Lecture 20: Reversible Processes and Queues

1 Examples of reversible processes

1.1 Birth-death processes

We define two non-negative sequences birth and death rates denoted by $\{\lambda_n : n \in \mathbb{N}_0\}$ and $\{\mu_n : n \in \mathbb{N}_0\}$. A Markov process $\{X_t \in \mathbb{N}_0 : t \in \mathbb{R}\}$ on the state space $\mathbb{N}_0$ is called a *birth-death process* if its infinitesimal transition probabilities satisfy

$$P_{n,n+m}(h) = \begin{cases} \lambda_n h + o(h), & \text{if } m = 1, \\ \mu_n h + o(h), & \text{if } m = -1, \\ o(h), & \text{if } |m| > 1. \end{cases}$$

We say $f(h) = o(h)$ if $\lim_{h \to 0} f(h)/h = 0$. In other words, a birth-death process is any CTMC with generator of the form

$$Q = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & 0 & \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & 0 & 0 & \cdots \\ 0 & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & 0 & \cdots \\ 0 & 0 & \mu_3 & -(\lambda_3+\mu_3) & \lambda_3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Proposition 1.1. *An ergodic birth-death process in steady-state is time-reversible.*

Proof. Since the process is stationary, the probability flux must balance across any cut of the form $A = \{0,1,2,\dots,i\}$, $i \geq 0$. But, this is precisely the equation $\pi_i \lambda_i = \pi_j \mu_j$ since there are no other transitions possible across the cut. So the process is time-reversible. $\square$

In fact, the following, more general, statement can be proven using similar ideas.

Proposition 1.2. *Consider an ergodic CTMC on a countable state space I with the following property: for any pair of states $i \neq j \in I$, there is a unique path $i = i_0 \to i_1 \to \cdots \to i_{n(i,j)} = j$ of distinct states having positive probability. Then the CTMC in steady-state is reversible.*

1.2 Truncated Markov Processes

Consider a transition rate matrix $(Q_{ij})_{i,j \in I}$ on the countable state space I. Given a nonempty subset $A \subseteq I$, the truncation of Q to A is the transition rate matrix $\{Q^A_{ij} : i,j \in A\}$, where for all $i,j \in A$

$$Q^A_{ij} = \begin{cases} Q_{ij}, & j \neq i, \\ -\sum_{k \neq i, k \in A} Q_{ik}, & j = i. \end{cases}$$

Proposition 1.3. *Suppose $\{X_t : t \in \mathbb{R}\}$ is an irreducible, time-reversible CTMC on the countable state space I, with generator $Q = \{Q_{ij} : i, j \in I\}$ and stationary probabilities $\pi = \{\pi_j : j \in I\}$. Suppose the truncation Q^A is irreducible for some $A \subseteq I$. Then, any stationary CTMC with state space A and generator Q^A is also time-reversible, with stationary probabilities*

$$\pi_j^A = \frac{\pi_j}{\sum_{i \in A} \pi_i}, \; j \in A.$$

Proof. It is clear that π^A is a distribution on state space A. We must show the reversibility with this distribution π^A. That is, we must show for all $i, j \in A$

$$\pi_i^A Q_{ij} = \pi_j^A Q_{ji}.$$

However, this is true since the original chain is time reversible. □

1.3 The Metropolis-Hastings algorithm

Let $\{a_j \in \mathbb{R}_+ : j \in [m]\}$ be a set of (known) positive numbers with $A = \sum_{i=1}^m a_i$. Suppose our goal is to build a sampler for a random variable with probability mass function $\pi_j = \frac{a_j}{A}$, for each $j \in [m]$, where m is large and A is difficult to compute directly. This rules out direct evaluation of the fraction $\frac{a_j}{A}$.

Idea. A clever way of (approximately) generating a sample from the distribution $\pi = \{\pi_j : j \in \mathbb{N}\}$ is by constructing an easy-to-simulate Markov chain with limiting (stationary) distribution π. We simply run this Markov chain long enough and return the sample (state) at the end.

Let M be an irreducible transition probability matrix on the integers $[m]$ such that $M = M^T$. An example is the transition matrix of an iid sequence of uniform random variables on $[m]$. Consider the Markov chain $\{X_i : i \in \mathbb{N}\}$ on the state space $[m]$ with the following transition probabilities:

$$P_{ij} = \begin{cases} M_{ij} \min\left(1, \frac{a_j}{a_i}\right), & j \neq i, \\ 1 - \sum_{k \neq i} M_{ik} \left\{1 - \min\left(1, \frac{a_k}{a_i}\right)\right\}, & j = i. \end{cases}$$

Note that the key property that allows us to easily simulate this Markov chain is that only the relative ratios a_j/a_i are required, and not A!

It can be directly verified that (1) this Markov chain is irreducible, and that (2) it is reversible with equilibrium distribution π!

