Overview: Stochastic Processes, Markov Chains, and Queues

Lecturer: Javad Ghaderi

Definition 1 (Stochastic Process)**.** A stochastic process is sequence of random variables $(X_t : t \in T)$, where X_t takes on values in a set S. In many applications, the index set T is a set of times, and S is the set of possible states for the system. The index set T could be discrete-time (consecutive integers), or continuous-time (real numbers).

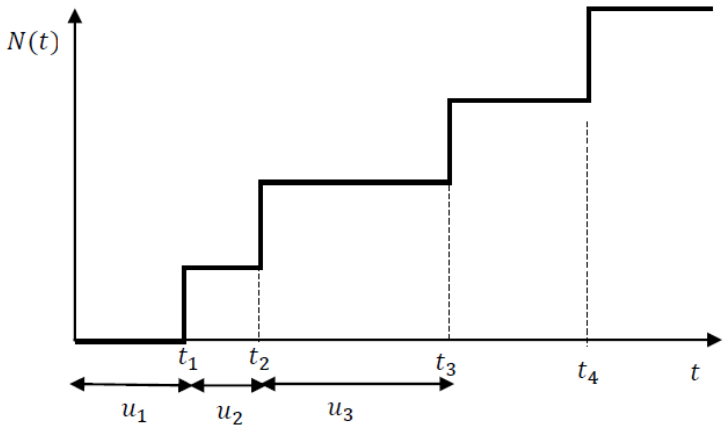
Any set of instances of $(X(t); t \in T)$ can be regarded as a path of a particle moving randomly in the state space S, its position at time t being $X(t)$. These paths are called sample paths of the stochastic process.

1 Counting Process and Poisson Process

Definition 2 (Counting Process)**.** A stochastic process is a counting process if its sample path is piecewise constant and increases by one at discrete instances in time (called count times or arrival times). Thus the process is described by a set of count (arrival) times $\{t_i\}$, or equivalently by the set of intercount (inter-arrival) times $u_1 = t_1, \cdots, u_i = t_i - t_{i-1}, \cdots$. Hence, the number of arrivals before time t, $N(t)$, is

$$N(t) = \sum_i \mathbb{1}(t_i \leq t),$$

where $\mathbb{1}(\cdot)$ is the indicator function.



Definition 3 (Poisson Process)**.** Consider a counting process $N(t)$ with $N(0) = 0$. If intercount times u_i, $i = 1, 2, \cdots$, are iid exponentially distributed with parameter λ, then $N(t)$ is called a Poisson process with rate λ.

It follows that this definition of Poisson process is equivalent to the following two definitions.

Definition 4 (Poisson Process)**.** A counting process $N(t)$ is called a Poisson process with rate λ if for any $t, s > 0$,
(i)$N(0) = 0$,
(ii)$N(t+s) - N(t)$ is independent of $N(t)$ (independent increments property),
(ii)$N(t+s) - N(t)$ has a Poisson distribution with parameter λs, i.e.,

$$\mathbb{P}\{N(t+s) - N(t) = k\} = \frac{(\lambda s)^k e^{-\lambda s}}{k!}, k = 0, 1, 2, \cdots$$

Another equivalent definition is the following.

Definition 5 (Poisson Process)**.** A counting process $N(t)$ is called a Poisson process with rate λ if

1. $N(0) = 0$.
2. $\mathbb{P}\{N(t+\delta) - N(t) = 1\} = \lambda\delta + \mathrm{o}(\delta)$,
3. $\mathbb{P}\{N(t+\delta) - N(t) > 1\} = \mathrm{o}(\delta)$.

Note that $\mathrm{o}(\cdot)$ means that $\frac{\mathrm{o}(\delta)}{\delta} \to 0$ as $\delta \to 0$ (for example $\delta^2 = \mathrm{o}(\delta)$). Hence in any small enough interval of length δ, there is either a count with probability $\lambda\delta$, or no counts with probability $1 - \lambda\delta$ (only these two events have non-negligible probability).

