Old exam questions related to the Analysis of Queueing Systems with Sample Paths and Simulation

Nicky D. van Foreest

April 2, 2020

This is pretty unorganized. If you feel inclined to help clean it up, do not hesitate to start sending PRs on github.

0.1 OLD EXAM QUESTIONS

0.1.1 *Multiple-choice Questions*

The questions in this section are actually claims that are either true of false. It is up to you to decide which of the two alternatives is correct. A (in)correct answers (costs) earns you a point.

0.1.1 (201703). *If the random variable $X \sim \text{Exp}(\lambda)$, then*

$$\mathsf{E}\left[X^2\right] = \frac{1}{\lambda^2}.$$

0.1.2 (201703). *If $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\mu)$ and X and Y are independent, then*

$$Z = \max\{X, Y\} \sim \text{Exp}(\lambda + \mu).$$

0.1.3 (201703). *If the interarrival times $\{X_k\}$ are i.i.d. and exponentially distributed with mean $1/\lambda$, and $A_k = \sum_{i=1}^k X_i$, then*

$$\mathsf{P}(A_{k+1} \leq t) = -\frac{(\lambda t)^k}{k!} e^{-\lambda t} + \mathsf{P}(A_k \leq t).$$

0.1.4 (201703). *Assume that $N_a(t) \sim P(\lambda t)$, $N_s(t) \sim P(\mu t)$ and independent. Then,*

$$\mathsf{P}(N_a(t) + N_s(t) = n) = e^{-(\mu+\lambda)t} \sum_{i=0}^n \frac{(\mu t)^{n-i}}{(n-i)!} \frac{(\lambda t)^i}{i!}.$$

0.1.5 (201703). *If $N(t) \sim P(\lambda t)$, then* $\mathsf{E}\left[N^2\right] = (\lambda t)^2$.

0.1.6 (201703). *For the $M/D/1$ queue, job service times are exponentially distributed and the interarrival times are deterministic.*

0.1.7 (201703). *A machine can switch on and off. If the queue length hits N, the machine switches on, and if the system becomes empty, the machine switches off. Let $I_k = 1$ if the machine is on in period k and $I_k = 0$ if it is off, let L_k be the number of items in the system at the end of period k, then,*

$$\begin{aligned} I_{k+1} &= \begin{cases} 1 & \text{if } L_k \geq N, \\ I_k & \text{if } 0 < L_k < N, \\ 0 & \text{if } L_k = 0, \end{cases} \\ I_{k+1} &= \mathbb{1}_{L_k \geq N} + I_k \, \mathbb{1}_{0<L_k<N}, \\ d_k &= \min\{L_{k-1}, c_k\}, \\ L_k &= L_{k-1} - (1 - I_k) d_k + a_k. \end{aligned}$$

Assume that $I_0 = 0$ at time $k = 0$.

It is true that the above recursions model this queueing system.

0.1.8 (201703). *For the $G/G/1$ queue the following recursion is true:*

$$\begin{aligned} W_k &= [W_{k-1} - X_k]^+ + S_k, \\ D_k &= A_k + W_k. \end{aligned}$$

0.1.9 (201703). *For the G/G/1 queue, suppose that the interarrival times $X_k \in \{1,3\}$ such that $\mathsf{P}(X_k = 1) = 1/5$ (hence, $\mathsf{P}(X_k = 3) = 4/5$) and the service times $S_k \in \{1,2\}$ with $\mathsf{P}(S_k = 1) = 1/3$ (hence, $\mathsf{P}(S_k = 2) = 2/3$). If $W_{L,0} = 3$,*

$$\mathsf{P}\left(W_{L,1} = 1\right) = \frac{4}{15}.$$

0.1.10 (201703). *Consider the random walk*

$$Z(t) = Z(0) + N_\lambda(t) - N_\mu(t),$$

where the arrival process is a Poisson process $N_\lambda(t)$ and the departure process is a Poisson process $N_\mu(t)$.

$$\mathsf{P}(m) Z(t) = n = \sum_{k=0}^{\infty} e^{-\mu t} \frac{(\mu t)^{k-n+m}}{(k-n+m)!} e^{-\lambda t} \frac{(\lambda t)^k}{k!},$$

where $\mathsf{P}(m)\cdot$ means that $Z(0) = m$.

0.1.11 (201704). *If $X \sim \text{Exp}(\lambda)$, $S \sim \text{Exp}(\mu)$ and independent, then*

$$\begin{aligned}
\mathsf{P}(X \le S) &= \mathsf{E}[\mathbb{1}_{X \le S}] \\
&= \int_0^\infty \int_0^\infty \mathbb{1}_{x \le y} f_{X,S}(x,y)\, dy\, dx \\
&= \lambda\mu \int_0^\infty \int_0^\infty \mathbb{1}_{x \le y} e^{-\lambda x} e^{-\mu y}\, dy\, dx \\
&= \lambda\mu \int_0^\infty e^{-\mu y} \int_0^y e^{-\lambda x}\, dx\, dy \\
&= \mu \int_0^\infty e^{-\mu y} e^{-\lambda y}\, dy \\
&= \mu \int_0^\infty e^{-(\lambda+\mu) y}\, dy \\
&= \frac{\mu}{\lambda + \mu}.
\end{aligned}$$

0.1.12 (201704). *For the G/G/1 it is true that*

$$W_k = [W_{k-1} - X_k + S_k]^+.$$

0.1.13 (201704). *For the G/G/1 queue the virtual waiting time process $\{V(t), t \ge 0\}$ satisfies*

$$V(t) = [V(A_{A(t)}) + (A_{A(t)} - t)]^+.$$

0.1.14 (201704). *For the G/G/1 queue, if $\mathsf{E}[B]$ is the expected busy time and $\mathsf{E}[I]$ is the expected idle time, then*

$$\mathsf{E}[B] = \frac{\rho}{1-\rho} \mathsf{E}[I].$$

0.1.15 (201704). *Consider a paint factory which contains a paint mixing machine that serves two classes of jobs, A and B. The processing times of jobs of types A and B are constant and require t_A and t_B hours. The job arrival rate is λ_A for type A and λ_B for type B jobs. It takes a setup time of S_{ij} hours to clean the mixing station when changing from paint type i to type j.*

The linear program below can be used to determine the minimal batch sizes. To keep the system (rate) stable,

$$\text{minimize } T$$

such that

$$\begin{aligned} T &= k_A t_A + S_{AB} + k_B t_B + S_{BA}, \\ \lambda_A T &< k_A, \\ \lambda_B T &< k_B. \end{aligned}$$

0.1.16 (201704)**.** *For the $G/G/1$ queue the stationary distribution of the waiting times at arrival times is equal to the empirical distribution*

$$\mathsf{P}(W \le x) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{W_k \le x}.$$

0.1.17 (201802)**.** *If $N(t)$ is Poisson distributed with parameter λt, i.e., $N(t) \sim P(\lambda t)$, the variance $\mathsf{V}[N(t)] = \lambda t$.*

0.1.18 (201802)**.**

$$\sum_{n=0}^{\infty} n \frac{\lambda^n}{n!} = \lambda e^{\lambda}.$$

0.1.19 (201802)**.** *Let N be a Poisson process with rate λ. Then,*

$$\{N(0,s] + N(s,t] = 1\} \cap \{N(0,t] = 1\} = \{N(0,s] = 1\}.$$

0.1.20 (201802)**.** *If $N \sim P(\lambda)$,*

$$M_N(s) = \mathsf{E}\left[e^{sN}\right] = \exp(\lambda(s-1)).$$

0.1.21 (201802)**.** *Suppose $N \sim P(\lambda)$ and N_1 is a random variable obtained by 'thinning' N with Bernoulli random variables with success probability p. The following reasoning is correct:*

$$\begin{aligned} \mathsf{P}(N_1 = k) &= \sum_{n=k}^{\infty} \mathsf{P}(N_1 = k, N = n) = \sum_{n=k}^{\infty} \mathsf{P}(N_1 = k \,|\, N = n)\mathsf{P}(N = n) \\ &= \sum_{n=k}^{\infty} \mathsf{P}(N_1 = k \,|\, N = n) e^{-\lambda} \frac{\lambda^n}{n!} \\ &= \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} e^{-\lambda} \frac{\lambda^n}{n!} \\ &= e^{-\lambda} \sum_{n=k}^{\infty} \frac{p^k(1-p)^{n-k}}{k!(n-k)!} \lambda^n = e^{-\lambda} \frac{(\lambda p)^k}{k!} \sum_{n=k}^{\infty} \frac{(\lambda(1-p))^{n-k}}{(n-k)!} \\ &= e^{-\lambda} \frac{(\lambda p)^k}{k!} \sum_{n=0}^{\infty} \frac{(\lambda(1-p))^n}{n!} = e^{-\lambda} \frac{(\lambda p)^k}{k!} e^{\lambda(1-p)} \\ &= e^{-\lambda p} \frac{(\lambda p)^k}{k!}. \end{aligned}$$

0.1.22 (201802)**.** *When many unrelated jobs arrive at a queueing system, like patients at a hospital, or customers at a shop, it is reasonable to model the interarrival times as exponentially distributed with a mean that is constant during short periods, (e.g. 10 minutes or 30 minutes, depending on the relevant context).*

0.1.23 (201802). *If X is exponentially distributed with rate λ,*

$$\mathsf{E}[X] = \int_0^\infty \lambda e^{-\lambda t}\, dt.$$

0.1.24 (201802). *Given a non-negative random variable B taking values in $\mathbb{N}$ and with $F(i) = \mathsf{P}(B \leq i)$, then, $\mathsf{E}[B] = \sum_{i=0}^\infty iF(i)$.*

0.1.25 (201802). *In the $D/M/1$ jobs have deterministic service times.*

0.1.26 (201802). *In a discrete-time queueing system, when job arrivals in period k cannot be served in period k, then $d_k = \min\{Q_{k-1}, c_k\}$, $Q_k = Q_{k-1} - d_k + a_k$.*

0.1.27 (201802). *For a continuous-time queueing system, if S_2 is the service time of job 2 and X_3 is the time between jobs 2 and 3, and S_2 and X_3 are independent, then*

$$\mathsf{P}(S_2 = 1, X_3 = 3) = \mathsf{P}(S_2 = 1)\mathsf{P}(X_3 = 3).$$

0.1.28 (201802). *For a single-server queueing system that starts empty, the number $L(t)$ of jobs in the system at time t satisfies*

$$L(t) = \sum_{k=1}^\infty \left[\mathbb{1}_{A_k \leq t} - \mathbb{1}_{D_k \leq t}\right].$$

0.1.29 (201802). *For three numbers a, b, c, $\max\{a, \max\{b, c\}\} = \max\{a, b, c\}$.*

0.1.30 (201802). *If a r.v. $X \sim \text{Exp}(\lambda)$, i.e., exponentially distributed with mean λ^{-1}, then the following shows that X has the memoryless property:*

$$\mathsf{P}(X > t + h | X > t) = \frac{\mathsf{P}(X > t + h, X > t)}{\mathsf{P}(X > t)} = \frac{\mathsf{P}(X > t + h)}{\mathsf{P}(X > t)} = \frac{e^{-\lambda(t+h)}}{e^{-\lambda t}} = e^{-\lambda h} = \mathsf{P}(X > h).$$

0.1.31 (201803). *Consider the $G/G/1$ queue. Under the 'shortest processing time first' scheduling rule predictions of the finish times (i.e., quoting due dates) are more accurate than under the 'first-in-first-out' rule.*

0.1.32 (201803). *Consider a single-server that serves two parallel queues A and B. Each queue receives a minimal service capacity every period. Reserved capacity unused for one queue can be used to serve the other queue. Any extra capacity beyond the reserved capacity is given to queue A with priority. The following set of recursions suffices to simulate this situation,*

$$\begin{aligned}
c_{k,A} &= \min\{Q_{k-1,A}, r_A\}, \\
d_{k,A} &= \min\{Q_{k-1,A}, c_k - c_{k,A}\}, \\
Q_{k,A} &= Q_{k-1,A} - d_{k,A} + a_{k,A}, \\
d_{k,B} &= \min\{Q_{k-1,B}, c_k - d_{k,A}\}, \\
Q_{k,B} &= Q_{k-1,B} - d_{k,B} + a_{k,B},
\end{aligned}$$

where r_A, r_B are the reserved capacities for each queue, and c_k the total service capacity available in time k and such that $c_k \geq r_A + r_B$.

0.1.33 (201803). *The Kolmogorov-Smirnov statistic between the distributions of the random variables $W_{Q,k}$ and $W_{Q,k-1}$ is given by*

$$\max_x\{\mathsf{P}\left(W_{Q,k} \leq x\right) - \mathsf{P}\left(W_{Q,k-1} \leq x\right)\}.$$

0.1.34 (201804). *Assume that the interarrival times $\{X_i\}$ are i.i.d. and $X_i \sim \text{Exp}(\lambda)$. Let $A_i = X_1 + X_2 + \cdots + X_i = \sum_{k=1}^{i} X_k$ with $i \geq 1$. Then,*

$$\mathsf{E}[A_i] = \frac{i}{\lambda}.$$

0.1.35 (201804). *We have a discrete-time queueing system with a server with capacity c per period. The server can serve 2 more jobs when the queue length is 24 or longer, and 1 less when the queue length is less than 12. We can use the following construction to simulate this queueing process:*

$$\begin{aligned} c_n &= c + 2\,\mathbb{1}_{Q_{n-1}\geq 24} - \mathbb{1}_{Q_{n-1}\leq 12}, \\ d_n &= \min\{Q_{n-1}, c_n\}, \\ Q_n &= Q_{n-1} + a_n - d_n, \end{aligned}$$

where the notation is the same as in the book.

