STOCHASTIC NETWORKS EXAMPLE SHEET 1 SOLUTIONS

ELENA YUDOVINA

Exercise 1. Determine the stationary distribution, π, of an $M/M/2$ queue. Show that the proportion of time both servers are idle is

$$\pi_0 = \frac{1-\rho}{1+\rho}, \text{ where } \rho = \frac{\nu}{2\mu}.$$

Proof. The transition rates are $q(j, j+1) = \nu$, $q(j, j-1) = 2\mu$ if $j \geq 2$, $q(1, 0) = \mu$. Detailed balance equations read

$$\pi_0\nu = \pi_1\mu, \quad \pi_n\nu = \pi_{n+1}2\mu \text{ for } n \geq 1$$

or

$$\pi_1 = 2\rho\pi_0, \quad \pi_{n+1} = \rho\pi_n = 2\rho^{n+1}\pi_0.$$

Normalizing for $\sum \pi_n = 1$ we get

$$1 = \pi_0(1 + 2\rho + 2\rho^2 + \dots) = \pi_0(1 + \frac{2\rho}{1-\rho}) = \pi_0\frac{1+\rho}{1-\rho},$$

hence $\pi_0 = \frac{1-\rho}{1+\rho}$ and $\pi_n = 2\rho^n\pi_0$. □

Exercise 2. Upon an $M/M/1$ queue is imposed the additional constraint that arriving customers who find N customers already present leave and never return. Find the stationary distribution of the queue.

Proof. The transition rates are $q(j, j+1) = \nu$ if $j < N$ and $q(j, j-1) = \mu$ if $j > 0$. The detailed balance equations are

$$\pi_n\nu = \pi_{n+1}\mu \quad n = 0, 1, \dots, N-1$$

with solution $\pi_{n+1} = \rho\pi_n = \rho^{n+1}\pi_0$. Normalizing for $\sum \pi_n = 1$ gives

$$\pi_0(1 + \rho + \dots + \rho^N) = 1 \implies \pi_0 = \frac{1-\rho}{1-\rho^{N+1}}, \ \pi_n = \rho^n\pi_0.$$

□

Exercise 3. A continuous time Markov process has transition rates $(q(j, k),\ j, k \in S)$, and equilibrium distribution $(\pi(j),\ j \in S)$. Write down the equations relating $q(\cdot, , \cdot)$ and $\pi(\cdot)$. A discrete time Markov chain is formed by observing the jumps of this process: at the successive jump times of the process, the state j just before, and the state k just after, the jump are recorded as an ordered pair $(j, k) \in S^2$. Write down the transition probabilities of the resulting Markov chain, and show that it has equilibrium distribution

$$\pi'(j, k) = G^{-1}\pi(j)q(j, k)$$

provided

$$G = \sum_j \sum_k \pi(j)q(j, k) < \infty.$$

Give an alternative interpretation of $\pi'(j,k)$ in terms of the conditional probability of seeing the original process jump from state j to state k in the interval $(t, t+h)$, given that a jump occurs in that interval.

Proof. For the continuous-time Markov process we have

$$\pi(j)\sum_k q(j,k) = \sum_k \pi(k)q(k,j).$$

For the jump chain,

$$p_{(j_1,k_1),(j_2,k_2)} = \begin{cases} 0, & j_2 \neq k_1 \\ q(j_2,k_2)/q(j_2), & j_2 = k_1, \end{cases}$$

where we define $q(j) = \sum_k q(j,k)$. We check that π' is the equilibrium distribution:

$$\pi'(j,k) \stackrel{?}{=} \sum_l \pi'(l,j)p_{(l,j),(j,k)}$$
$$G^{-1}\pi(j)q(j,k) \stackrel{?}{=} \sum_l G^{-1}\pi(l)q(l,j)q(j,k)/q(j)$$

Cancel G^{-1} and $q(j,k)$, move $q(j)$ to the left-hand side and expand into $\sum_l q(j,l)$, and this will be the equilibrium equations for the original chain.