1.1 Properties of Poisson Process

Memoryless property

Assume we have waited for some time 's' after t_i, what is the probability that we have to wait for an additional 't' time units before seeing the $(i+1)$-th count?

$$\mathbb{P}\{t_{i+1} > t_i + s + t | t_{i+1} > t_i + s\} = \frac{\mathbb{P}\{(t_{i+1} > t_i + s + t) \cap (t_{i+1} > t_i + s)\}}{\mathbb{P}\{t_{i+1} > t_i + s\}} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t}$$

i.e., the waiting time is still exponential independent of s.

Merging

Let $N_1(t)$ and $N_2(t)$ be two independent Poisson processes with rates λ_1 and λ_2. Then the process $N_1(t) + N_2(t)$ is Poisson with rate $\lambda_1 + \lambda_2$.
Clearly this can be extended to more than two processes.

Splitting

Suppose $N(t)$ is a Poisson process with rate λ. Create two new process $N_1(t)$ and $N_2(t)$ by assigning each count event to the first process with probability p and to the second one with probability $1-p$. Then $N_1(t)$ and $N_2(t)$ are independent Poisson processed with rates λp and $\lambda(1-p)$ respectively.

Clearly this can be extended to slitting into more than two processes.

2 Markov processes

Markov Processes naturally arise in the modeling of many systems where there is a notion of state for the system at each time. The state at time t contains all the relevant information about the system up to and including time t that is relevant to the future of the system. For example, the state of an aircraft at time t could consist of the position, velocity, and remaining fuel at time t. Think of t as the present time. Given the state at time t, the future part of the aircraft trajectory is determined, independently of the history up to time t, i.e, it does not matter how the aircraft has reached the current state at time t.

Definition 6 (Markov Process)**.** A process $X(t)$ is a Markov process if it has the memoryless property: Given the value of $X(t)$ at some time $t \in T$, the future path $X(s)$ for $s > t$ does not depend on knowledge of the past history $X(u)$ for $u < t$, i.e. for $t_1 < \cdots < t_n < t_{n+1}$,

$$\mathbb{P}\left\{X(t_{n+1}) = x_{n+1} | X(t_n) = x_n; \cdots; X(t_1) = x_1\right\} = \mathbb{P}\left\{X(t_{n+1}) = x_{n+1} | X(t_n) = x_n\right\}$$

The Markov processes that we will be considering in the course will have the following properties.

Definition 7 (Irreducibility)**.** $X(t)$ is irreducible if all states in S can be reached from all other states, by following the transitions of the process. If we draw a directed graph of the state space with a node for each state and an arc for each event, or transition, then for any pair of nodes there is a path connecting them, i.e. the graph is strongly connected.

Definition 8 (Time-homogeneity)**.** $X(t)$ is time homogeneous if behavior of the system does not depend on when it is observed. In particular, the transition probabilities between states are independent of the time at which the transitions occur. Thus, for all s and u,

$$\mathbb{P}\left\{X(s+t_1) = y | X(s) = x\right\} = \mathbb{P}\left\{X(u+t_1) = y | X(u) = x\right\}.$$

In this course, our primary objective with respect to a Markovian models will be to calculate the probability distribution of the random variable $X(t)$ over the state space S, as the time goes to infinity. The long run behavior of the system usually approaches a regular pattern called "the steady-state probability distribution". From this probability distribution we will derive performance measures based on subsets of states where some condition holds.

2.1 Discrete-Time Markov Chains

The Markov process $X(t)$ takes on values in a countable set S, for $t = 0, 1, 2, \cdots$. For simplicity, Assume that the elements of S are indexed by integer numbers. Associated with each Markov chain there is a probability transition matrix P, where $P_{ij} = \mathbb{P}\{X(t+1) = j | X(t) = i\}$, $i, j \in S$. For brevity we denote the pmf of $X(t)$ as $\pi(t) = (\pi_i(t) : i \in S)$, where $\pi_i(t) = \mathbb{P}\{X(t) = i\}$. Therefore the evolution of the pmf is given by

$$\pi(t+1) = \pi(t)P$$

Q: Is there exists a π such that $\pi(t) \to \pi$, starting from any initial condition $\pi(0)$?