0.1.36 (201804). *We have a discrete-time queueing system with a server with capacity c per period. The server can serve 2 more jobs when the queue length is 24 or longer, and 1 less when the queue length is less than 12. The following formula computes the number of periods in which the queue exceeds 30 for a simulation that starts at period 1 and stops at period N,*

$$N - \sum_{n=1}^{N} \mathbb{1}_{Q_n \leq 30}.$$

0.1.37 (201804). *The following is correct:*

$$(\lambda t)^k (\mu t)^{k+m-n} = (\lambda/\mu)^{(n-m)/2} (t\sqrt{\lambda\mu})^{2k+m-n}.$$

0.1.38 (201804). *Consider the G/G/1 queue. The total amount of service that arrived during the arrival times of the first and the $n+1$th job, i.e., between $[A_1, A_{n+1})$, is $\sum_{i=1}^{n} S_i$. Thus, the fraction of time that the server has been busy during $[A_1, A_{n+1})$ is*

$$\frac{\sum_{i=1}^{n} S_i}{A_{n+1} - A_1} = \frac{\sum_{i=1}^{n} S_i}{\sum_{i=1}^{n+1} X_i - X_1} = \frac{\sum_{i=1}^{n} S_i}{\sum_{i=2}^{n+1} X_i}.$$

0.1.39 (201807). *Let $\{N(t)\}$ denote the Poisson process with rate λ, and write $N(s,t]$ for the number of arrivals in the interval $(s,t]$. Claim: the following holds for $h \ll 1$,*

$$\begin{aligned} \mathsf{P}(N(t+h) = n | N(t) = n) &= \mathsf{P}(N(t+h) = n, N(t) = n)/\mathsf{P}(N(t) = n) \\ &= \mathsf{P}(N(t, t+h] = 0, N(t) = n)/\mathsf{P}(N(t) = n) \\ &= \mathsf{P}(N(t, t+h] = 0)\mathsf{P}(N(t) = n)/\mathsf{P}(N(t) = n), \\ &= \mathsf{P}(N(t, t+h] = 0) \\ &= \mathsf{P}(N(0, h] = 0) \\ &= e^{-\lambda h}(\lambda h)^0/0! \\ &= e^{-\lambda h} = 1 - \lambda h + o(h). \end{aligned}$$

0.1.40 (201807). *Assume that the arrival rate of customers is constant in a given half hour. Claim: it is reasonable to model the interarrival times of customers at call centers or hospitals as exponentially distributed.*

0.1.41 (201807). *If X is an exponentially distributed random variable with mean $1/\lambda$ and Y exp. dist. with mean $1/\mu$, and X and Y are independent. Claim:*

$$\mathsf{P}(X \leq Y) = 1 - \frac{\lambda}{\lambda+\mu}.$$

0.1.42 (201807). *Let S be a continuous random variable with survivor function G, density f, and finite second moment. Claim:*

$$\begin{aligned}\int_0^\infty yG(y)dy &= \int_0^\infty y \int_y^\infty f(x)\,dx dy = \int_0^\infty y \int_0^\infty 1\{y \leq x\} f(x)\,dx dy \\ &= \int_0^\infty f(x) \int_0^\infty y 1\{y \leq x\}\,dy dx = \int_0^\infty f(x) \frac{x^2}{2}\,dx = \frac{\mathsf{E}\left[S^2\right]}{2}.\end{aligned}$$

0.1.43 (201902). *Let $N \sim P(\lambda)$. Then its variance is λ.*

0.1.44 (201902). *Let $N \sim P(\lambda)$. Then its SCV is $1/\lambda^2$.*

0.1.45 (201902). *Let $f(x) = 10^6 x^2$. Then $f(x) = o(x)$.*

0.1.46 (201902). *Let N be a random variable taking values in $\mathbb{N}$. Let $\{Z_i\}$ be a set of i.i.d. Bernoulli random variables, and independent of N. Let $Y = \sum_{i=1}^N Z_i$. Then, for $s > 0$,*

$$\mathsf{E}\left[e^{sY}\right] = \sum_{n=0}^\infty \mathsf{E}\left[e^{s\sum_{i=1}^n Z_i}\right] \mathsf{E}\left[\mathbb{1}_{N \leq n}\right].$$

0.1.47 (201902). *Let Y be a random variable, exponentially distributed with rate $\lambda > 0$. Then $\mathsf{V}[Y] = 1/\lambda^2$.*

0.1.48 (201902). *Let A_i be the arrival time of customer i and set $A_0 = 0$. Assume that the inter-arrival times $\{X_i\}$ are i.i.d. with exponential distribution with mean $1/\lambda$ for some $\lambda > 0$. It it true that for the moment-generating function $M_{A_i}(t)$*

$$M_{A_i}(t) = \mathsf{E}\left[e^{tA_i}\right] = \mathsf{E}\left[\exp\left(t \sum_{k=1}^{i-1} X_k\right)\right].$$

0.1.49 (201902). *Let $S \sim U[0,7]$ and $X \sim U[0,10]$, where $U[I]$ stands for the uniform distribution concentrated on the interval I, and S and X independent. Then the joint density function is equal to*

$$f_{XS}(x,s) = \mathbb{1}_{0 \leq x \leq 10, 0 \leq s \leq 7}.$$

0.1.50 (201902). *Consider a check-in desk at an airport. There is one desk that is dedicated to business customers. However, when it is idle (i.e., no business customer in service or in queue), this desk also serves economy class customers. The other c desks are reserved for economy class customers. The queueing process as perceived by the economy class customers can be modeled as an $M/M/(c+1)$ queue.*

0.1.51 (201902). *In the $M/G/\infty$ queue jobs never spend time in queue.*

0.1.52 (201902). *The following Python code to simulate a queueing system in discrete time will work as is:*

```python
import numpy as np

Q = np.zeros_like(a)
d = np.zeros_like(a)
Q[0] = 10 # initial queue length

for k in range(1, len(a)):
 d[k] = min(Q[k - 1], c[k])
 Q[k] = Q[k - 1] - d[k] + a[k]
```

0.1.53 (201902)**.** *We have a queueing system in discrete time. Take $c_k = c\,\mathbb{1}_{Q_{k-1}>t}$ as service capacity in period k. If $c < t$, and $Q_0 > 0$, then $Q_k > 0$ for all k.*

0.1.54 (201902)**.** *With the recursion below we can simulate a queueing system in discrete time such that the arrivals in period k can also be served in period k. (Note: the question is not whether this code will run as is; it will not.)*

```python
for k in range(1, len(a)):
 Q[k] = max(Q[k - 1] - c[k] + a[k], 0)
 d[k] = Q[k - 1] + a[k] - Q[k]
```

0.1.55 (201902)**.** *Take $d_k = \min\{Q_{k-1}+a_k, c_k\}$, and assume that jobs are served in FIFO sequence. The largest possible waiting time $W_{+,k}$ for a job arriving in period k is given by*

$$W_{+,k} := \min\left\{ m : \sum_{i=k}^{k+m} c_i \geq Q_{k-1} + a_k \right\}.$$

0.1.56 (201902)**.** *Consider a random walk $Z_k = Z_{k-1} + X_k$ with $\{X_k\}$ a process of i.i.d. random variables such that $\mathsf{P}(X_k = 1) = 1 - \mathsf{P}(X_k = -1) = p \in (0,1)$. Then, for large n and $\alpha = 2$*

$$\mathsf{P}\left(Z_n > np + \alpha\sqrt{np(1-p)}\right) \approx 2.5\%.$$

0.1.57 (201903)**.** *Let $\{X_k, k = 1,2,\ldots\}$ be a set of i.i.d. exponentially distributed random variables with mean $1/\lambda$, and $\{N(t), t \geq 0\}$ with $N(0) = 0$ the associated counting process. Then $\mathsf{P}(X_1 \leq s, X_2 \leq t) = \mathsf{P}(N(s+t) = 2)$.*

0.1.58 (201903)**.** *Let $X \sim \mathrm{Exp}(\lambda)$ and $M_X(s)$ the associated moment-generating function. Then $M_X(s) < \infty$ for all $s \in \mathbb{R}$.*

0.1.59 (201903)**.** *We consider a discrete-time queueing system with a_k the number of arrivals in period k, and c_k the service capacity. Let*

$$d_k = \min\{Q_{k-1}, c_k\}, \qquad Q_k = Q_{k-1} - d_k + a_k. \tag{0.1.1}$$

Take a_k Poisson distributed with $\lambda = 2$, i.e., $a_k \sim P(2)$, and $c_k \sim P(1)$ and initialize $Q_0 = 0$. Then $\mathsf{P}(Q_{10000} \geq 100) \leq 1/2$.

0.1.60 (201903)**.** *A single-server queueing station processes customers. At the start of a period the server capacity is chosen, so that for period k the capacity is c_k. Demand that arrives in a period can be served in that period. It costs β per unit time per unit processing capacity to operate the machine, i.e., to have it switched on. There is also a cost h per unit time per job in the system. The total cost up to some time T is given by $\beta\sum_{k=1}^{T} c_k$.*

0.1.61 (201903)**.** *In python you need to set the seed of the random number generator to a fixed value in order to obtain the same random numbers for various simulation runs.*

0.1.62 (201903)**.** *Consider the* $G/G/1$ *queue in continuous time with inter-arrival times* X_k*, service times* S_k*, arrival times* A_k *and departure times* D_k*. We can compute the waiting times and sojourn times with the following recursion:*

$$W_{Q,k} = [W_{k-1} - X_k]^+, \qquad W_k = W_{Q,k} + S_k = [W_{k-1} - X_k]^+ + S_k. \tag{0.1.2}$$

0.1.63 (201904)**.** *One server serves two queues and has capacity* c *per period available. We divide the capacity* c *over the queues in proportion to the queue lengths* L^i_{k-1}, $i = 1,2$*. The following implements this rule:*

$$c^1_k = \left\lfloor \frac{L^1_{k-1}}{L^1_{k-1} + L^2_{k-1}} c + \frac{1}{2} \right\rfloor, \qquad c^2_k = c - c^1_k,$$

where c^i_k *be the capacity allocated to queue* i *in period* k *and we include the rounding to prevent the loss of capacity.*

0.1.64 (201904)**.** *When* $X \sim \text{Exp}(\lambda)$ *its SCV is larger than 1.*

0.1.65 (201904)**.** *For the* $G/G/1$ *queue the waiting time satisfies the recursion*

$$W_{Q,k} = \max\{W_{Q,k-1} + S_k - X_k, 0\}.$$

0.1.66 (201907)**.** $e^x = 1 + x^2 + o(x^2)$.

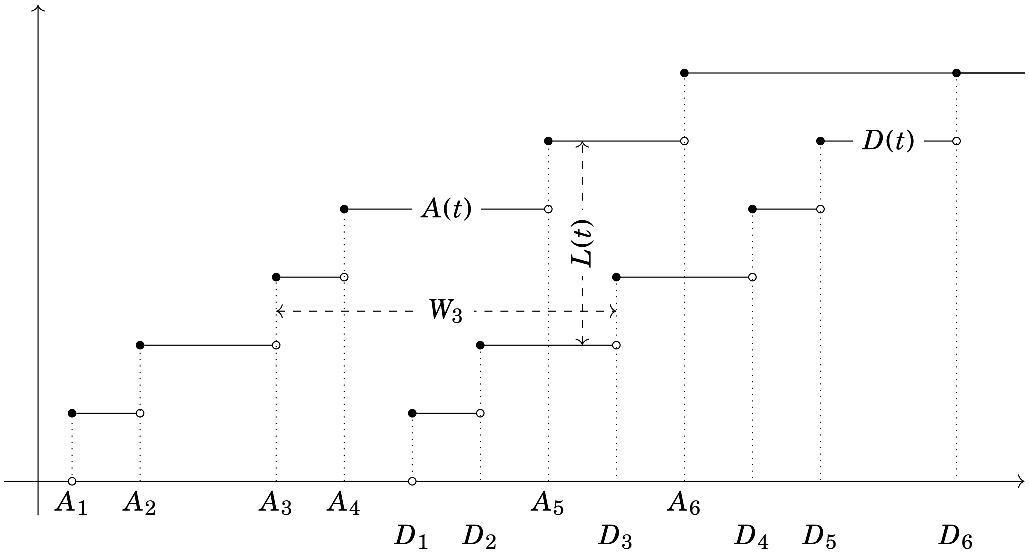
0.1.67 (201907)**.** *A machine produces items to serve customer demand. A fraction* p_k *of the items produced in period* k *turns out to be faulty, and has to be made anew. The following set of recursions models the queue of customers waiting to be served from the machine:*

$$\begin{aligned} d_k &= \min\{L_{k-1}, c_k\}, \\ L_k &= L_{k-1} - d_k(1 - p_k) + a_k. \end{aligned}$$

