STATISTICAL COMPUTATIONAL METHODS

Seminar Nr. 6, Queuing Systems

1. Performance of a car wash center is modeled by the B1SQP with 2-minute frames. Cars arrive every 10 minutes, on the average, and the average service time is 6 minutes. There are no cars at the center at 10:00 a.m., when the center opens. What is the probability that at 10:04 one car is being washed and another is waiting?

Solution:

We have

$$\begin{aligned}
\Delta &= 2 \text{ min.}, \\
\lambda_A &= 1/\mu_A = 1/10 \text{ min}^{-1}, \\
\lambda_S &= 1/\mu_S = 1/6 \text{ min}^{-1}, \text{ so} \\
p_A &= \lambda_A \Delta = 1/5 \text{ and} \\
p_S &= \lambda_S \Delta = 1/3.
\end{aligned}$$

There are no cars at the beginning, so the value of $X_0 = 0$, i.e. the initial distribution of X_0 is

$$P_0 = [1\ 0\ 0 \ldots].$$

There are 2 frames between 10:00 am and 10:04 am. We want the conditional probability

$$P(X_2 = 2 \mid X_0 = 0).$$

Since the number of cars at the wash center can change by at most 1 during each frame, this probability equals

$$\begin{aligned}
p_{02}^{(2)} &= [\text{1st row of } P][\text{3rd column of } P] = [1 - p_A\ \ p_A\ \ 0] \begin{bmatrix} 0 \\ p_A(1-p_S) \\ (1-p_A)(1-p_S) + p_A p_S \end{bmatrix} \\
&= p_A^2(1-p_S) = \frac{1}{25} \cdot \frac{2}{3} = \frac{2}{75} = 0.0267.
\end{aligned}$$

2. A metered parking lot with two parking spaces is modeled by a Bernoulli two-server queuing system with capacity limited by two cars and 30-second frames. Cars arrive at the rate of one car every 4 minutes and each car is parked for 5 minutes, on the average.

a) find the transition probability matrix for the number of parked cars;

b) find the steady-state distribution for the number of parked cars;

c) what fraction of the time are both parking spaces vacant?

d) what fraction of arriving cars will not be able to park?

e) every 2 minutes of parking costs 25 cents; assuming all drivers use all the parking time they pay for, how much money is the parking lot going to raise every 24 hours?

Solution:

This is a B2SQS with

$$C = 2, k = 2, \Delta = 1/2 \text{ min}, \lambda_A = 1/4 \text{ min}^{-1} \text{ and } \lambda_S = 1/5 \text{ min}^{-1}.$$

a) $p_A = \lambda_A \Delta = 1/8$ and $p_S = \lambda_S \Delta = 1/10$. There are 3 states $\{0, 1, 2\}$ and the transition probabilities are

$$\begin{aligned}
p_{00} &= 1 - p_A = 7/8, \\
p_{01} &= p_A = 1/8, \\
p_{02} &= 0; \\
p_{10} &= p_S(1 - p_A) = 7/80, \\
p_{11} &= (1 - p_A)(1 - p_S) + p_A p_S = 4/5, \\
p_{12} &= p_A(1 - p_S) = 9/80; \\
p_{20} &= p_S^2(1 - p_A) = 7/800, \\
p_{21} &= 2p_S(1 - p_A)(1 - p_S) + p_S^2 p_A = 127/800, \\
p_{22} &= (1 - p_A)(1 - p_S)^2 + 2p_A p_S(1 - p_S) + p_A(1 - p_S)^2 = 333/400.
\end{aligned}$$

So, the transition probability matrix is

$$P = \begin{bmatrix} 7/8 & 1/8 & 0 \\ 7/80 & 4/5 & 9/80 \\ 7/800 & 127/800 & 333/400 \end{bmatrix}.$$

b) The steady-state distribution is

$$\pi = [\pi_0 \ \ \pi_1 \ \ \pi_2] = [0.3089 \ \ 0.4135 \ \ 0.2777].$$

c) That would be

$$P(X = 0) \ = \ \pi_0 \ = \ 0.3089,$$

or 30.89% of the time.

d) A car cannot park if both spaces are taken, so

$$P(X = 2) \ = \ \pi_2 \ = \ 0.2777,$$

or 27.77% of cars.

e) The expected number of parked cars is

$$E(X) \ = \ \sum_0^2 x\pi_x \ = \ 0 \cdot 0.3089 + 1 \cdot 0.4135 + 2 \cdot 0.2777 \ = \ 0.9689.$$

Then the total revenue in 24 hours is

$$E(X) \cdot 24 \cdot 60/2 \cdot 0.25 \ = \ 174.4020 \ \text{ dollars.}$$

3. Trucks arrive at a weigh station according to a Poisson process with average rate of 1 truck every 10 minutes. Inspection times are Exponential with the average of 3 minutes. When a truck is on scale, the other arrived trucks stay in line waiting for their turn. Compute

a) the expected number of trucks at the weigh station at any time;

b) the proportion of time when the weigh station is empty;

c) the expected time each truck spends at the station, from arrival to departure;

d) the fraction of time there are fewer than 2 trucks in the weigh station.