If so, π is called the steady-state distribution. Note that π has to be one of the solutions (equilibrium probability vectors) of the fixed point equation $\pi = \pi P$. Equivalently, this matrix form can be written as

$$\sum_{j \neq i} \pi_j P_{ji} \;\; = \;\; \sum_{j \neq i} \pi_i P_{ij} \; \forall i \in S \text{ (flux into i = flux out of i)}$$

These equations are called global balance equations. We further need to impose that

$$\begin{aligned} \sum_i \pi_i &= 1 \\ \pi_i &\geq 0; \forall i \in S. \end{aligned}$$

Definition 9 (Periodicity)**.** A state i has period k if any return to state i occurs in multiples of k time steps. Formally, the period of a state is defined as

$$k = \gcd\{t : \mathbb{P}\{X(t) = i | X(0) = i\} > 0\},$$

where "gcd" is the greatest common divisor. For example, assume it is possible to return to the state i at times $\{6, 8, 10, 12, ...\}$, then $k = 2$, even though 2 does not appear in this list.

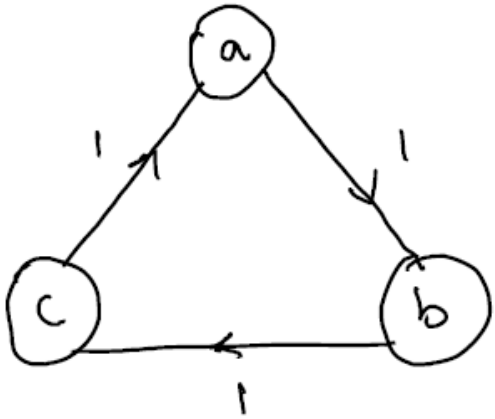
If $k = 1$, then the state is said to be aperiodic: returns to state i can occur at all the times large enough. Formally, a state i is aperiodic if there exists t_0 such that for all $t \geq t_0$,

$$\mathbb{P}\{X(t) = i | X(0) = i\} > 0.$$

A Markov chain is aperiodic if every state is aperiodic. All the states of an irreducible Markov chain have the same period. Hence, an irreducible Markov chain only needs one aperiodic state to imply all states are aperiodic.

Example 1 (irreducible but not aperiodic)**.** Consider the state diagram of the following Markov chain

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



First note that the Markov chain is obviously irreducible. Solving the equation $\pi = \pi P$, we see there is a unique equilibrium probability vector $\pi = (1/3, 1/3, 1/3)$. On the other hand, if $\pi(0) = (1, 0, 0)$, then

$$\pi(t) = \pi(0)P^t = \begin{cases} (1,0,0) & if \quad t \equiv 0 \bmod 3 \\ (0,1,0) & if \quad t \equiv \bmod 3 \\ (0,0,1) & if \quad t \equiv 2 \bmod 3 \end{cases}$$

Therefore, $\pi(t)$ does not converge as $t \to \infty$. This example shows a periodic Markov chain (with period 3).

For a state i, let $\tau_i = \min\{t : X(t) = i\}$, by the convention that the minimum of an empty set of numbers is $+\infty$. Let $M_i = \mathbb{E}[\tau_i | X(0) = i]$. If $\mathbb{P}\{\tau_i < +\infty | X(0) = i\} < 1$, state i is called transient (and by convention, $M_i = +\infty$). Otherwise $\mathbb{P}\{\tau_i < +\infty | X(0) = i\} = 1$, and i is said to be positive recurrent if $M_i < +\infty$ and to be null recurrent if $M_i = +\infty$.

Theorem 1. *Suppose $X(t)$ is irreducible and aperiodic. Then*

(a) All states are transient, or all are positive recurrent, or all are null recurrent.

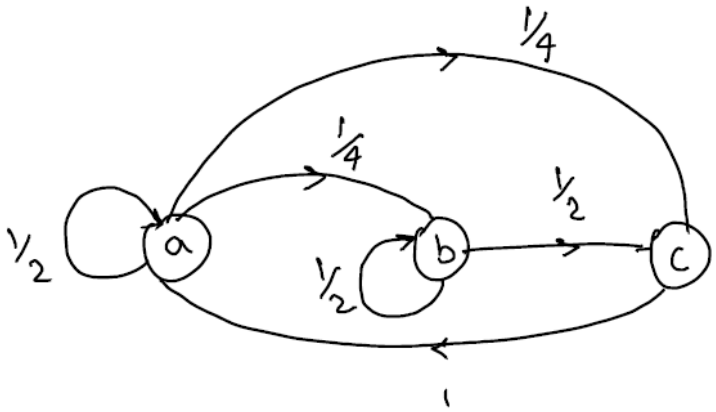
(b) For any initial distribution $\pi(0)$, $\lim_{t\to\infty} \pi_i(t) = 1/M_i$, with the understanding that the limit is zero if $M_i = +\infty$.