0.1.68 (201907)**.** *The departure process* $\{D(t)\}$ *can be computed from the set* $\{D_k\}$ *of departure times according to:*

$$D(t) = \sum_{k=1}^{\infty} \mathbb{1}_{D_k \leq t}.$$

0.1.69 (201907)**.** *The waiting time of the third job is correctly represented in the figure below.*



0.1.2 *Open Questions*

0.1.70 (201704). *One server serves n queues in parallel. The server has capacity $c_k \in \mathbb{R}_+$ for day k. The amount of work that arrives on day k for queue i is given by $a_k^i \in \mathbb{R}_+$. Jobs arriving on day k cannot be served on day k. Queue i receives service capacity in proportion to its queue length. Derive a set of recursions to compute the queue lengths $Q_k^i \in \mathbb{R}_+$ for days $k = 1,2,\ldots$.*

0.1.71 (201704). *Provide a real-world example for this queueing model.*

0.1.72 (201704). *For the situation of the previous question, find a recursion to compute a tight upper bound on the time the arriving work a_k^i spends in the system.*

0.1.73 (201807). *Consider a network of two stations in tandem. Jobs only arrive at station 1. Jobs arriving in period k are only accepted when the total number of jobs in the system is less than K at the start of the period. Otherwise, all arriving jobs are rejected. Provide a set of recursions to simulate the queue length process at both stations in discrete time.*

0.1.74 (201807). *We have a queueing system that operates under the following set of recursions:*

$$\begin{aligned} c_n &= 3\,\mathbb{1}_{Q_n \in [5,10]} + \mathbb{1}_{Q_n \notin [5,10]}, \\ d_n &= \min\{Q_{n-1} + a_n, c_n\} \\ Q_n &= Q_{n-1} + a_n - d_n, \end{aligned}$$

where the arrivals a_n are i.i.d. and have Poisson distribution with parameter $\lambda = 2$, c_n is the service capacity in period n and d_n corresponds to the number of departures. Why is this queueing system unstable?

0.1.75 (201807). *What is a disadvantage of using simulation to analyze queueing systems?*

0.1.76 (201904). *A machine produces items, but a fraction p of the items does not meet the quality requirements after the first pass at the server, but it requires a second pass. Assume that the repair of a faulty item requires half of the work of a new job, and that the faulty jobs are processed with priority over the new jobs. Also assume that faulty items do not need more than one repair (hence, faulty items that are repaired cannot be faulty anymore). Make a set of discrete-time recursions to analyze this case.*

0.1.77 (201907). *Why did we discuss the transient behavior of the random walk in a course on queueing theory?*

Solutions

s.0.1.1. Answer = B.

$$\mathsf{E}\left[X^2\right] = \frac{2}{\lambda^2}.$$

s.0.1.2. Answer = B.
$Z = \min\{X,Y\} \sim \text{Exp}(\lambda + \mu)$.

s.0.1.3. Answer = A.

I decided to withdraw this question. The initial wording of the question was like this: 'If the interarrival times $\{A_k\}$ are i.i.d. and exponentially distributed with mean $1/\lambda$, then...'. However, in the book we use the notation A_k for arrival times, not interarrival times. During the exam I changed the word 'interarrival' by 'arrival', hoping that it would clarify the meaning of the A_k, but then there is another problem, namely the arrival times are not i.i.d. Hence, there was no way in which the problem could be interpreted during the exam in an unambiguous way.

s.0.1.4. Answer = A.

s.0.1.5. Answer = B.

Set $t = 1$ for notational simplicity. Then

$$M(s) = \mathsf{E}\left[e^{sN}\right] = e^{-\lambda}\sum_{n=0}^{\infty}\frac{\lambda^n}{n!} = \exp\{\lambda(e^s - 1)\}.$$

Then,

$$M'(s) = \frac{\mathrm{d}}{\mathrm{d}t}M(s) = \exp\{\lambda(e^s - 1)\}\lambda e^s,$$

and

$$M''(s) = \exp\{\lambda(e^s - 1)\}(\lambda^2 e^{2s} + \lambda e^s).$$

Finally, $\mathsf{E}\left[N^2\right] = M''(0) = \lambda^2 + \lambda$.

s.0.1.6. Answer = B.

The interarrival times are exponentially distributed and the service times are deterministic.

s.0.1.7. Answer = B. It should be this: $d_k = I_k \min\{L_{k-1}, c_k\}$, $L_k = L_{k-1} - d_k + a_k$.

s.0.1.8. Answer = A.

s.0.1.9. Answer = A.

$$\mathsf{P}\left(W_{L,1} = 1\right) = \mathsf{P}(S_0 = 1, X_1 = 3) = \frac{1}{3}\frac{4}{5} = \frac{4}{15}.$$

I asked $< 1/2$ in case you would use your calculator which might have resulted in rounding errors. Like this, I was on the safe side.

s.0.1.10. Answer = A.

s.0.1.11. Answer = B.

$$
\begin{aligned}
\mathsf{P}(X \leq S) &= \mathsf{E}[\mathbb{1}_{X\leq S}] \\
&= \int_0^\infty \int_0^\infty \mathbb{1}_{x\leq y} f_{X,S}(x,y)\,\mathrm{d}y\,\mathrm{d}x \\
&= \lambda\mu \int_0^\infty \int_0^\infty \mathbb{1}_{x\leq y} e^{-\lambda x} e^{-\mu y}\,\mathrm{d}y\,\mathrm{d}x \\
&= \lambda\mu \int_0^\infty e^{-\mu y} \int_0^y e^{-\lambda x}\,\mathrm{d}x\,\mathrm{d}y \\
&= \mu \int_0^\infty e^{-\mu y}(1-e^{-\lambda y})\,\mathrm{d}y \\
&= \mu \int_0^\infty (e^{-\mu y} - e^{-(\lambda+\mu)y})\,\mathrm{d}y \\
&= \mu \int_0^\infty (e^{-\mu y} - e^{-(\lambda+\mu)y})\,\mathrm{d}y \\
&= 1 - \frac{\mu}{\lambda+\mu}.`
\end{aligned}
$$

s.0.1.12. Answer = B.

$$
\begin{aligned}
W_{Q,k} &= [W_{k-1} - X_k]^+, \\
W_k &= W_{Q,k} + S_k = [W_{k-1} - X_k]^+ + S_k.
\end{aligned}
\tag{0.1.3}
$$

s.0.1.13. Answer = A.

s.0.1.14. Answer = A.

s.0.1.15. Answer = A.

s.0.1.16. Answer = B.

$$
\mathsf{P}(W \leq x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n \mathbb{1}_{W_k \leq x}. \tag{0.1.4}
$$

s.0.1.17. Answer = A.

s.0.1.18. Answer = A.

s.0.1.19. Answer = B.

s.0.1.20. Answer = B.

s.0.1.21. Answer = A.

s.0.1.22. Answer = A.

s.0.1.23. Answer = B.

s.0.1.24. Answer = B.

s.0.1.25. Answer = B.

s.0.1.26. Answer = A.

s.0.1.27. Answer = A.

s.0.1.28. Answer = A.

s.0.1.29. Answer = A.

s.0.1.30. Answer = A.

s.0.1.31. Answer = B.

s.0.1.32. Answer = B. The following is correct:

$$\begin{aligned} c_{k,B} &= \min\{Q_{k-1,B}, r_B\}, \\ d_{k,A} &= \min\{Q_{k-1,A}, c_k - c_{k,B}\}, \\ Q_{k,A} &= Q_{k-1,A} - d_{k,A} + a_{k,A}, \\ d_{k,B} &= \min\{Q_{k-1,B}, c_k - d_{k,A}\}, \\ Q_{k,B} &= Q_{k-1,B} - d_{k,B} + a_{k,B}. \end{aligned}$$

s.0.1.33. Answer = B. It should be

$$\max_x\{|\mathsf{P}\left(W_{Q,k} \leq x\right) - \mathsf{P}\left(W_{Q,k-1} \leq x\right)|\}.$$

s.0.1.34. Answer = A.

s.0.1.35. Answer = B. It should be this:

$$\begin{aligned} c_n &= c + 2\,\mathbb{1}_{Q_{n-1}\geq 24} - \mathbb{1}_{Q_{n-1}<12}, \\ d_n &= \min\{Q_{n-1}, c_n\}, \\ Q_n &= Q_{n-1} + a_n - d_n. \end{aligned}$$

Note the *less than*.

s.0.1.36. Answer = A.

$$N - \sum_{i=n}^{N} \mathbb{1}_{Q_n \leq 30} = \sum_{i=n}^{N} (1 - \mathbb{1}_{Q_n \leq 30}) = \sum_{i=n}^{N} \mathbb{1}_{Q_n > 30}.$$

s.0.1.37. Answer = A.

s.0.1.38. Answer = B. The point here is that it is not true in general. Only when $L(A_{n+1}) = 0$, it is true. Otherwise, if $A_{n+1} < D_n$, there is a queue, hence the total amount of work arrived is larger than what has been served. In fact, the quantity in the formula may exceed 1, in which case it cannot be a fraction of time.

s.0.1.39. Answer = A.

s.0.1.40. Answer = A.

s.0.1.41. Answer = B.

s.0.1.42. Answer = A.

s.0.1.43. Answer = A.

s.0.1.44. Answer = B. The SCV is $1/\lambda$.

s.0.1.45. Answer = A.

s.0.1.46. Answer = B.

$$\mathsf{E}\left[e^{sY}\right] = \sum_{n=0}^{\infty} \mathsf{E}\left[e^{s\sum_{i=1}^{n} Z_i}\right] \mathsf{E}\left[\mathbb{1}_{N=n}\right].$$

We discussed this and similar expressions in class, at least in two lectures.

s.0.1.47. Answer = A.

s.0.1.48. Answer = B. $A_i = A_{i-1} + X_i$. Hence,

$$M_{A_i}(t) = \mathsf{E}\left[e^{tA_i}\right] = \mathsf{E}\left[\exp\left(t\sum_{k=1}^{i} X_k\right)\right].$$

s.0.1.49. Answer = B.

$$f_{XS}(x,s) = f_X(x)\cdot f_S(s) = \frac{1}{10}\mathbb{1}_{0\leq x\leq 10}\cdot\frac{1}{7}\mathbb{1}_{0\leq s\leq 7}.$$

s.0.1.50. Answer = B.

s.0.1.51. Answer = A.

s.0.1.52. Answer = B. The vectors a and c are not given. See this:

```
>>> import numpy as np
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
ImportError: No module named numpy

>>> Q = np.zeros_like(a)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'np' is not defined
>>> d = np.zeros_like(a)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'np' is not defined
>>> Q[0] = 10 # initial queue length
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'Q' is not defined
```

```
>>> for k in range(1, len(a)):
...   d[k] = min(Q[k - 1], c[k])
...   Q[k] = Q[k - 1] - d[k] + a[k]
...
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'a' is not defined
```

s.0.1.53. Answer = A.

s.0.1.54. Answer = A. This recursion is used to relate the queueing process to a random walk.

s.0.1.55. Answer = A.

The initial formulation of the question was like this: Take $d_k = \min\{Q_{k-1} + a_k, c_k\}$, and assume that jobs are served in FIFO sequence. The largest possible waiting time $W_{+,k}$ for the kth arriving job is given by

$$W_{+,k} := \min\left\{ m : \sum_{i=k}^{k+m} c_i \geq Q_{k-1} + a_k \right\}.$$

We removed this question from the exam as it was confusing. The problem was that the kth job need not arrive in period k.

The formulation in the question is adapted so that it can be used as an example.

s.0.1.56. Answer = B. Immediate consequence of the central limit theorem. It is in fact very interesting. It tells us that when $\mathsf{E}[X_k] < 0$ and Q_0 is very large, the relative variability of the queueing process decreases when n becomes large. To see this, observe that $\mathsf{E}[X_k] = p - (1-p) = 2p-1$ and $\mathsf{V}[X_k] = 1 - (2p-1)^2 = 2p(1-2p)$. Then consider $(Z_n - n\,\mathsf{E}[X_k])/\sqrt{n\,\mathsf{V}[X_k]}$. From the central limit theorem we know that this random variable is normally distributed with mean 0 and $\sigma = 1$.

The statement in the question uses random variables X_k such that $\mathsf{P}(X_k = 1) = p = 1 - \mathsf{P}(X_k = 0)$.

s.0.1.57. Answer = B.

When $X_1 \leq s$ and $X_2 \leq t$ then still $N(s+t) = 3$ is possible. For $N(s+t) = 2$ it is necessary that $X_1 + X_2 + X_3 > s + t$.

s.0.1.58. Answer = B.

If $s > \lambda$ the integral does not converge.

s.0.1.59. Answer = B.

s.0.1.60. Answer = B.

The total cost is

$$\sum_{k=1}^{T} \left(\beta c_k + h Q_k\right).$$

s.0.1.61. Answer = A.

s.0.1.62. Answer = A.

s.0.1.63. Answer = A, **1.2.11**. We introduce rounding to prevent the service of 'partial' customers. For instance, if the first queue contains 1 job, and the second 2, then without rounding we would server 2/3 customer of the second type.

