\infty} k (1-p)^{k-1}$$

$$= p (1-p) \sum_{k=1}^{\infty} k (1-p)^{k-1} =$$

$$= -p \left(1-p\right) \sum_{k=1}^{\infty} \frac{d}{dp} (1-p)^k = -p (1-p) \frac{d}{dp} \sum_{k=1}^{\infty} (1-p)^{k}$$

$$= -p (1-p) \frac{d}{dp} \left\{ \frac{1}{1-(1-p)} - 1 \right\} \left(\frac{1}{1-p} - 1\right) = -p (1-p) \frac{d}{dp} \frac{1}{p} =$$

$$= p (1-p) \frac{1}{p^2} = \frac{1-p}{p}.$$
• Set $\rho = 1 - \frac{\lambda}{\mu}$, and get:

for large $t$,

$$L = \mathbb{E}[X(t)] = \frac{1 - 1 + \frac{\lambda}{\mu}}{1 - \frac{\lambda}{\mu}} = \frac{\lambda}{\mu - \lambda} = \frac{\lambda / \mu}{1 - \lambda / \mu}.$$

• For M/M/1 queues we introduce the **utilization factor**:

$$\rho = \frac{\lambda}{\mu} \quad \left( \frac{\lambda}{\mu} < 1 \implies \rho < 1 \right);$$

since $\rho = \frac{\lambda}{\mu} = 1 - \pi_0$, we have $\rho = \text{FRACTION OF TIME THAT THE SERVER IS BUSY.}$

• We may write:

$$L = \frac{\rho}{1 - \rho}.$$

• $L_q = \sum_{k=1}^{\infty} (k-1) \pi_k = \sum_{k=1}^{\infty} k \pi_k - \sum_{k=1}^{\infty} \pi_k$

$$= L - (1 - \pi_0) = \frac{\lambda}{\mu - \lambda} - \frac{\lambda}{\mu} = \frac{\lambda^2}{\mu (\mu - \lambda)}$$

or:

$$L_q = L - (1 - \pi_0) = \frac{\rho}{1 - \rho} - \rho = \frac{\rho - \rho + \rho^2}{1 - \rho} = \frac{\rho^2}{1 - \rho}.$$
Note: we may introduce:

$L_s = \text{average \# of customers being served}$

$L = L_q + L_s$

For $M/M/1$ queue: $L_s = \frac{\lambda}{\mu}$.

- In fact, if $L_s = \# \text{ of customers being served}$,

  $L_s = \begin{cases} 
  0 & \text{if } X(t) = 0, \text{ i.e. w.p. } \pi_0 \\
  1 & \text{if } X(t) > 0, \text{ i.e. w.p. } \pi_1 + \pi_2 + \ldots = 1 - \pi_0 
  \end{cases}$

  so $L_s = 1E[L_s] = \lambda (1 - \pi_0) = 1 - \pi_0 = \frac{\lambda}{\mu} = \rho$.

- $W$: waiting time in system:
  i.e., if I join the line at a certain time, how long do I have to wait?

  $W = 1E[W]$.

Suppose that when I join queue there are $X(t)$ (for simplicity) customers in line:

$W = T + T_1 + T_2 + \ldots + T_x$

↑ service time of first customer in line, etc.
$T \sim E(\mu)$
$T_i \sim E(\mu)$, for $i = 1, 2, \ldots, X$.

$W$ is not ERLANG, because $X$ is random
(sum of a random number of random variables!)

If I fix $X = k$, then $W$ (under condition $X = k$) \( \sim E(\mu, k+1) \),
i.e. $\mathbb{E}[W \mid X = k] = \frac{k+1}{\mu} \pi_k$

So:

$$W = \mathbb{E}[W] = \sum_{k=0}^{\infty} \mathbb{E}[W \mid X = k] \cdot \mathbb{P}[X = k]$$

$$= \sum_{k=0}^{\infty} \frac{k+1}{\mu} \pi_k$$

$$= \frac{1}{\mu} \sum_{k=0}^{\infty} k \pi_k + \frac{1}{\mu} \sum_{k=0}^{\infty} \pi_k$$

$$= \frac{1}{\mu} \frac{\lambda}{\mu - \lambda} + \frac{1}{\mu} = \frac{1}{\mu} \left( \frac{\lambda}{\mu - \lambda} + 1 \right)$$

$$= \frac{1}{\mu} \frac{\lambda + \mu - \lambda}{\mu - \lambda} = \frac{1}{\mu - \lambda}$$

Note that (since $L = \frac{\lambda}{\mu - \lambda}$):

$$\boxed{L = \lambda W} \quad \text{LITTLE'S FORMULA}$$

(has more general validity than $M/M/1$).
It turns out that: $W \sim E(\mu - \lambda)$.

$$F_W(t) = P[W \leq t] =$$

$$= \sum_{k=0}^{\infty} P[W \leq t \mid X = k] \cdot P[X = k]$$

Erlang Prob. distrib $\sim E(\mu, k+1)$

$$f_W(t) = \sum_{k=0}^{\infty} \frac{\mu^{k+1}}{k!} t^k e^{-\mu t} (1 - \frac{\lambda}{\mu})^k$$

$$= \mu (1 - \frac{\lambda}{\mu}) e^{-\mu t} \sum_{k=0}^{\infty} \frac{1}{k!} (\lambda t)^k$$

$$= (\mu - \lambda) e^{-\mu t} e^{\lambda t} = (\mu - \lambda) e^{-(\mu - \lambda)t}$$

---

$W_q = E[\text{waiting time in queue}]$

$$= W - W_s = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{\mu - \mu + \lambda}{\mu (\mu - \lambda)} = \frac{\lambda}{\mu (\mu - \lambda)}.$$

(expected service time)
Next time: \[ M/M/1/K \]
\[ M/M/s \]
\[ M/M/S/K \]
\[ (M/G/1 ?) \]

\[ \lambda = 10 \quad \mu = 15 \quad \Rightarrow \quad X(t) \text{ has an (asymptotic) long term prob. distr.} \]

\* Proportion of time during which no-one is waiting to be served:
\[ \Pi_0 = 1 - \frac{\lambda}{\mu} = 1 - \frac{15-10}{15} = \frac{1}{3} \]

\* Average # of customers in system:
\[ L = \frac{\lambda}{\mu - \lambda} = \frac{10}{5} = 2 \]

\* Average # in line:
\[ L_q = 2 - \frac{10}{15} = 2 - \frac{2}{3} = \frac{4}{3} \]

\* Average waiting time in system:
\[ W = \frac{1}{\mu - \lambda} = \frac{1}{5} \text{ (12 minutes)} \]