(c) There exists a unique equilibrium probability distribution π such that $\pi = P\pi$ and $\lim_{t\to\infty} \pi_i(t) = \pi_i$, if and only if all states are positive recurrent.

Example 2 (calculation of equilibrium (steady sate /stationary) distribution)**.** The Markov chain is irreducible and aperiodic.

$$\begin{aligned} \pi_a(1/2) &= \pi_c(1) \\ \pi_b(1/2) &= \pi_a(1/4) \\ \pi_c(1) &= \pi_a(1/4) + \pi_b(1/2) \\ \pi_a + \pi_b + \pi_c &= 1 \end{aligned}$$

It has a unique stationary solution $\pi_a = 1/2, \pi_b = \pi_c = 1/4$.



In general, it might be difficult to solve the global balance equations when S is a large space. Instead, one can first try a solution that satisfy the following set of equations called detailed balance equations:

$$\begin{aligned} \pi_i P_{ij} &= \pi_j P_{ji} \ \forall i, j \in S \\ \sum_{i \in S} \pi_i &= 1 \end{aligned}$$

It is easy to see that if the system of detailed balance equations has a solution, it would satisfy the global balance equations. If the chain is irreducible and aperiodic, then we know that this is the only possible solution. A Markov chain that satisfies the detailed balance equations is called reversible.

Example 3 (Model of wireless link, Srikant-Ying's book)**.** Here is a simple model of wireless link in discrete time.

Suppose at each time slot either one packet can be transmitted over the wireless channel or zero packets (due to wireless fading). Let $s(t)$ denote the number of packets transmitted over the channel and suppose that $s(t)$'s are iid Bernoulli with mean μ. Suppose packets arrive to the link according to a Bernoulli process with mean λ, i.e., $a(t)$ is Bernoulli with mean λ. There is a buffer (queue) at the link to contain packets waiting for transmission. Let $q(t)$ be the number of packets in the queue at the beginning of time t, then

$$q(t+1) = (q(t) + a(t) - s(t))^+$$

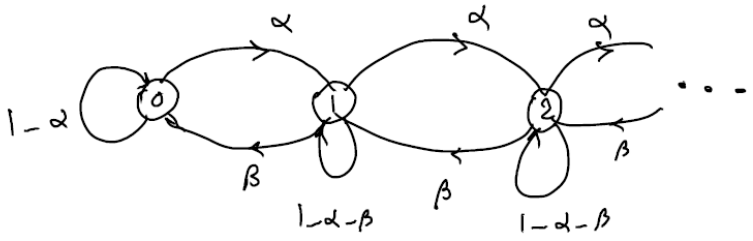
where $(x)^+ := \max\{0, x\}$. The above equation assumes that first arrivals happen, then possible departures. A transition from $q(t) = i$ to $q(t+1) = i+1$ occurs when there is an arrival but no departures. Hence $P_{i,i+1} = \lambda(1-\mu) =: \alpha$. Similarly, $P_{i,i-1} = \mu(1-\lambda) =: \beta$. Otherwise, $q(t)$ remains unchanged.

We try the detailed balance equations

$$\pi_i \alpha = \pi_{i+1} \beta \ \forall i$$

thus

$$\pi_i = (\alpha/\beta)^i \pi_0$$



Since we need $\sum_i \pi_i = 1$, we obtain

$$\sum_{i=0}^{\infty} (\alpha/\beta)^i \pi_0 = 1$$

If $\lambda < \mu$, then $\alpha < \beta$, and $\pi_0 = 1 - \frac{\alpha}{\beta}$. If $\lambda \geq \mu$, then there is no π that solves the detailed balance equations. In fact, one can see that there is no equilibrium distribution in this case. Thus the Markov chain is not positive recurrent (it is unstable) if $\lambda \geq \mu$.