1.4 Random walks on edge-weighted graphs

Consider an undirected graph $G = (I, E)$ with the vertex set I and the edge set E being a subset of unordered pairs of elements from I. Assume having a positive number w_{ij} associated with each edge $\{i, j\}$ in E. Further the edge weight w_{ij} is defined to be 0 if $\{i, j\}$ is not an edge of the graph. Suppose that a particle moves in discrete time, from one vertex to another in the following manner: If the particle is presently at vertex i then it will next move to vertex j with probability

$$P_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}.$$

The Markov chain describing the sequence of vertices visited by the particle is a random walk on an undirected edge-weighted graph. Google's PageRank algorithm, to estimate the relative importance of webpages, is essentially a random walk on a graph!

Proposition 1.4. *Consider an irreducible Markov chain that describes the random walk on an edge weighted graph with a finite number of vertices. In steady state, this Markov chain is time reversible with stationary probability of being in a state $i \in I$ given by*

$$\pi_i = \frac{\sum_j w_{ij}}{\sum_j \sum_k w_{kj}}. \tag{1}$$

Proof. Using the definition of transition probabilities for this Markov chain, we notice that the detailed balance equation for each pair of states $i, j \in I$ reduces to

$$\frac{\alpha_i w_{ij}}{\sum_k w_{ik}} = \frac{\alpha_j w_{ji}}{\sum_k w_{jk}}.$$

From the symmetry of edge weights in undirected graphs, it follows that $w_{ij} = w_{ji}$. Hence, we see that the distribution π defined as in (1) solves the equation, and we get the desired result. □

The following 'dual' result also holds:

Lemma 1.5. *Let $\{X\}_n$ be a reversible Markov chain on a finite state space I and transition probability matrix P. Then, there exists a random walk on a weighted, undirected graph G with the same transition probability matrix P.*

Proof. We create a graph $G = (I, E)$, where $(i, j) \in E$ if and only if $P_{ij} > 0$. We then set edge weights

$$w_{ij} \triangleq \pi_i P_{ij} = \pi_j P_{ji} = w_{ji},$$

where π is the stationary distribution of X. With this choice of weights, it is easy to check that $w_i = \sum_j w_{ij} = \pi_i$, and the transition matrix associated with a random walk on this graph is exactly P. □

2 General queueing theory

The notation A/B/C/D/E for a queueing system indicates

A: Inter-arrival time distribution,

B: Service time distribution,

C: Number of servers,

D: Maximum number of jobs that can be waiting and in service at any time (∞ by default), and

E: Queueing service discipline (FIFO by default).

Theorem 2.1 (PASTA). *Poisson arrivals see time averages. At any time t, we denote a system state by $N(t)$ and the number of arrivals in $[0,t)$ by $A(t)$. The nth arrival instant is denoted by A_n and B a Borel measurable set in $\mathbb{R}_+$, then*

$$\bar{\tau}_B \triangleq \lim_{t \in \mathbb{R}_+} \frac{1}{t} \int_0^t 1\{N(u) \in B\} du = \lim_{n \in \mathbb{N}} \frac{1}{n} \sum_{i=1}^{n} 1\{N(A_i-) \in B\} \triangleq \bar{c}_B.$$

Proof. We will show the special case when $N(t)$ is the number of customers in the system at time t, and $B = \{k\}$. We define for $k \in \mathbb{N}_0$

$$P_k \triangleq \lim_{t \in \mathbb{R}_+} \Pr\{N(t) = k\} \qquad A_k \triangleq \lim_{t \in \mathbb{R}_+} \Pr\{N(t-) = k | A_{k+1} = t\}.$$

Using independent increment property of Poisson arrivals, Baye's rule, and continuity of probabilities, we can write the second limiting probability as

$$A_k = \lim_{t\in\mathbb{R}_+} \lim h \downarrow 0 \frac{\Pr\{N(t-) = k, A(t+h) - A(t) = 1\}}{\Pr\{A(t+h) - A(t) = 1\}} = \lim_{t\in\mathbb{R}_+} \Pr\{N(t) = k\} = P_k.$$

□

Theorem 2.2 (Little's law). *Consider a stable single server queue. Let T_i be waiting time of customer i, $N(t)$ be the number of customers in the system at time t, and $A(t)$ be the number of customers that entered system in duration $[0,t)$, then*

$$\lim_{t\to\infty} \frac{\int_0^t N(u)du}{t} = \lim_{t\to\infty} \frac{\sum_{i=1}^{A(t)} T_i}{A(t)}.$$

Proof. Let $A(t), D(t)$ respectively denote the number of arrivals and departures in time $[0,t)$. Then, we have

$$\sum_{i=1}^{D(t)} T_i \le \int_0^t N(u)du \le \sum_{i=1}^{A(t)} T_i.$$

Further, for a stable queue we have

$$\lim_{t\to\infty} \frac{D(t)}{t} = \lim_{t\to\infty} \frac{A(t)}{t}.$$

Combining these two results, the theorem follows. □

2.1 The M/M/1 queue

The M/M/1 queue is the simplest and most studied models of queueing systems. We assume a continuous-time queueing model with following components.