<u>Solution:</u>

A Poisson process of arrivals implies Exponential interarrival times, so the described system is M/M/1 with $\mu_A = 10$ min and $\mu_S = 3$ min. Hence, $\lambda_A = 1/10 \text{ min}^{-1}$, $\lambda_S = 1/3 \text{ min}^{-1}$ and $r = \lambda_A/\lambda_S = 0.3 < 1$.

a) The expected number of trucks at the weigh station is

$$E(X) \;=\; \frac{r}{1-r} \;=\; 3/7 \;=\; 0.4286.$$

b) The proportion of time when the weigh station is empty is

$$P(X=0) \;=\; 1-r \;=\; 0.7,$$

or 70% of time.

c) This is the expected response time

$$E(R) \;=\; \frac{\mu_S}{1-r} \;=\; \frac{3}{0.7} \;=\; 30/7,$$

or 4.2857 minutes.

d) This is the probability

$$\begin{aligned} P(X<2) \;&=\; P(\{X=0\} \cup \{X=1\}) \;=\; P(0)+P(1) \\ &=\; \pi_0+\pi_1 \;=\; (1-r)+r(1-r) \\ &=\; 1-r^2 \;=1-0.09 \;=\; 0.91, \end{aligned}$$

or 91% of the time.

4. A toll area on a highway has three toll booths and works as an M / M / 3 queuing system. On the average, cars arrive at the rate of one car every 5 seconds, and it takes 12 seconds to pay the toll, not including the waiting time. Compute the fraction of time when there are ten or more cars waiting in the line.

Solution:

We have

$$\lambda_A \;=\; 1/5\,\mathrm{sec}^{-1},\; \lambda_S \;=\; 1/12\,\mathrm{sec}^{-1},\; k \;=\; 3 \text{ and } r \;=\; \lambda_A/\lambda_S \;=\; 12/5 \;=\; 2.4 \;<3.$$

We want to compute

$$P(X_w \geq 10).$$

Since there are 3 toll booths (all busy), this is the same as

$$P(X \geq 13) \;=\; \sum_{x=13}^{\infty} \pi_x.$$

Now, for the steady-state distribution, we have

$$\begin{aligned}
\pi_0 &= \frac{1}{1 + r + \dfrac{r^2}{2} + \dfrac{r^3}{6(1 - r/3)}} \;=\; 0.0562, \\
\pi_1 &= \ldots, \\
\pi_2 &= \ldots, \\
\pi_x &= \frac{r^3}{3!}\pi_0 \left(\frac{r}{3}\right)^{x-3}, \text{ for any } x \geq 3.
\end{aligned}$$

So,

$$\begin{aligned}
P(X_w \geq 10) &= \frac{r^3 \pi_0}{3!} \sum_{x=13}^{\infty} \left(\frac{r}{3}\right)^{x-3} \\
&= \frac{r^3 \pi_0}{6} \cdot \frac{(r/3)^{10}}{1 - r/3} \;=\; 0.0695,
\end{aligned}$$

or 6.95% of the time. Note that we used the formula for the sum of a Geometric series with ratio $q < 1$,

$$\sum_{k=a}^{\infty} q^k \;=\; q^a \sum_{k=a}^{\infty} q^{k-a} \;=\; q^a \sum_{i=0}^{\infty} q^i \;=\; \frac{q^a}{1-q}.$$

5. Sports fans tune to a local sports radio station according to a Poisson process with the rate of three fans every two minutes and listen to it for an Exponential amount of time with the average of 20 minutes.
a) what queuing system is the most appropriate for this situation?
b) compute the expected number of concurrent listeners at any time;
c) find the fraction of time when 40 or more fans are tuned to this station.

Solution:

We have

$$\lambda_A \;=\; 3/2 \,\text{min}^{-1},\; \lambda_S \;=\; 1/20 \,\text{min}^{-1},\; k \;=\; \infty \text{ and } r \;=\; \lambda_A/\lambda_S \;=\; 60/2 \;=\; 30.$$

a) We have a Poisson process of arrivals (so Exponential interarrival times), Exponential service times and infinitely many servers (because any number of people can listen a radio station simultaneously), therefore, this is an M / M / ∞ queuing system.

b) The expected number of concurrent listeners is

$$E(X) \;=\; r \;=\; 30 \;\text{ listeners.}$$

c) The number of concurrent listeners, X, has Poisson distribution with parameter $r = 30$. So,

$$P(X \geq 40) \;=\; 1 - P(X < 40) \;=\; 1 - P(X \leq 39) \;=\; 1 - \mathit{poisscdf}(39, 30) \;=\; 0.0463,$$

or 4.63% of the time.