Remark. In discrete time, equilibrium balance equations are

$$\pi(n) = \sum_{k\in S} \pi(k)p_{kn},$$

where k may be equal to n. This is equivalent to the form

$$\pi(n)\sum_{k\in S} p_{nk} = \sum_{k\in S} \pi(k)p_{kn},$$

since $\sum_{k\in S} p_{nk}$ is the probability of going somewhere from state n, and is equal to 1. (In the discrete case, we allow $p_{nn} > 0$.) In the continuous case, we use the second form,

$$\pi(n)\sum_{k\neq n} q_{nk} = \sum_{k\neq n} \pi(k)q_{kn},$$

because the sum of the rates $\sum_{k\in S} q_{nk}$ can be anything it likes. Also, definitionally, $q_{nn} = 0$.

$\pi'(j,k)$ is precisely the (limit as $h \to 0$ of the) conditional probability that, given that a jump occurs in $(t, t+h)$, it is from j to k. Indeed, the probability that there is a jump from j to k in $(t, t+h)$ is

$$\pi(j)q(j,k)(h + o(h)) + o(h).$$

The first $o(h)$ term comes from approximating $\exp(q(j,k)h) - 1$ by a first order Taylor expansion, and the second $o(h)$ corresponds to the probability of two or more jumps of the system in the time interval $(t, t+h)$. The probability that there is some jump in $(t, t+h)$ is the sum of these over all j and k. Taking the ratio and sending $h \to 0$ gives the result. □

Exercise 4. An $M/M/1$ queue has arrival rate ν and service rate μ, where $\rho = \nu/\mu < 1$. Show that the sojourn time (= queueing time + service time) of a typical customer is exponentially distributed with parameter $\mu - \nu$.

Proof. We compute the moment generating function of W, the sojourn time, by conditioning on the number of people present in the queue when the customer arrives:

$$\mathbb{E}[e^{zW}|n \text{ customers in queue}] = \mathbb{E}[e^{z(S_1+\ldots+S_n+S_{n+1})}] = (\mathbb{E}[e^{zS_1}])^{n+1} = (1-z/\mu)^{-(n+1)}$$

Here, I'm using the fact that for an exponential random variable S_1,

$$\mathbb{E}[e^{zS_1}] = \int_0^\infty e^{zx}\mu e^{-\mu x}dx = \int_0^\infty \mu e^{-(\mu-z)x}dx = \frac{\mu}{\mu-z} = (1-z/\mu)^{-1}.$$

Now, the number of people in the queue when the typical customer arrives has the stationary geometric distribution (since Poisson arrivals see time averages). Therefore,

$$\mathbb{E}[e^{zW}] = \sum_{n=0}^{\infty}(1-\rho)\rho^n(1-z/\mu)^{-(n+1)} = (1-\rho)(1-z/\mu)^{-1}\frac{1}{1-\rho/(1-z/\mu)}$$

Rewriting (using the fact that $\rho = \nu/\mu$), we get

$$\mathbb{E}[e^{zW}] = \frac{\mu-\nu}{\mu-\nu-z},$$

which is the moment generating function of an exponential random variable with rate $\nu-\mu$. (Recall that the moment generating function determines the distribution uniquely.)

Alternatively, you can show that the hazard rate,

$$\mathbb{P}(\text{waiting time ends in } (t,t+\delta t)|\text{still waiting at time } t),$$

is $(\mu-\nu)\delta t + o(\delta t)$. Note that if the waiting time has probability density f and cumulative probability $F = \int f$, then this conditional probability is (approximately, as $\delta \to 0$) equal to $f/(1-F) = (-\log(1-F))'$, so it determines F. On the other hand, we can compute the conditional probability by noting that the probability that some departure happens in $(t,t+\delta t)$ is $\mu\delta t + o(\delta t)$, and the probability that our customer is the first in line (so is the one departing) is the probability that the geometric number of customers in the queue in front of him is actually 0, which is $1-\rho$. Thus, the probability that our customer leaves in $(t,t+\delta t)$ given that he hasn't left by time t is $\mu(1-\nu/\mu)\delta t + o(\delta t) = (\nu-\mu)\delta t + o(\delta t)$ as required. □

Exercise 5. Consider an $M/M/\infty$ queue with servers numbered $1,2,\ldots$ On arrival a customer chooses the lowest numbered server which is free. Calculate the equilibrium probability that j out of the first n servers are busy. For what fraction of the time is each server busy?