Note that the case with empty queues does not lead to a problem. When there are no jobs, the service distribution is irrelevant.

s.0.1.64. Answer = B, see **1.3.8**.

s.0.1.65. Answer = B, (1.4.5).

s.0.1.66. Answer = B, see **0.2.2**.

s.0.1.67. Answer = A, see**1.2.6**.

s.0.1.68. Answer = A.

s.0.1.69. Answer = A.

s.0.1.70. Let c_k^i be the capacity allocated to queue i in period k. The fair rule gives that

$$c_k^i = \frac{Q_{k-1}^i}{\sum_{i=1}^n Q_{k-1}^i} c_k.$$

Then,

$$\begin{aligned} d_k^i &= \min\{Q_{k-1}^i, c_k^i\}, \\ Q_k^i &= Q_{k-1}^1 + a_k^i - d_k^i. \end{aligned}$$

s.0.1.71. A machine serving work of different types of jobs. Jobs of the same type are put in the same queue. When the machine has to switch from one type of job to another, it might need to change a tool. For this reason it works on jobs of the same queue for some time. Then it changes to the next queue, and so on.

s.0.1.72. We need to compute the time it takes to clear Q_k^i. This is

$$\min\left\{r : \sum_{j=k+1}^r c_j^i \geq Q_k^i\right\}.$$

Some wrong answers:

- Computing $\mathsf{E}[W]$. The question is not to compute the expectation of waiting time...
- $W_k^i = (Q_{k-1}^i + a_k^i)/C_k$. This is partly ok, intuitively at least. However, the capacity c_k is the total capacity for all queues, hence the division should be by c_k^i. This is still not completely ok, because c_k^i is not constant as a function of k.
- $W_k = [W_{k-1} - X_k]^+ + S_k$. Here we are not dealing with a continuous-time queueing model.

s.0.1.73.

$$\begin{aligned}
a'_{k,1} &= a_{k,1}\,\mathbb{1}_{Q_{k-1,1}+Q_{k-1,2}<K},\\
d_{k,1} &= \min\{c_{k,1}, Q_{k-1,1}\},\\
Q_{k,1} &= Q_{k-1,1} + a'_{k,1} - d_{k,1},\\
d_{k,2} &= \min\{c_{k,2}, Q_{k-1,2}\},\\
Q_{k,2} &= Q_{k-1,2} + d_{k,1} - d_{k,2}.
\end{aligned}$$

Some students just model one queue, or do not mention that the arrivals at the second station are the departures of the first station.

s.0.1.74. This system is in fact unstable. Once $Q_n > 10$, there is a drift upwards with rate $2-1$. Eventually, the queue will drift to infinity.

Now even when $Qn > 10$, it may happen that some time later the queue length gets below 10. This happens when $a_i = 0$ for a sufficient number of consecutive periods. However, whenever the queue is less than 10, the state 10 will be hit with probability one. Eventually there will be time that it does not get back to state 10 again.

The process is a random walk with drift. The study of its probabilistic properties is interesting. Chapter 3 of Feller 1 is an interesting read.

s.0.1.75. It can take a long simulation time to obtain relevant answers. It is hard to get structural insights into systems. When you make a bug, it may be hard to find it out.

Some students say that the fact that a simulation is a model is itself a disadvantage. This is of course not a good answer. Anything we say is a model: the word 'apple' is not an apple itself, and all apples are slightly different...

Others say that with simulation it is only possible to model systems that operate in discrete time. This is nonsense, just check the book.

s.0.1.76. See 1.2.7.

s.0.1.77. The random walk acts as a queueing system without a boundary at $y = 0$; the random walk can be negative while the queueing system cannot be negative. Often a queueing system is called a reflected random walk. We analyzed the transient behavior of the random walk, and saw that that was already quite complicated. Including reflections makes the analysis of the transient behavior harder. Thus we decided to focus on the stationary behavior instead.

0.2 OLD EXAM QUESTIONS

0.2.1 *Multiple-choice Questions*

0.2.1 (201703)**.** *Let $L(s)$ be the number of items in the system at time s. Define*

$$\alpha = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} L(A_k).$$

For the $M/G/1$ queue we can use PASTA to see that $1-\rho = \alpha$.

0.2.2 (201703). *For the G/G/1 queue,*

$$\frac{D(n-1,t)}{t} = \frac{D(n-1,t)}{Y(n,t)}\frac{Y(n,t)}{t} \to \mu(n)p(n), \tag{0.2.1}$$

if $n \geq 1$.

0.2.3 (201703). *For the M/M/1 queue with* $\lambda = 3$, $\mu = 5$, $\mathsf{E}\left[L_Q\right] \leq 1$

0.2.4 (201703).

$$\begin{aligned}
\sum_{n=0}^{\infty} n^2\rho^n &= \sum_{n=0}^{\infty}\left(\sum_{i=1}^{\infty} 2i\,\mathbb{1}_{i\leq n} - n\right)\rho^n = \sum_{n=0}^{\infty}\sum_{i=0}^{\infty} 2i\,\mathbb{1}_{i\leq n}\rho^n - \sum_{n=0}^{\infty} n\rho^n \\
&= \sum_{i=0}^{\infty} 2i \sum_{n=i}^{\infty}\rho^n - \frac{\mathsf{E}[L]}{1-\rho} = \sum_{i=0}^{\infty} 2i\rho^i \sum_{n=0}^{\infty}\rho^n - \frac{\mathsf{E}[L]}{1-\rho} \\
&= \frac{2}{1-\rho}\sum_{i=0}^{\infty} i\rho^i - \frac{\mathsf{E}[L]}{1-\rho} = \frac{2}{(1-\rho)^2}\mathsf{E}[L] - \frac{\mathsf{E}[L]}{1-\rho} \\
&= \frac{\mathsf{E}[L]}{1-\rho}\left(\frac{2}{1-\rho} - 1\right) = \frac{\mathsf{E}[L]}{1-\rho}\frac{1+\rho}{1-\rho} \\
&= \frac{\rho}{1-\rho}\frac{1+\rho}{(1-\rho)^2}.
\end{aligned}$$

0.2.5 (201703). *For the M/G/1 queue with* $G = U[0,A]$*, i.e., the uniform distribution on* $[0,A]$*:*

$$C_s^2 = \frac{1}{3}.$$

0.2.6 (201704). *Consider the following queueing process. At times* $0,2,4,\ldots$ *customers arrive, each customer requires* 1 *unit of service, and there is one server. Then, for* $t \in [0,3)$

$$Y(1,t) = \int_0^t \mathbb{1}_{L(s)=1}\,ds = \begin{cases} t & t \in [0,1), \\ 1 & t \in [1,2), \\ 1+(t-2) & t \in [2,3), \end{cases}$$

0.2.7 (201704). *For the M/M/c queue with* $\rho = c\lambda/\mu$,

$$p(n+1) = \frac{\Pi_{k=1}^{n+1}\min\{c,k\}}{\rho}^{n+1} p(0).$$

0.2.8 (201704). *Consider a queueing system in which each customer requires precisely 59 minutes of service. At the start of each hour, one customer arrives. Then* $\pi(0) = 1/60$.

0.2.9 (201704). *For the G/G/1 queue, when* T *is a moment in time in which the system is empty, we have*

$$\begin{aligned}
\int_0^T L(s)\,ds &= \int_0^T \sum_{k=1}^{A(T)} 1\{A_k \leq s < D_k\}\,ds \\
&= \sum_{k=1}^{A(T)} \int_0^T 1\{A_k \leq s < D_k\}\,ds = \sum_{k=1}^{A(T)} W_k.
\end{aligned}$$

In words, the area between the graphs of $A(s)$ *and* $D(s)$ *must be equal to the total waiting time spent by all jobs in the system until* T.

0.2.10 (201704). *For the number of jobs in the system L for the $M/G/1$ queue we have that*

$$\phi(z) = \mathsf{E}\left[z^L\right] = \sum_{n=0}^{\infty} z^n p(n) = (1-\rho)\sum_{n=0}^{\infty}(\rho z)^n = \frac{1-\rho}{1-\rho z}.$$

Then

$$\mathsf{E}[L] = \frac{d}{dz}\phi(z)\bigg|_{z=0}.$$

0.2.11 (201704). *For the $M^X/M/1$ queue,*

$$\frac{\lambda}{\mu}\mathsf{E}\left[B^2\right] = \rho\frac{\mathsf{V}[B]}{(\mathsf{E}[B])^2}\mathsf{E}[B] = \rho C_s^2\,\mathsf{E}[B].$$

0.2.12 (201704). *For the $M^X/M/1$ queue with $G(n) = \mathsf{P}(B > n)$,*

$$\begin{aligned}\sum_{n=0}^{\infty} G(n) &= \sum_{n=0}^{\infty}\mathsf{P}(B>n) = \sum_{n=0}^{\infty}\sum_{i=n+1}^{\infty}\mathsf{P}(B=i)\\ &= \sum_{n=0}^{\infty}\sum_{i=0}^{\infty} 1\{n<i\}\mathsf{P}(B=i) = \sum_{i=0}^{\infty}\sum_{n=0}^{\infty} 1\{n<i\}\mathsf{P}(B=i)\\ &= \sum_{i=0}^{\infty} i\mathsf{P}(B=i) = \mathsf{E}[B].\end{aligned}$$

0.2.13 (201704). *We model a workstation with just one machine as a $G/G/1$ queue. The coefficient of variation of the interarrival times of the jobs is 1 and the arrival rate $\lambda = 3/8$ per hour. The average service time $\mathsf{E}[S] = 2$ hours, $C_s^2 = 1/2$. Then, $\mathsf{E}\left[W_Q\right] \in [4,5]$ hours.*

0.2.14 (201802). *Let A_1 be the arrival time of the first job at a queueing system. Assume $L(A_1-) = 0$. Suppose that the $n+1$th job is the first job after A_1 that sees an empty system. Thus, $L(D_n) = 0$. The fraction of time that the server has been busy during $[A_1, A_{n+1})$ is*

$$\frac{\sum_{i=1}^n S_i}{A_{n+1} - A_1}$$

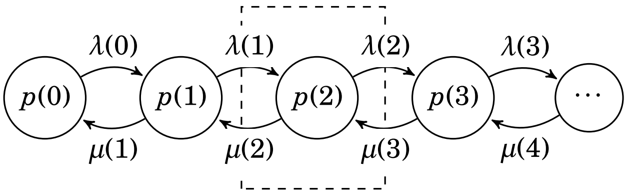
0.2.15 (201802). *The number of arrivals $A(t)$ up to time t is equal to* $\sup\{k : A_k \leq t\}$.

0.2.16 (201802). *In a discrete-time queueing model, L_k is the number of jobs in the system at the end of period k. Thus, $\sum_{k=1}^n \mathbb{1}_{L_k>m}$ is the number of jobs that see more than m jobs in the system upon arrival.*

0.2.17 (201802). *In a level-crossing analysis of a queueing system, the departure rate from state n is*

$$\mu(n) = \lim_{t\to\infty}\frac{D(n-1,t)}{Y(n,t)},$$

0.2.18 (201802). *To obtain the* balance equations *we do not count the number of up- and down crossings of a level. Instead we count how often a box around a state, such as state 2 in the figure below, is crossed from inside and outside.*



0.2.19 (201802)**.** *The process $L(t)$ that counts the number of jobs in system is* right continuous.

0.2.20 (201803)**.** *If the limit exists,*

$$\frac{1}{t}\sum_{k=1}^{A(t)} \mathbb{1}_{W_k \leq x} \to \mathsf{P}(W \leq w),$$

as $t \to \infty$.

0.2.21 (201803)**.** *Using the definitions of the book, for the $M/M/1$ queue and $t > 0$,*

$$\left|\frac{A(n,t)}{Y(n,t)}\frac{Y(n,t)}{t} - \frac{D(n,t)}{Y(n+1,t)}\frac{Y(n+1,t)}{t}\right| \leq 1.$$

0.2.22 (201803)**.** *For the $M/M/1$ queue, the following reasoning leads to the expected number of jobs in the system.*

$$\begin{aligned} M_L(s) = \mathsf{E}\left[e^{sL}\right] &= \sum_{n=0}^{\infty} e^{sn}p(n) = (1-\rho)\sum_n e^{sn}\rho^n \\ &= \frac{1-\rho}{1-e^s\rho}, \end{aligned}$$

where we assume that s is such that $e^s\rho < 1$. Then,

$$M_L'(s) = (1-\rho)\frac{1}{(1-e^s\rho)^2}e^s\rho.$$

Hence, $\mathsf{E}[L] = M_L'(0) = \rho/(1-\rho)$.

0.2.23 (201803)**.** *Customers of fast-food restaurants prefer to be served from stock. For this reason such restaurants often use a 'produce-up-to' policy: When the on-hand inventory I is equal or lower than some threshold $S-1$, the company produces items until the inventory level equals S again. The level S is known as the order-up-to level, and $S-1$ as the reorder level.*

Suppose that customers arrive as a Poisson process with rate λ and the production times of single items are i.i.d. and exponentially distributed with parameter μ. Assume also that customers who cannot be served from on-hand stock are backlogged, that is, they wait until their item has been produced.