The Markov chain here is called a *Geo/Geo/1* queue (1 server, inter-arrival and service times are geometrically distributed)

2.2 Foster-Lyapunov Stability Criterion

Markov Chains can be very complex and therefore it is useful to determine the stability (positive recurrence) without solving the set of equations. The Foster-Lyapunov method provides such a tool to determine the stability and at the same time can be used to derive approximations or bounds on key performance parameters.

Theorem 2 (Foster-Lyapunov)**.** *Consider an irreducible discrete-time Markov process X on a countable state space S. Suppose there exists a function $V : S \to \mathbb{R}^+$ and a finite set $B \subseteq S$ such that*

$$\mathbb{E}\left[V(X(t+1)) - V(X(t))|X(t) = x\right] \leq -\epsilon;\ \forall x \in B^c,\ \textit{for some } \epsilon > 0, \textit{and}$$
$$\mathbb{E}\left[V(X(t+1)) - V(X(t))|X(t) = x\right] \leq A;\ \forall x \in B,\ \textit{for some } A < \infty.$$

Then the Markov chain $X(t)$ is positive recurrent.

2.3 Continuous-Time Markov Chains

For a continuous-time Markov Chain, transition from the current state to another state occurs after some (continuous) holding time at the current state. In general, these holding times represent the time duration that processing is occurring in the system, and the transitions represent events in the system. Assume the Markov chain enters a state $i \in S$ at time t, then the next state transition occurs at time $t + T_i$, where T_i is the holding time in state i. For the Markov property to hold, at any point of time, the distribution of the time until the next change of state, must be independent of the age of the system in the current state. This means that holding times are memoryless. Since the only probability distribution

function which has this property is the exponential distribution, $\mathbb{P}\{T_i > x\} \leq e^{-q_i x}$ where q_i is the parameter (rate) of the holding time in state i. Hence at time s, the probability that there is a state transition in the interval $(s, s+\delta)$ is $q_i\delta + \mathrm{o}(\delta)$; otherwise, with probability $1 - q_i\delta + \mathrm{o}(\delta)$, no transition occurs.

Now assume that when a transition from state i occurs, the new state is j with probability p_{ij}. Therefore, for $i \neq j$,

$$\mathbb{P}\{X(t+\delta) = j | X(t) = i\} = q_i p_{ij}\delta + \mathrm{o}(\delta)$$

We define $q_{ij} = q_i p_{ij}$. Hence q_{ij} can be thought of the transition rate from i to j,and the average time for a transition from i to j is exponentially distributed with parameter q_{ij}. Note that from these definitions $q_i = \sum_{j \neq i} q_{ij}$, hence to describe the sequence of events (state transitions) we only need q_{ij}'s.

Thus the results of discrete-time Markov chains can be extended to the continuous-time case by discretizing the time by multiples of δ. For example, the global balance equations to find the equilibrium distribution is given by

$$\sum_{j \neq i} \pi_i q_{ij}\delta + \mathrm{o}(\delta) = \sum_{j \neq i} \pi_j q_{ji}\delta + \mathrm{o}(\delta),$$

dividing both sides by δ and noting that $\mathrm{o}(\delta)/\delta \to 0$ as $\delta \to 0$, yields

$$\sum_{j \neq i} \pi_i q_{ij} = \sum_{j \neq i} \pi_j q_{ji}$$

Definition 10 (Generator Matrix of continuous-time Markov chain)**.** The generator matrix is defined as a matrix Q, where $Q_{ij} = q_i j$ and $Q_{ii} = -q_i$.

Then the global balanced equations is equivalently given by $\pi Q = 0$. Since we are looking for solutions that are probability vectors, we also need $\sum_i \pi_i = 1$.

Note that the continuous-time chain is always aperiodic (by construction of discretized version), thus we only need incredibility.

Theorem 3. *Suppose X is irreducible. Then*

(a) All states are transient, or all are positive recurrent, or all are null recurrent.

(b) An equilibrium probability distribution π (a probability vector π that solves $\pi Q = 0$) exists if and only if all states are positive recurrent, and further π is unique.

Similar to the discrete-time case, a reversible continuous-time Markov chain is the one that satisfies the detailed balance equations

$$\pi_i q_{ij} = \pi_j q_{ji} \ \forall i \neq j$$

It is easy to see that the solution to the detailed balance equations also satisfies the global balance equations.