- There is a single queue for waiting that can accommodate arbitrarily large number of customers.
- Arrivals to the queue occur according to a Poisson process with rate $\lambda > 0$. That is, let A_n be the arrival instant of the nth customer, then the sequence of inter-arrival times $\{A_n - A_{n-1} : n \in \mathbb{N}\}$ is *iid* exponentially distributed with rate λ.
- There is a single server and the service time of nth customer is denoted by a random variable S_n. The sequence of service times $\{S_n : n \in \mathbb{N}\}$ are *iid* exponentially distributed with rate $\mu > 0$, independent of the Poisson arrival process.
- We assume that customers join the tail of the queue, and hence begin service in the order that they arrive *first-in-queue-first-out* (FIFO).

Let $X(t)$ denote the number of customers in the system at time $t \in \mathbb{R}_+$, where "system" means the queue plus the service area. For example, $X(t) = 2$ means that there is one customer in service and one waiting in line. Due to continuous distributions of inter-arrival and service times, a transition can only occur at customer arrival or departure times. Further, departures occur whenever a service completion occurs. Let D_n denote the nth departure from the system. At an arrival time A_n, the number $X(A_n) = X(A_n-) + 1$ jumps up by the amount 1, whereas at a departure time D_n, then number $X(D_n) = X(D_n-) - 1$ jumps down by the amount 1.

For the M/M/1 queue, one can argue that $\{X(t) : t \in \mathbb{R}_+\}$ is a CTMC on the state space $\mathbb{N}_0$. We will soon see that a *stable* M/M/1 queue is time-reversible.

2.1.1 Transition rates

Given the current state $\{X(t) = i\}$, the only transitions possible in an infinitesimal time interval are (a) a single customer arrives, or (b) a single customer leaves (if $i \geq 1$). It follows that the infinitesimal generator for the CTMC $\{X(t)\}_t$ is

$$Q_{ij} = \begin{cases} \lambda, & j = i+1, \\ \mu, & j = i-1, \\ 0, & |j-i| > 1. \end{cases}$$

Since $\lambda, \mu > 0$, this defines an irreducible CTMC.

2.1.2 Equilibrium distribution and reversibility

The M/M/1 queue's generator defines a birth-death process. Hence, if it is stationary, then it must be time-reversible, with the equilibrium distribution π satisfying the detailed balance for each $i \in \mathbb{N}_0$

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

This yields $\pi_{i+1} = \frac{\lambda}{\mu} \pi_i$. Since $\sum_{i \geq 0} \pi = 1$, we must have $\rho \triangleq \frac{\lambda}{\mu} < 1$, giving for each $i \in \mathbb{N}_0$

$$\pi_i = (1-\rho)\rho^i.$$

In other words, if $\lambda < \mu$, then the equilibrium distribution of the number of customers in the system is geometric with parameter $\rho = \lambda/\mu$. We say that the M/M/1 queue is in the *stable* regime when $\rho < 1$. We have thus shown

Corollary 2.3. *The number of customers in an M/M/1 queueing system at equilibrium is a reversible Markov process.*

Further, since M/M/1 queue is a reversible CTMC, the following theorem follows.

Theorem 2.4 (Burke). *Departures from a stable M/M/1 queue are Poisson with same rate as the arrivals.*

2.1.3 Limiting waiting room: M/M/1/K

Consider a variant of the M/M/1 queueing system that has a finite buffer capacity of at most k customers. Thus, customers that arrive when there are already k customers present are 'rejected'. It follows that the CTMC for this system is simply the M/M/1 CTMC truncated to the state space $\{0, 1, \ldots, K\}$, and so it must be time-reversible with stationary distribution $\pi_i = \rho^i / \sum_{j=0}^{k} \rho^j$, $0 \leq i \leq k$.

(Two queues with joint waiting room). *Consider two independent M/M/1 queues with arrival and service rates λ_i and μ_i respectively for $i \in [2]$. Then, joint distribution of two queues is*

$$\pi(n_1, n_2) = (1-\rho_1)\rho_1^{n_1}(1-\rho_2)\rho_2^{n_2}, \quad n_1, n_2 \in \mathbb{N}_0.$$

Suppose both the queues are sharing a common waiting room, where if arriving customer finds R waiting customer then it leaves. In this case,

$$\pi(n_1, n_2) = (1-\rho_1)\rho_1^{n_1}(1-\rho_2)\rho_2^{n_2}, \quad (n_1, n_2) \in A \subseteq \mathbb{N}_0 \times \mathbb{N}_0.$$