6. Messages arrive at an electronic mail server according to a Poisson process with the average frequency of 5 messages per minute. The server can process only one message at a time and messages are processed on a "first come $-$ first serve" basis. It takes an Exponential amount of time M_1 to process any text message, plus an Exponential amount of time M_2, independent of M_1, to process attachments (if there are any), with $E(M_1) = 2$ seconds and $E(M_2) = 7$ seconds. Forty percent of messages contain attachments. Use Monte Carlo methods to estimate
a) the expected response time of this server;
b) the expected waiting time of a message before it is processed.

Solution:

The process of arrivals is Poisson, so the interarrival times are Exponential with an average of 5 messages per minute (i.e. $\mu_A = 5$ minutes). Notice that because of the attachments, the overall service time is *not* Exponential, so this system is *not* M / M / 1, it is M / G / 1, where "G" stands for "general". For service times, we have $\mu_{M_1} = 2$ seconds and $\mu_{M_2} = 7$ seconds.
We will keep track of arrival times, of times when service starts and times when service finishes (departure times), as arrays. In addition, we need to know when the server becomes available. We start with the parameters and initialization of variables

```
% Generate an M/G/1 queuing system.
clear all

% Parameters of the system
```

```
lamA = 5/60; % arrival rate lambdaA, per sec.
lamM1 = 1/2; % parameter of M1 (texts); lambdaM1
lamM2= 1/7; % parameter of M2 (attachments); lambdaM2
p = 0.4; % proportion of emails with attachments
N = input('size of MC study = '); % size of the MC study:
                                  % number of generated jobs

% Initialize variables
arrival = zeros(1, N); % arrival times
start = zeros(1, N); % times when service starts
finish = zeros(1, N); % times when service finishes;
                      % departure times
T = 0; % arrival time of a new job
A = 0; % time when the server becomes available
```

Then each new arrival time is an $Exp(\lambda_A)$ variable. For service time, it is $M_1 \in Exp(\lambda_{M_1})$ plus the service time for attachments $M_2 \in Exp(\lambda_{M_2})$, with probability $p = 0.4$. Whether or not we add the time for processing attachments is a Bernoulli variable with parameter $p = 0.4$,

$$\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}.$$

Remember that this is simulated as $U < p$ ("success"), for a Standard Uniform variable U. So the *total* service time will be $M_1 + (U < p) * M_2$. For each job, the service starts either when the job arrives or when the server becomes available, whichever happens *last*.

```
for j = 1 : N; % do-loop over N jobs
    T = T - 1/lamA*log(rand); % arrival time of the j-th job
    S = -1/lamM1*log(rand) - (rand < p)*1/lamM2*log(rand);
                      % service time of the j-th job
    arrival(j) = T; % arrival time of the j-th job
    start(j) = max(A, T); % time when service starts
    finish(j) = start(j) + S; % departure time
    A = finish(j); % time when the server becomes available
                   % to take the (j+1)st job
end
```

Then we summarize the results.

```
% Output
fprintf('a) expected response time E(R) is %3.5f sec.\n', ...
            mean(finish - arrival))
fprintf('b) expected waiting time E(W) is %3.5f sec.\n', ...
            mean(start - arrival))
```

Run it a few times with $N = 5e4, 1e5, 5e5$. We can estimate the expected response time as ≈ 8.8 seconds and the expected waiting time ≈ 4 seconds.

7. A small clinic has several doctors on duty, but only one patient is seen at a time. Patients are scheduled to arrive at equal $15-$minute intervals, are then served in the order of their arrivals and each of them needs a Gamma time with the doctor, that has parameters $\alpha = 4$ and $\lambda = 10/3 \text{ min}^{-1}$. Use Monte Carlo simulations to estimate
a) the probability that a patient has to wait before seeing the doctor;
b) the expected waiting time for a patient;

Solution:

This system is *not* M / M / 1 either, both because the interarrival times are not random and because the service times are not Exponential. This is a D / G / 1 queuing system, where "D" stands for "deterministic" (fixed).
Now arrival times are fixed, at increments of 15 minutes. Service times have a $Gamma(\alpha, \lambda)$ distribution, which can be simulated as a sum of α $Exp(1/\lambda)$ variables. Let us notice that the mean interarrival time is $\mu_A = 15$ minutes, while the average service time is $\mu_S = \alpha \cdot \lambda = 40/3 = 13.33$, so the system is functional $r = \mu_S/\mu_A < 1$. Also, the standard deviation of service times is $\sqrt{\alpha \cdot \lambda^2} = \dfrac{20}{3} \approx 6.67$, quite large, so we can expect quite a bit of variation.
Again, we keep track of arrival times, of times when service starts, times when service finishes (departure times), as arrays and time when the server becomes available.

```
% Generate a  D/G/1 queuing system.
clear all

% Parameters of the system
alpha = 4; lambda = 10/3; % parameters of service times
t = 15; % fixed interarrival times
N = input('size of MC study = '); % size of the MC study:
```

```
                                  % number of generated jobs
arrival = 0 : t : (N - 1)*t; % arrival times = 0, t, 2t, 3t, ...

% Initialize variables
start = zeros(1, N); % times when service starts
service = zeros(1, N); % service time for each job
finish = zeros(1, N); % times when service finishes;
                      % departure times
A = 0; % time when the doctor becomes available

for j = 1 : N % do-loop over N jobs
    start(j) = max(A, arrival(j)); % time when service starts
    service(j) = -lambda*sum(log(rand(alpha,1)));
      % service time for each job, Gamma(alpha, lambda) distr.
    finish(j) = start(j) + service(j); % departure time
    A = finish(j); % time when the server (doctor) becomes
                  % available to take the (j+1)st job (patient)
end

% Output
fprintf('a) prob. that a patient has to wait P(W > 0) ...
         is %3.5f\n', mean(start > arrival))
fprintf('b) expected waiting time E(W) is %3.5f min.\n', ...
         mean(start - arrival))
```

After several runs, we find the probability that a patient has to wait to be ≈ 0.66 and the average waiting time for a patient ≈ 10.6 minutes.

8. Assume that the clinic in Problem 7. is only open between 8 a.m. and 6 p.m. to receive patients. Use Monte Carlo methods to estimate

a) the expected waiting time for a patient;

b) the longest waiting time for a patient;

c) the number of patients still in the clinic at 6 p.m.

Solution:

Now we have a limit on the arrival time. Other than that, we basically put the previous code in a loop.

```
% Generate a  D/G/1 queuing system with time limit.
clear all
% Parameters
alpha = 4; lambda = 10/3; % parameters of service times
t = 15; % fixed interarrival times

% Initialize variables
Tmax = input('maximum time (in minutes) = '); % limited max. time
N = input('size of MC study = '); % size of the MC study:
                                  % number of generated jobs
Nmax = Tmax/15; % max nr. of scheduled patients
arrival = 0 : t : (Nmax - 1)*t; % arrival times = 0, t, 2t,
                                % 3t, ...
expw = zeros(1, N); % expected waiting time
maxw = zeros(1, N); % maximum waiting time
nr_p = zeros(1, N); % number of patients in the clinic at
                    % the end of the day
for i = 1 : N
    j = 0; % job number
    T = 0; % arrival time of a new job
    A = 0; % time when the doctor becomes available
    start = zeros(1, Nmax); % times when service starts
    service = zeros(1, Nmax); % service time for each job
    finish = zeros(1, Nmax); % times when service finishes;
                             % departure times
    while T < Tmax; % until the end of the day
        j = j + 1; % next job
        T = T + t; % arrival time of job j
        start(j) = max(A, arrival(j)); % time when service starts
        S = -lambda*sum(log(rand(alpha,1)));
        % service time for each job, Gamma(alpha, lambda) distr.
        finish(j) = start(j) + S; % departure time
        A = finish(j); % time when the server (doctor) becomes
```

```
            % available to take the (j+1)st job (patient)
    end

    expw(i) = mean(start - arrival);
    maxw(i) = max(start-arrival);
    nr_p(i) = sum(finish > Tmax);
end

% Output
fprintf('a) expected waiting time E(W) is %3.2f min.\n', ...
         mean(expw)); % expected waiting time
fprintf('b) longest waiting time is %3.2f min.\n', ...
         mean(maxw)); % the longest waiting time
fprintf('c) number of patients at time Tmax is %3.2f \n', ...
         mean(nr_p)); % number of jobs at time Tmax
```

Between 8 a.m. and 6 p.m. there are 10 hours, so $Tmax = 600$ minutes. Try a smaller number of simulations $N = 1e4, 5e4, 1e5$. We estimate the average waiting time for a patient ≈ 7.9 minutes (slightly smaller than in the unlimited capacity case), longest waiting time ≈ 28.4 minutes and there seems to be at least 1 patient in the clinic at the end of the workday.