Car parking spaces are labelled $n = 1,2,\ldots,$ where the label indicated the distance (in car lengths) to walk to a shop, and an arriving car parks in the lowest numbered free space. Cars arrive as a Poisson process of rate ν, and parking times are exponentially distributed with unit mean, and are independent of each other and of the arrival process. Show that the distance parked from the shop has mean

$$\sum_{C=0}^{\infty}\mathcal{E}(\nu,C),$$

where $\mathcal{E}(\nu,C)$ is Erlang's formula.

Proof. If we are interested in the number of busy servers among the first n servers, then we have the same model as for the telephone exchange with n lines. Therefore, the probability that j out of the first n servers are busy is

$$\pi_{jn} = \frac{\nu^j/j!}{\sum_{i=0}^n \nu^i/i!}.$$

To determine when a server $n \geq 1$ is busy, let I_n be the indicator of this event: then $\mathbb{P}(n \text{ is busy}) = \mathbb{E}I_n$. Also, let N_n be the number of servers busy among the first n; that is, $\pi_{jn} = \mathbb{P}(N_n = j)$. Now, clearly $I_n = N_n - N_{n-1}$, so $\mathbb{E}[I_n] = \mathbb{E}[N_n] = \mathbb{E}[N_{n-1}]$, or

$$\mathbb{E}[I_n] = \sum_{j=0}^{n} j\pi_{jn} - \sum_{j=0}^{n-1} j\pi_{j,n-1}.$$

(Of course, we can start the sums at 1 rather than at 0.) For example,

$$\begin{aligned} \mathbb{E}[I_1] &= \pi_{11} &&= \tfrac{\nu}{1+\nu}, \\ \mathbb{E}[I_2] &= \pi_{12} + 2\pi_{22} - \pi_{11} &&= \tfrac{\nu+\nu^2}{1+\nu+\nu^2/2} - \tfrac{\nu}{1+\nu}, \end{aligned}$$

and so on.

For the car parking process, the probability that I park in spot $\geq n+1$ is the probability that all of the first n spots are occupied. Therefore,

$$\mathbb{E}d = \sum_{n=0}^{\infty} \mathbb{P}(d \geq n+1) = \sum_{n=0}^{\infty} \pi_{nn} = \sum_{n=0}^{\infty} \mathcal{E}(\nu, n).$$

as required.

Remark. I was asked in the examples class why we shouldn't consider the following question instead: at a fixed time, look at all the cars parked, and compute their average distance from the shop (set to 0 if the system is empty), then take the expectation of the result. That is, we would be interested in

$$\mathbb{E}\frac{\sum jI_j}{\sum I_j} \text{ rather than } \mathbb{E}\min\{j : \ I_j = 0\}.$$

Which one we consider is a matter of taste or interpretation of the wording of the question. There is certainly no reason to expect that the two would be equal as random variables, or even in expectation: I don't really want to compute the expectation for this problem, but will do it for a simpler model.

Consider a system with N parking spaces, and suppose that space n is busy or free with probability $1/2$ independently of all the other spaces. Then

$$\mathbb{E}\min\{j : \ I_j = 0\} = 2$$

since this is a geometric random variable with parameter $1/2$; on the other hand,

$$\mathbb{E}\frac{\sum jI_j}{\sum I_j} = \mathbb{E}\mathbb{E}\left[\frac{\sum jI_j}{\sum I_j} | \sum I_j = n\right].$$

Now, conditional on their being n busy spaces, they are distributed independently, so the average index of a busy space is $N/2$ independently of n. Therefore, in this example

$$\mathbb{E}\frac{\sum jI_j}{\sum I_j} = N/2.$$

(We would generally expect that while the first gap is at quite a small parking space, there are also cars in quite large locations.) Incidentally, we could also look at

$$\frac{\mathbb{E}\sum jI_j}{\mathbb{E}\sum I_j},$$

which in this case would give us the same answer, but in general would give a third potential "average distance".