The average on-hand inventory level is S minus the average number of jobs at the cook. i.e., $\mathsf{E}[I] = \sum_{i=0}^{S}(S-i)p(i)$.

0.2.24 (201803)**.** *A queueing system with balking, at level b say, behaves the same as a queueing system with finite calling population of size b.*

0.2.25 (201803)**.** *Consider the $M/G/1$ queue. By the PASTA property, a fraction ρ, $\rho < 1$, of the arrivals sees the server occupied, while a fraction $1-\rho$ sees a free server.*

0.2.26 (201803)**.** *For the $M/G/1$ queue we can use the PASTA property to see that the expected waiting time in the system is equal to* $\mathsf{E}[W] = \sum_{n=0}^{\infty}\mathsf{E}\left[W_Q \,\middle|\, N=n\right]\pi(n) + \mathsf{E}[S]$.

0.2.27 (201803). *The following computation is correct for the $M/M/1$ queue:*

$$\begin{aligned}\sum_{n=0}^{\infty} n^2\rho^n &= \sum_{n=0}^{\infty}\left(\sum_{i=1}^{\infty} 2i\,\mathbb{1}_{i\leq n} - n\right)\rho^n = \sum_{n=0}^{\infty}\sum_{i=0}^{\infty} 2i\,\mathbb{1}_{i\leq n}\rho^n - \sum_{n=0}^{\infty} n\rho^n \\ &= \sum_{i=0}^{\infty} 2i\sum_{n=i}^{\infty}\rho^n - \frac{\mathsf{E}[L]}{1-\rho} = \sum_{i=0}^{\infty} 2i\rho^i\sum_{n=0}^{\infty}\rho^n - \frac{\mathsf{E}[L]}{1-\rho} \\ &= \frac{2}{1-\rho}\sum_{i=0}^{\infty} i\rho^i - \frac{\mathsf{E}[L]}{1-\rho} = \frac{2}{(1-\rho)^2}\mathsf{E}[L] - \frac{\mathsf{E}[L]}{1-\rho} \\ &= \frac{\mathsf{E}[L]}{1-\rho}\left(\frac{2}{1-\rho} - 1\right) = \frac{\mathsf{E}[L]}{1-\rho}\frac{1+\rho}{1-\rho} \\ &= \frac{\rho}{1-\rho}\frac{1+\rho}{(1-\rho)^2}.\end{aligned}$$

0.2.28 (201803). *For the $M/M/1$ queue, if* $\mathsf{E}[L] = \rho/(1-\rho)$ *then* $\mathsf{E}[W] = \lambda\,\mathsf{E}[L]$.

0.2.29 (201803). *For the $M/M/1$ queue,* $\mathsf{P}(L \leq n) = \sum_{k=0}^{n}\rho^k$.

0.2.30 (201803). *For the $M/M/c$ queue,* $\mathsf{E}[L_Q] = \sum_{n=0}^{\infty}\max\{n-c,0\}p(n)$.

0.2.31 (201803). *The load of the $M^X/M/1$ queue is $\rho = \lambda\,\mathsf{E}[S]$ where $\mathsf{E}[S]$ is the service time of a single item in a batch.*

0.2.32 (201803). *For the $M/G/1$ queue the following is true:*

$$\lambda\,\mathsf{E}\left[S^2\right] = (1+C_s^2)\mathsf{E}[S], \quad \textit{where } C_s^2 = \frac{\mathsf{V}[S]}{(\mathsf{E}[S])^2} \tag{0.2.2}$$

is the square coefficient of variation.

0.2.33 (201803). *If a job arrives at time A, and $\{L(t)\}$ is the queue length process, then the random variable $L(A)$ denotes the number in the system seen by this job upon arrival.*

0.2.34 (201803). *For the $M^X/M/1$ queue, define*

$$A(m,n,t) = \sum_{k=1}^{A(t)} \mathbb{1}_{L(A_k-)=m}\,\mathbb{1}_{B_k>n-m}$$

as the number of jobs up to time t that see m in the system upon arrival and have batch size larger than $n-m$. Then, $A(n-1,n,t)$ is the number of batches that arrived up to time t.

0.2.35 (201803). *Consider the $M^X/M/1$ queue with partial acceptance: the system can contain at most K jobs, so that when a batch arrives, accept whatever fits in the queue, and reject the rest. The level-crossing equations are then as follows:*

$$\mu\pi(n+1) = \lambda\sum_{m=0}^{n}\pi(m)G(n-m),$$

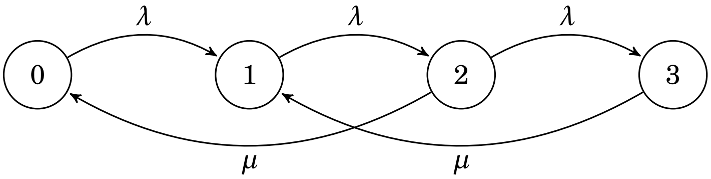
for $n = 0,1,\dots,K-1$, where $G(n-m) = \mathsf{P}(B > n-m)$ is the survivor function of the random batch size B.

0.2.36 (201804). *For the $G/G/1$ queue, the average number of jobs in the system as seen by arrivals is given by*

$$\frac{1}{t}\int_0^t L(s)ds = \frac{1}{t}\int_0^t (A(s)-D(s))ds, \tag{0.2.3}$$

where we use that $L(s) = A(s) - D(s) + L(0)$ is the total number of jobs in the system at time s and $L(0) = 0$.

0.2.37 (201804). *Consider the $M^2/M^2/1/3$ queue. The graph below shows all relevant transitions.*



0.2.38 (201804). *For the $M/M/1$ queue, when the server is busy at time 0, the time to the next departure has density $f_D(t) = \mu e^{-\mu t}$.*

0.2.39 (201804). *When $X \sim \text{Exp}(\lambda)$ and $S \sim \text{Exp}(\mu)$, and X and S are independent, their joint density is $f_{X,S}(x,y) = \lambda\mu e^{-\lambda x - \mu y}$. With this,*

$$\begin{aligned}
\mathsf{P}(X+S \le t) &= \lambda\mu \int_0^\infty \int_0^\infty e^{-\lambda x - \mu y} \mathbb{1}_{x+y\le t}\, dx\, dy \\
&= \lambda\mu \int_0^t e^{-\lambda x} \int_0^{t-x} e^{-\mu y}\, dy\, dx \\
&= \lambda \int_0^t e^{-\lambda x}(1 - e^{-\mu(t-x)}) dx \\
&= \lambda \int_0^t e^{-\lambda x}\, dx - \lambda e^{-\mu t} \int_0^t e^{(\mu-\lambda)x}\, dx
\end{aligned}$$

0.2.40 (201804). *For the $G/G/1$ queue and the definitions of the book, consider state n, i.e., the system contains n jobs. If we count the transitions into and out of state n, the following is true. The number of transitions into state n during $[0,t]$ are given by $A(n,t) + D(n-1,t)$, and the number of transitions out of state n up during $[0,t]$ is given $A(n-1,t) + D(n,t)$.*

0.2.41 (201804). *For the $M^X/M/1$ queue, if B_r is the number of items of the batch currently at the server and $L_{Q,b}$ the number of batches in queue, then*

$$\mathsf{E}[L] = \mathsf{E}\left[L_{Q,b}\right] \mathsf{E}[B] + \mathsf{E}[B_r].$$

0.2.42 (201804). *For the $M^X/M/1$ queue, Let $\tilde{A}_k$ be the moment the kth batch moves to the server and D_k its departure time. When $S_{k,i}$ is the service time of the ith item of batch k,*

$$\int_{\tilde{A}_k}^{D_k} \mathbb{1}_{L_S(s)=i}\, ds = S_{k,i}\, \mathbb{1}_{B_k \ge i}.$$

0.2.43 (201804). *For the $M/G/1$ queue and $S \sim \text{Exp}(\mu)$,*

$$\begin{aligned}
\mathsf{E}\left[S^2\right] = \mu \int_0^\infty x^2 e^{-\mu x}\, dx &= x^2 e^{-\mu x}\big|_0^\infty + 2\int_0^\infty x e^{-\mu x}\, dx \\
&= 2\frac{x}{\mu} e^{-\mu x}\Big|_0^\infty + \frac{2}{\mu}\int_0^\infty e^{-\mu x}\, dx = \frac{2}{\mu^2}.
\end{aligned}$$

0.2.44 (201804). *For the $M/G/1$ queue, define for $m = 1, \dots, n$,*

$$D(m,n,t) = \sum_{k=1}^{D(t)} \mathbb{1}_{L(D_{k-1})=m}\, \mathbb{1}_{Y_k > n-m+1},$$

with Y_k the number of arrivals during the service time of the kth job. Then,

$$\begin{aligned}\lim_{t\to\infty}\frac{D(m,n,t)}{t} &= \lim_{t\to\infty}\frac{D(t)}{t}\frac{D(m,n,t)}{D(t)}\\ &= \delta\lim_{t\to\infty}\frac{D(m,n,t)}{D(t)}\\ &= \delta\mathsf{P}(Y>n-m+1)\\ &= \delta G(n-m+1),\end{aligned}$$

where G is the survivor function of Y and δ the departure rate.

0.2.45 (201807)**.** *In the level-crossing analysis of the $M(n)/M(n)/1$ queue we claim it is necessary that the interarrival times of jobs are i.i.d.*

0.2.46 (201807)**.** *To prove Little's law for any input-output system we claim that we need for all $T \geq 0$ the property*

$$\int_0^T L(s)\,ds = \sum_{k=1}^{A(T)} W_k.$$

0.2.47 (201807)**.** *For the $M/M/1$ queue we claim that $\mathsf{E}\left[L_Q\right] = \sum_{n=1}^{\infty}(n-1)\pi(n)$.*

0.2.48 (201807)**.** *A repair / maintenance facility would like to determine how many employees should be working in its tool crib. The service time is exponential, with mean 4 minutes, and customers arrive by a Poisson process with rate 28 per hour. With one employee we claim that the system is not rate stable.*

0.2.49 (201807)**.** *In the notes we derived that*

$$\frac{\mathsf{E}\left[L(M^X/M/1)\right]}{\mathsf{E}[L(M/M/1)]} = \frac{\mathsf{E}\left[B^2\right]}{2\,\mathsf{E}[B]} + \frac{1}{2},$$

when the loads in both queueing systems are the same. We claim that this implies for such systems that $\mathsf{E}\left[L(M^X/M/1)\right] \geq \mathsf{E}[L(M/M/1)]$.

0.2.50 (201807)**.** *For the $G/G/1$ the difference between the number of 'out transitions' and the number of 'in transitions' is at most 1 for all t. As a consequence,*

$$\begin{aligned}\textit{transitions out} &\approx \textit{transitions in} \iff\\ A(n,t)+D(n-1,t) &\approx A(n-1,t)+D(n,t) \iff\\ \frac{A(n,t)+D(n-1,t)}{t} &\approx \frac{A(n-1,t)+D(n,t)}{t} \iff\\ \frac{A(n,t)}{t}+\frac{D(n-1,t)}{t} &\approx \frac{A(n-1,t)}{t}+\frac{D(n,t)}{t}.\end{aligned}$$

Thus, under proper technical assumptions (which you can assume to be satisfied) this becomes for $t \to \infty$,

$$(\lambda(n)+\mu(n))p(n) = \lambda(n-1)p(n-1)+\mu(n+1)p(n+1).$$

We claim that if we specialize this result for the $M/D/1$ queue we have that $\lambda(n)=\lambda$ and $\mu(n)=\mu$, hence using PASTA,

$$(\lambda+\mu)\pi(n) = \lambda\pi(n-1)+\mu\pi(n+1).$$

0.2.51 (201807)**.** *For the $M^X/M/1$ queue we have shown in the notes that*

$$\mu \mathsf{E}[L] = \lambda \frac{\mathsf{E}[B^2]}{2} + \lambda \mathsf{E}[B]\mathsf{E}[L] + \lambda \frac{\mathsf{E}[B]}{2},$$

With a proper definition for the load ρ we claim that it can be rewritten to

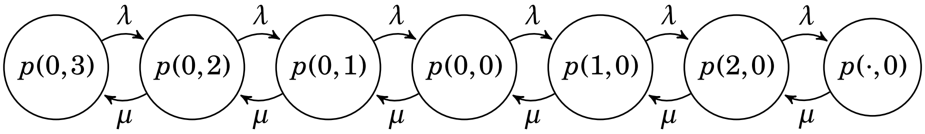
$$(1-\rho)\mathsf{E}[L] = \frac{\rho}{2}\left(\frac{\mathsf{E}[B^2]}{\mathsf{E}[B]}+1\right).$$

0.2.52 (201807)**.** *For the $M/G/1$ queue, let us concentrate on a down-crossing of level n; recall that level n lies between states n and $n+1$. We claim that job k only generates a down-crossing of level n this job leaves n jobs behind right after its service completion.*