2.4 Continuous-Time Queueing Systems

In general, we use Kendall's notation

$$A/S/n/c$$

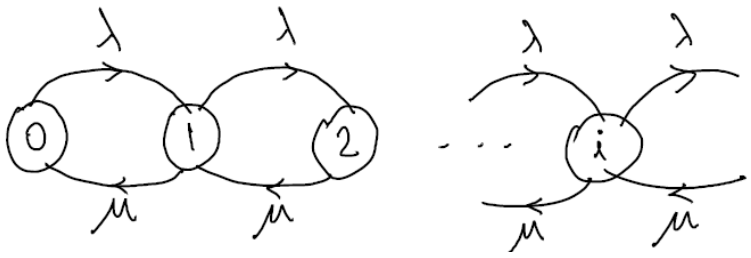
for a queueing system, where

- A stands for the description of the arrival process, (e.g., M stands for Markovian interarrivals, GI for general (any distribution) independent arrivals, etc).
- S stands for the service time distribution, (e.g., M stands for Markovian service, G for general (any distribution) service, etc).
- n stands for the number of servers in the system and can be any integer equal or larger than 1.
- c stands for the maximum number of jobs that can be queued in the system ($c \geq 0$). If this argument is missing, then, by default, the buffer size is infinity.

M/M/1 Queue

The simplest and the easiest queueing system to analyze is the M/M/1 queue. Here the first M stands for Markov arrival (iid exponentially distributed inter-arrival times with rate λ), the second M stands for Markov service (customers need iid exponentially distributed service time with rate μ), and 1 means that there is one server.

The dynamics of the queue is given by a continuous-time Markov chain.



Global balance equations are given by the following. For state $i = 0$,

$$\lambda\pi_0 = \mu\pi_1,$$

for the state $i = 1$

$$(\lambda + \mu)\pi_1 = \lambda\pi_0 + \mu\pi_2$$

therefore

$$\lambda\pi_1 = \mu\pi_2,$$

Proceeding similarly we obtain

$$\lambda\pi_i = \mu\pi_{i+1},$$

Therefore

$$\pi_i = \pi_0(\lambda/\mu)^i$$

Let $\lambda/\mu = \rho$ be the load on the system (utilization). If we impose $\sum_i \pi_i = 1$, then $\pi_i = (1-\rho)\rho^i$, $i \geq 0$, if $\rho < 1$ (stability requirement).

Average queue size:

$$\begin{aligned}
\mathbb{E}[Q] &= \sum_{i=0}^{\infty} i\pi_i \\
&= \sum_{i=0}^{\infty}(1-\rho)i\rho^i \\
&= \rho(1-\rho)\sum_{i=1}^{\infty} i\rho^{i-1} \\
&= \rho(1-\rho)\frac{\mathrm{d}}{\mathrm{d}\rho}(\sum_{i=1}^{\infty}\rho^i) \\
&= \rho(1-\rho)\frac{\mathrm{d}}{\mathrm{d}\rho}(\frac{1}{1-\rho}) \\
&= \rho(1-\rho)(\frac{1}{(1-\rho)^2}) \\
&= \frac{\rho}{1-\rho}
\end{aligned}$$

Average delay: we use Little's Law

$$\mathbb{E}[D] = \mathbb{E}[Q]/\lambda = \frac{1}{\mu-\lambda}$$

Assume we scale the server rate and arrival rate by a factor of n. Then the load remains the same. Thus $\mathbb{E}[Q]$ remains unchanged but $\mathbb{E}[D] = \frac{1}{n}\frac{1}{\mu-\lambda}$ which reduces by a factor of n.