□

Exercise 6. Goldie's Restaurant remains open 24 hours per day, 365 days per year. The total number of customers served in the restaurant during 2007 was 21% greater than the total for 2006. In each year, the number of customers in the restaurant was recorded at a large number of randomly selected times, and the average of those numbers in 2007 was 16% greater than the average in 2006. By how much did the average duration of a customer visit to the restaurant increase or decrease?

Proof. Recall Little's Law $L = \lambda W$, where L is the average number in the system, λ is the average arrival rate, and W is the average waiting time. We are given $L_{2007} = 1.16L_{2006}$ and $\lambda_{2007} = 1.21\lambda_{2006}$. Therefore, $W_{2007} = \frac{1.16}{1.21}W_{2006} \approx 0.95W_{2006}$, i.e. the average duration of a visit decreased by about 5%. □

Exercise 7. Does the Post Office arrangement (one queue for S counters rather than S queues) reduce the expected waiting time? What advantage does it have?

Proof. It doesn't reduce the expected waiting time, but does reduce the variance of it.

We are going to model the situation as follows: suppose the service times of customers are iid, and are drawn not by the customers themselves but by the servers when the customer comes up to be served. That is, each server has a sequence of (random) times, and he will pull out one of them at the start of each service. We will also assume that if a server is idle, then in fact nobody is queueing – that is, we allow customers to jump queues, and insist that they jump queues if otherwise a server would go idle.

In this case, it is easy to see that any pair of systems with the same arrival times $t_1 < t_2 < \dots$ will have the same set of departure times $d_1 < d_2 < \dots$ at which customers leave the queue and go into service. (We will also have the same set of times at which someone leaves the system entirely.) In this case, it's clear that the average waiting time is the same in both systems. For example, if there is a regeneration point when the system is empty (this means $t_{N+1} > d_N$) then sum of the waiting times for the N customers during a single regeneration period is simply $\sum_{j=1}^{N} d_j - t_j$, which doesn't depend on the order in which the customers leave.

On the other hand, the Post Office arrangement means that if customer j comes into the queue before k, then j will also leave the queue (and go into service) before k. We claim that this is optimal in terms of minimizing the expectation of the square of the waiting time (and since the mean is the same, this means optimizing the variance).

We work in the model as above (same set of arrival times and same set of departure times), over a single regeneration period (empty to empty), and look at two customers $j < k$ who arrived at times $t_j < t_k$, and left at times $d_m > d_n$ respectively. Then the waiting time of j in the queue is $w_j = (d_m - t_j)$ and the waiting time of k is $w_k = (d_n - t_k)$, where $t_j < t_k < d_n < d_m$.

Now consider an alternative arrangement, which is closer to first-in-first-out, where we swap j and k as they go into service. This will mean that the customer that leaves the queue (and goes into service) at time d_n is j, and the customer that leaves at time d_m is k. In that case, the waiting time of j is $\tilde{w}_j = (d_n - t_j)$ and the waiting time of k is $\tilde{w}_k = (d_m - t_k)$, where still $t_j < t_k < d_n < d_m$.

It is not hard to check that $\tilde{w}_j^2 + \tilde{w}_k^2 < w_j^2 + w_k^2$ (this is convexity of the function $x \mapsto x^2$).

We have shown that if j and k depart out of order, then by swapping them we can reduce the variance of the waiting time distribution.

Note that we have considered the times that the customers wait in the queue, ignoring their own service time. However, because their service time is independent of the time they've spent waiting in the queue, this doesn't make a difference (the means and the variances will just add).

Also, our result will no longer be true if the customers are allowed to know each other's processing times as they come into the system. For example, if the jobs go into service in order of shortest processing time first, the average waiting time will be reduced. (So it's socially optimal to let that person with a really small shopping basket go in front of you!)

Remark. This was originally published by John Kingman in the Mathematial Proceedings of the Cambridge Philosophical Society, in 1962, when John Kingman was a Part III student.

□