0.2.53 (201807)**.** *In the notes we derive a recursion to be satisfied by the queue length distribution of the $M/G/1$ queue. To check whether this recursion holds for the $M/M/1$ we are lead to the computation below. Given that $\alpha = \rho/(1+\rho)$, we claim that the computation below entirely correct.*

$$\begin{aligned}\alpha^{n+1} + \sum_{m=1}^{n} \rho^m \alpha^{n-m+2} &= \alpha^{n+1} + \alpha^{n+2} \sum_{m=1}^{n} (\rho/\alpha)^m \\ &= \alpha^{n+1} + \alpha^{n+1}\rho \sum_{m=0}^{n-1} (\rho/\alpha)^m \\ &= \alpha^{n+1} + \alpha^{n+1}\rho \frac{1-(\rho/\alpha)^n}{1-\rho/\alpha} \\ &= \alpha^{n+1} - \alpha^{n+1}\frac{\rho}{\alpha}(1-(\rho/\alpha)^n).\end{aligned}$$

0.2.54 (201807)**.** *(Hall 5.22). At a large hotel, taxi cabs arrive at a rate of 15 per hour, and parties of riders arrive at the rate of 12 per hour. Whenever taxicabs are waiting, riders are served immediately upon arrival. Whenever riders are waiting, taxicabs are loaded immediately upon arrival. A maximum of three cabs can wait at a time (other cabs must go elsewhere). Let $p(i,j)$ be the steady-state probability of there being i parties of riders and j taxicabs waiting at the hotel. Claim: the transitions are modeled by the graph below.*



0.2.55 (201807)**.** *Just assume that the figure in the previous question is correct. Claim: the balance equations are as follows:*

$$\begin{aligned}\lambda p(0,3) &= \mu p(0,2) \\ (\lambda+\mu)p(0,2) &= \mu p(0,1) + \lambda p(0,3) \\ (\lambda+\mu)p(0,1) &= \mu p(0,0) + \lambda p(0,2) \\ (\lambda+\mu)p(0,0) &= \mu p(1,0) + \lambda p(0,1) \\ (\lambda+\mu)p(i,0) &= \mu p(i+1,0) + \lambda p(i-1,0)\end{aligned}$$

for $i \geq 1$

0.2.56 (201902). *A machine produces items, but a fraction p of the items produced in each period turns out to be faulty. Faulty items have to be repaired. The service time of faulty items is just as long as entirely new items. Repaired items are always ok, in other words, they cannot be faulty again. To keep the system stable, the average service capacity must satisfy*

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} c_i > \lambda(1+p)$$

where λ is the arrival rate of requests for items.

0.2.57 (201902). *Define $A(A_n-) = \lim_{h\downarrow 0} A(A_n - h)$. Then $A(A_n-) = n-1$.*

0.2.58 (201902). *The number of jobs in the system at time t is equal to*

$$L(t) = A(t) - D(t) = \sum_{k=1}^{\infty} \mathbb{1}_{D_k < t < A_k}.$$

0.2.59 (201902). *Let $N_{\lambda+\mu}$ be a Poisson process with rate $\lambda+\mu$. If $\{a_k\}$ is an i.i.d. sequence of Bernoulli random variables such that $\mathsf{P}(a_k = 1) = \lambda/(\lambda+\mu) = 1 - \mathsf{P}(a_k = 0)$, the random variable*

$$N(t) = \sum_{k=1}^{\infty} a_k \mathbb{1}_{k \leq N_{\lambda+\mu}(t)},$$

is Poisson distributed with rate μt.

0.2.60 (201902). *We consider the M/G/1 queue such that $\mathsf{E}[X] > \mathsf{E}[S]$. In general the expected busy time of the server is $\mathsf{E}[B] = \rho$. (Recall, ρ is the long-run fraction of time the server is busy.)*

0.2.61 (201902). *Consider a G/G/1 queue that is rate-stable. The distribution of the waiting times at arrival times can be sensibly defined as*

$$\mathsf{P}(W \leq x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{W_k \leq x}.$$

0.2.62 (201903). *Consider the G/G/1 queue of the previous exercise. Define*

$$U(t) = \max\{n : D_n \leq t\}.$$

The service rate is

$$\mu = \lim_{t\to\infty} \frac{U(t)}{t}.$$

0.2.63 (201903). *Consider the G/G/1 queue of the previous exercise. Then $(\mathsf{E}[X] - \mathsf{E}[S])/\mathsf{E}[X]$ is the fraction of time the server is idle.*

0.2.64 (201903). *For the G/G/1 queue define*

$$A(n,t) = \sum_{k=1}^{\infty} \mathbb{1}_{A_k \leq t} \mathbb{1}_{L(A_k -)=n},$$

where $L(s)$ is the number of customers in the system at time s. Then $A(n,t)$ counts the number of arrivals up to time t that saw n customers in the system at their arrival.

0.2.65 (201903). *The data below specifies the arrival time of customers, e.g., Jan arrives at time 21. The following code is guaranteed to print the customers in order of arrival*

```
from heapq import heappop, heappush


stack = []

heappush(stack, (21, "Jan"))
heappush(stack, (20, "Piet"))
heappush(stack, (18, "Klara"))
heappush(stack, (25, "Cynthia"))

print(stack)
```

0.2.66 (201903)**.** *For the $M/M/c$ queue we can take*

$$\lambda(n) = \lambda,$$
$$\mu(n) = \begin{cases} n\mu, & \text{if } n \leq c, \\ c\mu, & \text{if } n \geq c. \end{cases}$$

Then $p(n) = p(0)(c\rho)^n/n!$ for all n.

0.2.67 (201903)**.** *We can use the PASTA property to conclude that $\sum_{n=0}^{\infty} n p(n) = \sum_{n=0}^{\infty} n\pi(n)$ for any $G/M/1$ queue.*

0.2.68 (201903)**.** *When $\lambda > \delta$, then $\pi(n) < \delta(n)$.*

0.2.69 (201903)**.** *For the $M/G/1$ queue,*

$$\mathsf{E}[W_Q] = \mathsf{E}[L]\,\mathsf{E}[S],$$

that is, the expected time in queue is the expected number of customers in the system times the expected service time of these customers.

0.2.70 (201903)**.** *For the $M/M/1$ queue we have that $\mathsf{P}(L = n) = (1-\rho)\rho^n$. We can use the relation $\sum_{i=1}^{n} i = n(n+1)/2$ to see that $n^2 = -n + 2\sum_{i=1}^{n} i$. Using this result it follows that*

$$\begin{aligned} \mathsf{E}[L^2] &= (1-\rho)\sum_{n=0}^{\infty} n^2\rho^n \\ &= (1-\rho)\sum_{n=0}^{\infty}\left(\sum_{i=1}^{\infty} 2i\,\mathbb{1}_{i\leq n} - n\right)\rho^n \\ &= (1-\rho)\sum_{n=0}^{\infty}\sum_{i=0}^{\infty} 2i\,\mathbb{1}_{i\leq n}\rho^n - \sum_{n=0}^{\infty} n\rho^n. \end{aligned}$$

0.2.71 (201903)**.** *A single-server queueing system is known to have Poisson arrivals and exponential service times. However, the arrival rate and service time are state dependent. The manager observes that when the queue becomes longer, servers work faster, and the arrival rate declines. The following choice for $\lambda(n)$ and $\mu(n)$ are consistent with the manager's observation: $\lambda(0) = 5$, $\lambda(1) = 3$, $\lambda(2) = 2$, $\lambda(n) = 0, n \geq 3$, $\mu(0) = 0$, $\mu(1) = 2$, $\mu(2) = 3$, $\mu(n) = 4, n \geq 3$.*

0.2.72 (201903)**.** *Consider the $G/G/2$ queue with $P(0) = 0.4$, $P(1) = 0.3$, $P(2) = 0.2$, $P(3) = 0.05$, $P(4) = 0.05$. (Here, $P(n)$ is the fraction of time the system contains n jobs.) Then with the following code we can compute the (time) average number of jobs in queue.*

```
>>> P = [0.4, 0.3, 0.2, 0.05, 0.05]
>>> ELQ = sum(n*P[n] for n in range(len(P)))
```

0.2.73 (201903)**.** *Consider a queueing system in which we normally have 1 server working at rate $\mu = 4$. When the queue becomes longer than a threshold at 20, we hire one extra server that also works at rate $\mu = 4$, and when the queue is empty again, we send the extra servers home, until the queue hits 20 again, and so on. The following code implements this behavior of the extra server in a correct way.*

```
import numpy as np

from scipy.stats import poisson

labda = 3
mu = 4
a = poisson(labda).rvs(100000)
Q = np.zeros_like(a)
d = np.zeros_like(a)

threshold = 20

for i in range(1, len(a)):
 if Q[i-1] < threshold:
 c = poisson(mu).rvs()
 elif Q[i-1] >= threshold:
 c = poisson(2*mu).rvs()
 d[i] = min(Q[i-1],c)
 Q[i] = Q[i-1] + a[i] - d[i]
```

0.2.74 (201903)**.** *Consider the $M^X/M/1$ queue with B denoting the batch size of an arriving batch. Then*

$$\frac{\mathsf{E}[B^2]}{\mathsf{E}[B]} = (1 + C_s^2)\mathsf{E}[B], \quad \textit{where } C_s^2 = \frac{\mathsf{V}[B]}{(\mathsf{E}[B])^2},$$

0.2.75 (201903)**.** *Consider the $M/G/1$ queue. Denote by $\tilde{A}_k$ the time job k starts service and by D_k its departure time, $k = 1, \ldots, n$. Then, the expression*

$$\sum_{k=1}^{n} \int_0^{D_n} (D_k - s) \mathbb{1}_{\tilde{A}_k \leq s < D_k} \, ds$$

computes the total remaining service time up to time $t = D_n$.

0.2.76 (201904)**.** *A machine serves two types of jobs. The processing time of jobs of type i, $i = 1, 2$, is exponentially distributed with parameter μ_i. The type T of a job is random and independent of anything else, and such that $\mathsf{P}(T = 1) = p = 1 - q = 1 - \mathsf{P}(T = 2)$. Then,*

$$\mathsf{E}[S] = p/\mu_1 + q/\mu_2.$$

0.2.77 (201904)**.** *The $M/G/c/K$ shorthand means that jobs arrive as a Poisson process, job service times are exponentially distributed, and there are c servers.*

0.2.78 (201904). *Consider the server of the G/G/1 queue as a system by itself. The time jobs stay in this system is* $\mathsf{E}[S]$*, and jobs arrive at rate* λ*. It follows from Little's law that the fraction of time the server is busy is* $\lambda\,\mathsf{E}[S]$.

0.2.79 (201904). *If* $A(n,t) = \sum_{k=1}^\infty \mathbb{1}_{A_k \leq t}\, \mathbb{1}_{L(A_k-)=n}$, *is* $A(t) = \sum_{n=0}^\infty A(n,t)$*?*

0.2.80 (201904). *If* $\lambda > \delta$ *it can happen that* $\lim_{t\to\infty} A(n,t)/t > 0$ *for some (finite) n.*

0.2.81 (201904). *Let*

$$D(n,t) = \sum_{k=1}^\infty \mathbb{1}_{D_k \leq t}\, \mathbb{1}_{L(D_k)=n}, \qquad Y(n,t) = \int_0^t \mathbb{1}_{L(s)=n}\, ds$$

denote the number of departures up to time t that leave n customers behind and the total time the system contains n jobs during $[0,t]$*. Then, the departure rate from state* $n+1$ *is*

$$\mu(n+1) = \lim_{t\to\infty} \frac{D(n+1,t)}{Y(n+1,t)},$$

0.2.82 (201904). *As* $K \to \infty$*, the performance measures of the M/M/1/K converge to those of the M/M/1 queue.*

0.2.83 (201904). *To model the* $M/M/c/c+K$ *queue as an* $M(n)/M(n)/1$ *queue we need to take* $\lambda(n) = \lambda$ *for all n.*

0.2.84 (201904). *Take*

$$\pi(n) = \lim_{t\to\infty} \frac{A(n,t)}{A(t)}, \qquad \delta(n) = \lim_{t\to\infty} \frac{D(n,t)}{D(t)}.$$

Then, $\lambda \neq \delta \implies \pi(n) > \delta(n)$.

0.2.85 (201904). *Consider a* $M/D/1$ *queue with* $\lambda = 1$ *and* $\mathsf{E}[S] = 0.10$. *Then the SCV of its departure process is smaller than* 0.5.

0.2.86 (201904). *For a given single-server queueing system the average number of customers in the system is* $\mathsf{E}[L] = 10$, *customers arrive at rate* $\lambda = 4$ *per hour and are served at rate* $\mu = 5$ *per hour. At the moment you join the system, the number of customers in the system is 10. Your expected time in the system is, by Little's law,* $\mathsf{E}[W] = \mathsf{E}[L]/\lambda = 2.5$ *hour.*

0.2.87 (201904). *When* $\mathsf{V}[S] = 0$, *it follows for the remaining service time* S_r *that*

$$\mathsf{E}[S_r \,|\, S_r > 0] = \frac{\mathsf{E}\left[S^2\right]}{2\mathsf{E}[S]} \implies \mathsf{E}[S_r \,|\, S_r > 0] = \frac{\mathsf{E}[S]}{2}$$

0.2.88 (201907). *For the* $M/G/1$ *queue with rate* $\lambda = 1$ *per hour,* $\mathsf{P}(A_k = k \text{ for all } k) > 0$.