PASTA: Poisson Arrivals See Time Averages

Consider the M/M/1 queue in stationary regime then we know that at any time t is $\mathbb{P}\{Q(t) = k\} = \pi_k$ which is the average fraction of time that the Queue contains k customers. Now suppose we are interested in the behavior of the system from the point of view of the customers that arrive. Assume a customer arrives just after time t, i.e,, it arrives in an interval $(t, t+\delta)$ for some δ small enough. Then

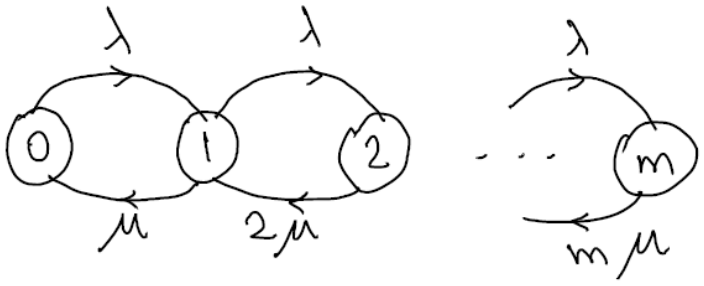
$$\begin{aligned}
\mathbb{P}\{Q(t) = k | A(t,t+\delta) = 1\} &= \frac{\mathbb{P}\{Q(t) = k, A(t,t+\delta) = 1\}}{\mathbb{P}\{A(t,t+\delta) = 1\}} \\
&= \frac{\mathbb{P}\{Q(t) = k\}\mathbb{P}\{A(t,t+\delta) = 1\}}{\mathbb{P}\{A(t,t+\delta) = 1\}} \\
&= \mathbb{P}\{Q(t) = k\} = \pi_k
\end{aligned}$$

where we have used the fact that the arrivals after time t are independent of the queue size at time t (independent increment property of Poisson process). Notice that the same type of argument holds for any other queueing system in stationary regime. The discrete-time counterpart of PASA is called BASTA (Bernoulli Arrivals See Time Averages).

Example 4 (PASTA is not true for general arrival processes)**.** Consider a queue where two customers arrive back-to-back in every $2T$ time units (thus arrivals are correlated not Poisson). Assume each customer needs 1 unit of service. Then $\mathbb{E}[D] = \frac{1}{2}(1) + \frac{1}{2}(2) = \frac{3}{2}$. Arrival rate is $\lambda = 2/(2T) = 1/T$. By Little's Law, $\mathbb{E}[Q] = \lambda \mathbb{E}[D] = \frac{3}{2T}$ which goes to 0 as $T \to \infty$. However, the average queue size seen by the arrivals is $\frac{1}{2}(0) + \frac{1}{2}(1) = \frac{1}{2}$.

M/M/m/m Queue

Arrivals occur according to a Poisson process with rate λ. There are m servers, each working at unit rate. Each arriving customer is assigned to a free server if available, otherwise it is dropped. Each customer requires exponentially distributed service time with parameter μ. Thus the queue size $Q(t)$ is the number of busy servers. $S = \{0, 1, , \cdots, m\}$. Note that the a



transition from state i to $i-1$ occurs whenever any of the i busy servers finishes its service. Therefore the transition rate from i to $i-1$ is $i\mu$. Checking the detailed balance equations, the stationary distribution is given by

$$\pi_k = \frac{\frac{\rho^k}{k!}}{\sum_{i=0}^{m} \frac{\rho^i}{i!}}; k \in S$$

The probability that all the servers are busy is therefore $\pi_m = \frac{\frac{\rho^m}{m!}}{\sum_{i=0}^{m} \frac{\rho^i}{i!}}$. The blocking probability is the probability that a customer does not find a free server upon arrival. By PASTA, this is equal to π_m. This expression for π_m is called the *Erlang-B formula*.

M/G/1 Queue and Pollaczek-Khinchine formula

In the previous examples, we considered queueing system with Poisson arrivals and exponentially distributed service times. Poisson arrivals are in many cases a relatively accurate

model for the arrival process, but exponential service times are not very common in practice, in fact in many applications service times are simply some pre-determined constant. Pollaczek-Khinchine formula extends the theory to the case of generally distributed service times.

In $M/G/1$ queue, the arrivals are per a Poisson process of rate λ. Service times $X_1, X_2, \cdots$ are independent and identically distributed according to a "general" (G) distribution, where X_i is the service time of the i-th arrival. Let $\mathbb{E}[X] = \frac{1}{\mu}$ be the average service time and $\mathbb{E}[X^2]$ be the second moment of service time distribution. Let $\rho = \frac{\lambda}{\mu}$ be the effective load (utilization) of the system, with $\rho < 1$. Then the Pollaczek-Khinchine (PK) formula states that

$$\mathbb{E}[W] = \frac{\lambda \mathbb{E}[X^2]}{2(1-\rho)}$$

where $\mathbb{E}[W]$ is the average waiting time of a customer before it starts getting service.