0.2.89 (201907). *For the computation of the waiting time of the single-server queue we assume that all random variables in the sequences* $\{X_k\}$ *and* $\{S_k\}$ *are independent. This a necessary condition to compute the set of waiting times* $\{W_k\}$.

0.2.90 (201907). *Let* $N_{\lambda+\mu}$ *be a Poisson process with rate* $\lambda + \mu$. *If* $\{a_k\}$ *is an i.i.d. sequence of Bernoulli random variables such that* $\mathsf{P}(a_k = 1) = \mu/(\lambda+\mu) = 1 - \mathsf{P}(a_k = 0)$, *the random variable*

$$N(t) = \sum_{k=1}^{\infty} a_k \, \mathbb{1}_{k \leq N_{\lambda+\mu}(t)},$$

has a Poisson distribution with rate λt.

0.2.91 (201907). *Take the stable* $M(n)/M/1$ *queue with* $\lambda(15) = 0$. *Suppose that the queue length starts at* 100, *i.e.,* $Q(0) = 100$. *Then* $\pi(90) > 0$.

0.2.92 (201907). *If* $L(t)/t \to 0$ *as* $t \to \infty$ *it can still be true that* $0 < \mathsf{E}[L] < \infty$.

0.2.93 (201907). *Consider the (stable)* $M/G/1$ *queue. The density* f_D *of the interdeparture times is equal to the density* f_S *of the service times.*

0.2.94 (201907). *Suppose there are 10 jobs present at the* $M/M/1$ *queue with arrival rate* $\lambda = 3$ *and service rate* $\mu = 4$ *per hour. The time to clear the system follows from Little's law and is* $3 \cdot 10 = 30$ *hours.*

0.2.95 (201907). *For the* $M^X/M/1$ *we have the recursion*

$$\pi(n) = \frac{\lambda}{\mu} \sum_{i=0}^{n-1} \pi(n-1-i) G(i).$$

The following code can be used to compute the unnormalized *probabilities for* $p(1), \ldots, p(4)$ *of the* $M/M/1$ *queue. (Recall that* `range[1,5]` *goes up to, but does not include,* 5.*)*

```
p[0] = 1
G = [1, 0]
for n in range(1, 5): # this goes up to, but does not include, 5
 p[n] = labda / mu * sum(p[n - 1 - i] * G[i] for i in range(min(n, len(G))))
```

0.2.96 (201907)**.** *For the $M/G/1$ queue the up-crossing rate of level n is equal to*

$$\delta\delta(0)G(n)+\delta\sum_{m=1}^{n}\delta(m)G(n-m+1), \tag{0.2.4}$$

where $G(j)=\mathsf{P}(Y>j)$ is the survivor function of the number of arrivals Y during a service time, δ is the long-run departure rate and $\delta(n)$ the probability to leave n jobs behind.

0.2.97 (201907)**.** *Consider a single-server queueing in discrete time, $k=0,1,\dots$. The machine can switch between a high and a slow production speed. When the queue is larger than or equal to M at the start of period k, the machine switches to a high speed c_+; when the queue becomes smaller or equal to m, $0\leq m<M$, the machine switches to the low speed $c_-<c_+$, otherwise the machine's speed remains the same. The variable I_k keeps track of the state of the server and satisfies*

$$I_{k+1}=c_+\,\mathbb{1}_{Q_k\geq M}+I_k\,\mathbb{1}_{m<Q_k<M}+c_-\,\mathbb{1}_{Q_k\leq m}.$$

0.2.98 (201907)**.** *Consider the (stable) $M/D/1$ queue. The SCV of the departure process is always less than 1.*

0.2.2 *Open Questions*

0.2.99 (201704)**.** *Show with a level-crossing argument that*

$$\lambda\pi(n)=\mu(n+1)p(n+1) \tag{0.2.5}$$

for a queueing system in which jobs arrive and depart in single units.

0.2.100 (201704)**.** *What condition should be satisfied in the above Equation* (0.2.5) *so that PASTA holds?*

0.2.101 (201704)**.** *The server of an $M/M/1$ queue fails, with constant failure rate, once per 10 days. The repair times are exponentially distributed with a mean of one day. The job service times without failures have a mean of 2 days. Jobs arrive with rate $\lambda=1/3$ per day. What is $\mathsf{E}[W]$?*

0.2.102 (201704)**.** *Consider the $M^X/M/1$ queue with $G(n)=\mathsf{P}(B>n)$. With level-crossing arguments we* nearly *get*

$$\lambda\sum_{m=0}^{n}G(n-1-m)\pi(m)=\mu\pi(n+1).$$

What is wrong with this formula, and repair it.

0.2.103 (201706)**.** *Show that an exponentially distributed random variable is memoryless.*

0.2.104 (201706)**.** *Suppose an $M/M/1$ queue contains 5 jobs at time some $t>0$. Jobs arrive at rate λ and have average service time μ^{-1}. What is the distribution of the time until the next event (arrival or departure epoch)?*

0.2.105 (201706)**.** *Suppose an $M/M/4$ queue contains 5 jobs at time some $t > 0$. Jobs arrive at rate λ and have average service time μ^{-1}. What is the distribution of the time until the next event (arrival or departure epoch)?*

0.2.106 (201706)**.** *If the inter-arrival times $\{X_i\}$ are i.i.d. and exponentially distributed with mean $1/\lambda$, prove that the number $N(t)$ of arrivals during interval $[0,t]$ is Poisson distributed. Use that $\mathsf{P}(A_k \leq t) = \int_0^t \lambda e^{-\lambda s}\frac{(\lambda s)^{k-1}}{(k-1)!}\,ds$, where A_k is the arrival time of the kth arrival.*

0.2.107 (201706)**.** *Consider a single-server queueing in discrete time, $k = 0,1,\ldots$. The machine can switch between a high and a low production speed. When the queue is larger than or equal to M at the start of period k, the machine switches to the high speed c_+; when the queue becomes smaller or equal to m, $0 \leq m < M$, the machine switches to the low speed $c_- < c_+$, otherwise the machine's speed remains the same. The arrival process is given by $\{a_k\}$. Assuming the arrivals a_k in period k cannot be served on day k, establish a set of recursions to enable a simulation of this system. Assume that $Q_0 = 0$.*

0.2.108 (201706)**.** *For the previous problem, provide a formula to compute (count) the number of times the machine switches to the fast state during the first n periods.*

0.2.109 (201706)**.** *Related to the previous problem, assuming that the average arrival rate of jobs $a = \lim_{n\to\infty} 1/n \sum_k^n a_k$ is such that $c_- < a < c_+$. What is the average cycle time, i.e., the average time between two moments the machine switches to the fast state? (More specifically, let τ_n be the nth time the machine switches to the fast state. What is $\lim_{n\to\infty} n^{-1}\,\mathsf{E}[\tau_n]$?)*

0.2.110 (201706)**.** *Provide a real-world example for the queueing model of the previous three problems.*

0.2.111 (201706)**.** *Can you make an arrival process such that $\frac{A(t)}{t}$ as $t \to \infty$ does not have a limit?*

0.2.112 (201706)**.** *Consider the $M/M/1$ queue with arrival rate $\lambda = 3$ per hour and average service times $\mathsf{E}[S] = 10$ minutes. What is the value of*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{L(A_k-)=3}?$$

0.2.113 (201706)**.** *State explicitly all relevant assumptions you needed to make in the previous problem to obtain an answer.*

0.2.114 (201706)**.** *The queueing system at a fast-food stand behaves in a peculiar fashion. When there is no one in the queue, people are reluctant to use the stand, fearing that the food is unsavory. People are also reluctant to use the stand when the queue is long. This yields the following arrival rates (in numbers per hour): $\lambda(0) = 10$, $\lambda(1) = 15$, $\lambda(2) = 15$, $\lambda(3) = 10$, $\lambda(4) = 5$, $\lambda(n) = 0, n \geq 5$. The stand has two servers, each of which can operate at 5 customers per hour. Service times are exponential, and the arrival process is Poisson. Calculate the steady-state probabilities.*

0.2.115 (201706)**.** *For the previous question, customers spend 10 Euro on average on an order. What is the rate at which the stand makes money? What theorem(s) do you need to use to compute this?*

0.2.116 (201804)**.** *Show with a level-crossing argument that for the $M/M/1$ queue*

$$\pi(n) = \delta(n). \tag{0.2.6}$$

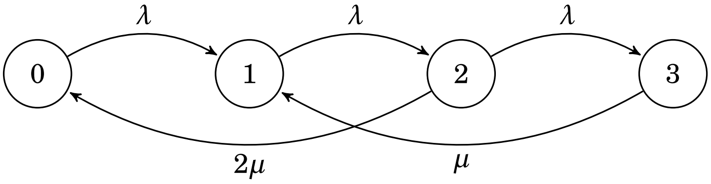
0.2.117 (201804)**.** *What condition should be satisfied in* (0.2.6) *so that it holds?*

0.2.118 (201804)**.** *Does* (0.2.6) *also hold for the* $G/G/1$ *queue (motivate).*

0.2.119 (201804)**.** *Does* (0.2.6) *imply the PASTA property (motivate).*

0.2.120 (201804)**.** *Use the Pollack-Khintchine equation to show for the* $M/M/1$ *queue that* $\mathsf{E}[L] = \rho/(1-\rho)$.

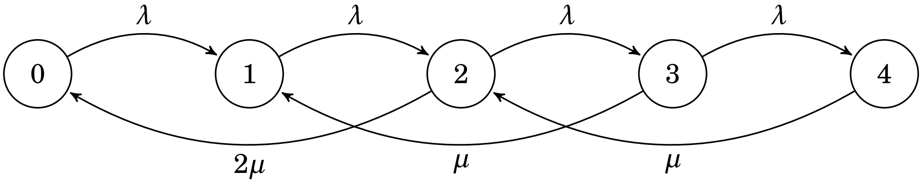
0.2.121 (201807)**.** *Consider the queueing system below. Jobs arrive as a Poisson process with rate* λ. *Service times are exponential service with mean* $(2\mu)^{-1}$ *when there are two jobs in the system, and mean* μ^{-1} *when there are 3 jobs. What is* $p(2)$ *if* $\lambda = \mu = 1$?



0.2.122 (201807)**.** *What is the fraction of lost jobs?*

0.2.123 (201807)**.** *Suppose the manager is unsatisfied with the loss rate. Writing* $x = \lambda/\mu$ *for ease, what condition should* x *satisfy such that the loss probability is less than some threshold* α? *(Just show the condition on* x*; you do not have to solve for* x*.)*

0.2.124 (201807, Continuation of previous exercise, 1)**.** *Assume again that* $\lambda = \mu = 1$. *Would the loss be reduced if the manager extends the system to four* 4 *positions like this:*



0.2.125 (201807)**.** *Consider the* $G/G/n/K$ *queue (specifically, the system can contain at most* K *jobs). Let* λ *be the arrival rate,* μ *the service rate,* β *the long-run fraction of customers lost, and* ρ *the average number of busy / occupied servers Show that*

$$\beta = 1 - \rho\frac{\mu}{\lambda}. \tag{0.2.7}$$

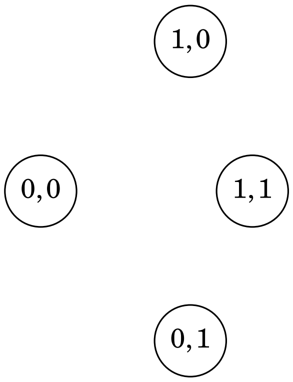
0.2.126 (201807)**.** *Explain why* (0.2.7) *not applicable for the queueing system of question* **0.2.121**.

0.2.3 *Parallel servers*

In the $M/M/c$ queue it is assumed that all servers work at the same rate. However, in practice, this is not always the case. In this section, from **0.2.127**–**1.1.8**, we make a model by which we can analyze the effects of servers with different speeds.

A station has two parallel servers, the service times of the first are $\text{Exp}(\mu_1)$ and of the second $\text{Exp}(\mu_2)$. Jobs arrive as a Poisson process with rate λ. The queue cannot contain a job, hence any job arriving when both servers are busy is rejected. When the system is empty, jobs are routed to the first server.

0.2.127 (201904). *Explain that the figure below contains all the relevant states.*



0.2.128 (201904). *Add arrows to the figure of the state space to indicate the possible transitions and add the rates.*

0.2.129 (201904). *Find an expression for the long-run fraction of time the system is empty. (You might want to use balance equations here.)*

0.2.130 (201904). *Express the long-run fraction of lost jobs in terms of one of the probabilities* $p(0,0)$, $p(1,0)$, $p(0,1)$, $p(1,1)$. *Explain your reasoning.*

0.2.131 (201904). *By making the simplifying assumption that* $\mu_1 = \mu_2$ *the above system reduces to a simpler queueing system for which we have closed-form solutions for the state probabilities. What is this simpler queueing system?*

Suppose we can extend the system such that one job can be queued.