If we want to model the system by a Markov process, then the state of the system at each time t is a pair $(N(t), A(t))$ where $N(t)$ is the number of customers in the queue and $A(t)$ is the service time already received by the customer in service. Note that if the service times are exponential then the state is simply $N(t)$ because the residual time (the remaining service time of the customer in service) is independent of $A(t)$ by the memoryless property. However for general service time distributions this is not true. The analysis of the Markov process $(N(t), A(t))$ to find the stationary probability is not simple because $A(t)$ takes values in $\mathbb{R}$. Here we use a mean value approach to derive the PK formula for the average quantities.

Consider an arbitrary arrival to the queue in the stationary regime. Let the random variable Q_A denote the number of customers in the queue (excluding the one in service) seen by the arrival. Then the waiting time of the new arrival to get service is

$$W_A = \sum_{i=1}^{Q_A} X_i + R_A,$$

where R_A is the residual service time seen by the arrival. By this we mean the remaining time until the end of the service of the customer which is already in service when the new customer arrives. Taking expectations,

$$\begin{aligned}
\mathbb{E}[W_A] &= \mathbb{E}\left[\sum_{i=1}^{Q_A} X_i\right] + \mathbb{E}[R_A] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{Q_A} X_i | Q_A\right]\right] + \mathbb{E}[R_A] \\
&= \mathbb{E}[Q_A \mathbb{E}[X]] + \mathbb{E}[R_A] \text{ (because service times are independent of } Q_i) \\
&= \mathbb{E}[Q_A] \frac{1}{\mu} + \mathbb{E}[R_A]
\end{aligned}$$

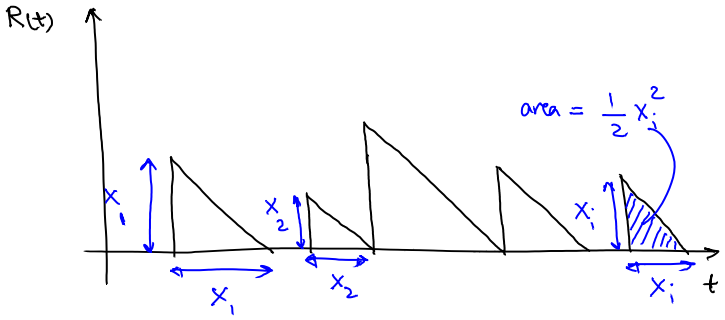
By PASTA, in the stationary regime, the average quantities seen at the arrivals are identical to the time average quantities. Hence, assuming that the time average quantities exist,

$$\mathbb{E}[W] = \mathbb{E}[Q] \frac{1}{\mu} + \mathbb{E}[R],$$

where the expectations above can be understood in terms of time averages. By the Little's Law $\mathbb{E}[Q] = \lambda \mathbb{E}[W]$, hence

$$\mathbb{E}[W] = \frac{\mathbb{E}[R]}{1-\rho}$$

Next, we calculate $\mathbb{E}[R]$ through a graphical argument. Let $R(t)$ denote the residual service time at time t. Then note that when a customer i starts its service, its residual time is X_i and then decays linearly with X_i time units.



Hence over a time interval $[0, T]$

$$\begin{aligned} \frac{1}{T}\int_0^T R(t)\mathrm{d}t &\approx \frac{1}{2T}\sum_{i=1}^{M(T)} X_i^2 \\ &= \frac{1}{2}\frac{M(T)}{T}\frac{1}{M(T)}\sum_{i=1}^{M(T)} X_i^2 \end{aligned}$$

where $M(T)$ is the number of service completions within $[O, T]$. Taking the limit from both sides, as $T \to \infty$, and again assuming that the time averages can be replaced by statistical averages, we have

$$\mathbb{E}[R] = \lim_{T\to\infty}\frac{1}{T}\int_0^T R(t) = \frac{1}{2}\lambda \mathbb{E}\left[X^2\right]$$

This concludes the proof of the PK formula.