0.2.132 (201904). *Make a sketch of the state space and include the transitions and rates.*

0.2.133 (201904). *Show how you can use level-crossing arguments to express the probability to find a job in queue in terms of one of the probabilities* $p(0,0)$, $p(1,0)$, $p(0,1)$, $p(1,1)$.

0.2.134 (201904). *Suppose now that the queue is infinite so that jobs are never lost. Approximate the queueing system by a suitable G/G/c queue and use Sakasegawa's formula to estimate* $\mathsf{E}[W_Q]$. *Take* $\mu_1 = 6$, $\mu_2 = 3$, $\lambda = 8$.

0.2.135 (201904). *For the* $M^X/M/1$ *queue we derived the expression that* $p(n)$, *i.e., the fraction of time the system contains* n *jobs, must satisfy*

$$\lambda \sum_{m=0}^{n} G(n-m)p(m) = \mu p(n+1), \tag{0.2.8}$$

where $G(k) = \mathsf{P}(B > k)$. *Explain this formula, by a drawing or in words (or both).*

0.2.136 (201904). *For the* $M^X/M/1$ *queue we have that*

$$\mu \mathsf{E}[L] = \mu \sum_{n=0}^{\infty} n\pi(n) = \lambda \frac{\mathsf{E}[B^2]}{2} + \lambda \mathsf{E}[B]\,\mathsf{E}[L] + \lambda \frac{\mathsf{E}[B]}{2}. \tag{0.2.9}$$

Show that this reduces to $\mathsf{E}[L] = \rho/(1-\rho)$ *for the* $M/M/1$ *queue.*

0.2.137 (201904). *Explain (briefly) why checks such as in the previous exercise are important.*

Solutions

s.0.2.1. Answer = B. We should define

$$\alpha = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} L(A_k-).$$

s.0.2.2. Answer = A.

s.0.2.3. Answer = A.

$$\mathsf{E}[L_Q] = \frac{\rho^2}{1-\rho}.$$

s.0.2.4. Answer = A.

s.0.2.5. Answer = A.

s.0.2.6. Answer = A.

s.0.2.7. Answer = B. $\rho = \lambda/(c\mu)$, and the rest of the formula is also wrong:

$$p(n+1) = \frac{\lambda}{\mu(n+1)}p(n) = \frac{\lambda}{\min\{c, n+1\}\mu}p(n)$$

etc.

s.0.2.8. Answer = B. $\pi(0) = 1$.

s.0.2.9. Answer = A.

s.0.2.10. Answer = B.

$$\mathsf{E}[L] = \left.\frac{d}{dz}\phi(z)\right|_{z=1} = \frac{\rho}{1-\rho}.$$

s.0.2.11. Answer = B. We have

$$\begin{aligned}\frac{\lambda}{\mu}\mathsf{E}\left[B^2\right] &= \frac{\lambda\,\mathsf{E}[B]}{\mu}\frac{\mathsf{E}\left[B^2\right]}{(\mathsf{E}[B])^2}\mathsf{E}[B] = \rho\frac{\mathsf{E}\left[B^2\right]}{(\mathsf{E}[B])^2}\mathsf{E}[B] \\ &= \rho\frac{(\mathsf{E}[B])^2 + \mathsf{V}[B]}{(\mathsf{E}[B])^2}\mathsf{E}[B] = \rho(1+C_s^2)\mathsf{E}[B].\end{aligned}$$

s.0.2.12. Answer = A.

s.0.2.13. Answer = A.
$\rho = \lambda\,\mathsf{E}[S] = 6/8 = 3/4$. Hence, $\rho/(1-\rho) = 3$.

$$\mathsf{E}\left[W_Q\right] = \frac{1+1/2}{2}3\,\mathsf{E}[S] = 3/4\cdot 3\cdot 2 = 9/2 = 4.5$$

s.0.2.14. Answer = A.

s.0.2.15. Answer = A.

s.0.2.16. Answer = B. When $a_k = 0$, there are no job arrivals, but still L_k can be positive.

s.0.2.17. Answer = A.

s.0.2.18. Answer = A.

s.0.2.19. Answer = A.

s.0.2.20. Answer = B.

$$\frac{1}{t}\sum_{k=1}^{A(t)} \mathbb{1}_{W_k \leq x} = \frac{A(t)}{t}\frac{1}{A(t)}\sum_{k=1}^{A(t)} \mathbb{1}_{W_k \leq x} \to \lambda\mathsf{P}(W \leq x).$$

Moreover, the last w in the equation in the exercise is also wrong, and should be an x.

s.0.2.21. It is wrong. Suppose that the system starts empty and only one job arrives at $t = 1/3$. Consider now a time $t = 2/3$. Then $Y(0,t) = 1/3$, $Y(1,t) = 1/3$, $A(0,t) = 1$ and $D(n,t) = 0$. Then the expression at the left-hand side becomes 3/2. Hence, the statement is not correct for all $t > 0$;

s.0.2.22. Answer = A.

s.0.2.23. Answer = A.

s.0.2.24. Answer = B. There were some questions about this. So lets explain this a bit better. In a system with finite calling population, b say, the arrival rate is $\lambda(n) = \lambda(b-n)$ when there are n jobs in service or in queue. For a queueing system with balking *at* level b, $\lambda(n) = \lambda$ for all $n < b$, and $\lambda(n) = 0$ for $n \geq b$. Thus, the arrival processes are different, hence the queueing systems must behave differently.

s.0.2.25. Answer = A.

s.0.2.26. Answer = A.

s.0.2.27. Answer = A.

s.0.2.28. Answer = B.

s.0.2.29. Answer = B. $\mathsf{P}(L \leq n) = (1-\rho)\sum_{k=0}^{n} \rho^k$. It's evident in the question that the normalization misses.

s.0.2.30. Answer = A.

s.0.2.31. Answer = B.

s.0.2.32. Answer = B.

$$\lambda\,\mathsf{E}\left[S^2\right] = (1+C_s^2)\rho\,\mathsf{E}[S], \quad \text{where } C_s^2 = \frac{\mathsf{V}[S]}{(\mathsf{E}[S])^2}. \tag{0.2.10}$$

s.0.2.33. Answer = B. It should be $L(A-)$.

s.0.2.34. Answer = B. It should be $A(n,n,t)$.

s.0.2.35. Answer = A.

s.0.2.36. Answer = B. The expressions become (in the limit $t \to \infty$) to the time-average of the number of jobs in the system. In general, this time-average is not what the jobs see upon arrival.

s.0.2.37. Answer = B. The figure sketches the $M/M^2/1/3$ queue.

s.0.2.38. Answer = A. When the server is busy at time 0, we need to wait until the job in service departs, as this is the next departure. As service times are Exp(μ) and memoryless, the expression follows.

s.0.2.39. Answer = A.

s.0.2.40. Answer = B. It's precisely the other way around.

s.0.2.41. Answer = A.

s.0.2.42. I decided to accept any answer to this exercise, as it is possible to read it in two ways, which I did not realize at first. One way is that the ith item is the one such that still i items remain (in which case the equation is correct). The other is that $i-1$ items have been served, and that the server is processing the ith item in the sequence, in which case the system contains $B-i-1$.

s.0.2.43. Answer = B. It should be this:

$$\begin{aligned}\mathsf{E}\left[S^2\right] &= \mu\int_0^\infty x^2 e^{-\mu x}\,\mathrm{d}x = -\,x^2 e^{-\mu x}\big|_0^\infty + 2\int_0^\infty x e^{-\mu x}\,\mathrm{d}x \\ &= -\,2\frac{x}{\mu}e^{-\mu x}\Big|_0^\infty + \frac{2}{\mu}\int_0^\infty e^{-\mu x}\,\mathrm{d}x \\ &= -\frac{2}{\mu}\,e^{-\mu x}\big|_0^\infty = \frac{2}{\mu^2}.\end{aligned}$$

Note the minus signs.

s.0.2.44. Answer = B.

s.0.2.45. Answer = B.

s.0.2.46. Answer = B. This only holds at times T at which the system is empty.

s.0.2.47. Answer = A.

s.0.2.48. Answer = A.

s.0.2.49. Answer = A.

s.0.2.50. Answer = B.

s.0.2.51. Answer = A.

s.0.2.52. Answer = B. In fact, two events must be satisfied:

$$\text{Down-crossing of level } n \iff \mathbb{1}_{L(D_{k-1})=n+1}\,\mathbb{1}_{L(D_k)=n} = 1.$$

s.0.2.53. Answer = B. The last line should be

$$\alpha^{n+1} - \alpha^{n+1}(1-(\rho/\alpha)^n).$$

s.0.2.54. Answer = A.

s.0.2.55. Answer = A.

s.0.2.56. Answer = A.

s.0.2.57. Answer = A.

s.0.2.58. Answer = B. $D_k \geq A_k$ always.

$$L(t) = A(t) - D(t)\sum_{k=1}^\infty \mathbb{1}_{A_k\leq t} - \sum_{k=1}^\infty \mathbb{1}_{D_k\leq t}.$$

s.0.2.59. Answer = B. It is Poisson distributed with rate λt.

s.0.2.60. Answer = B. The server utilization is ρ. Check exercise 1.7.10.

s.0.2.61. Answer = A.

s.0.2.62. Answer = B.

s.0.2.63. Answer = A.

s.0.2.64. Answer = A.

s.0.2.65. Answer = B.

```
>>> from heapq import heappop, heappush


>>> stack = []

>>> heappush(stack, (21, "Jan"))
>>> heappush(stack, (20, "Piet"))
>>> heappush(stack, (18, "Klara"))
>>> heappush(stack, (25, "Cynthia"))

>>> print(stack)
[(18, 'Klara'), (21, 'Jan'), (20, 'Piet'), (25, 'Cynthia')]
```

s.0.2.66. Answer = B.

s.0.2.67. Answer = B. In the $G/M/1$ queue jobs don't arrive as a Poisson process.

s.0.2.68. Answer = B.
In fact, if $\lambda > \delta$, then $p(n) = 0 = \delta(n)$ for all n.

s.0.2.69. Answer = B. In the $M/G/1$ queue the remaining service time of the job in service (if there is any), is not $\mathsf{E}[S]$.

s.0.2.70. Answer = B.

$$\begin{aligned}
\mathsf{E}\left[L^2\right] &= (1-\rho)\sum_{n=0}^{\infty} n^2 \rho^n \\
&= (1-\rho)\sum_{n=0}^{\infty}\left(\sum_{i=1}^{\infty} 2i\,\mathbb{1}_{i\leq n} - n\right)\rho^n \\
&= (1-\rho)\sum_{n=0}^{\infty}\sum_{i=0}^{\infty} 2i\,\mathbb{1}_{i\leq n}\rho^n - (1-\rho)\sum_{n=0}^{\infty} n\rho^n.
\end{aligned}$$

s.0.2.71. Answer = A.

s.0.2.72. Answer = B. It's the number of jobs in the system, not in queue.

s.0.2.73. Answer = B. The extra server switches off when the queue length becomes below 20.

s.0.2.74. Answer = A.

s.0.2.75. Answer = A.

s.0.2.76. Answer = A, **2.1.16**.

s.0.2.77. Answer = B. **1.3.8**

s.0.2.78. Answer = A, **4.4.5**.

s.0.2.79. Answer = A, **4.2.2**

s.0.2.80. Answer = B, **4.2.4**

s.0.2.81. Answer = B, **4.2.8**

s.0.2.82. Answer = A, **5.2.2**

s.0.2.83. Answer = B, **5.2.11**

s.0.2.84. Answer = B, **4.3.4**.

s.0.2.85. Answer = B.

```
>>> labda = 1
>>> ES = 0.10
>>> Ca = 1
>>> Cs = 0
>>> rho = labda*ES
>>> rho
0.1
>>> Cd = (1-rho*rho)*Ca + rho*rho*Cs
>>> Cd
0.99
```

s.0.2.86. Answer = B, **4.4.7**. Depending on the service distribution, the expected time can be 10/2, but it can also be something else. This depends on the distribution of the remaining service time of the job in service at the moment you arrive.

s.0.2.87. Answer = A, **5.4.19**

s.0.2.88. Answer = B. $\mathsf{P}(A_k = k \text{ for all } k) = 0$.

s.0.2.89. Answer = B. Of course this is not necessary. For instance, in a simulation these random variables are not independent.

s.0.2.90. Answer = B, **2.2.2**

s.0.2.91. Answer = B. It is zero. Note that the services are described by an M, hence the service rate is μ at all stations. Since the queue is stable, it is necessary that $\mu > 0$.

s.0.2.92. Answer = A, **2.4.2**.

s.0.2.93. Answer = B, **6.3.7**.

s.0.2.94. Answer = B.

s.0.2.95. Answer = A.

s.0.2.96. Answer = A, see (5.6.5).

s.0.2.97. Answer = A.

s.0.2.98. Answer = A.

s.0.2.99. $|A(n,t) - D(n,t)| \leq 1$. Hence

$$\frac{A(t)}{t}\frac{A(n,t)}{A(t)} \approx \frac{D(n+1,t)}{Y(n+1,t)}\frac{Y(n+1,t)}{t}.$$

Now take the limit $t \to \infty$.

Just saying that level-crossing implies $\lambda\pi(n) = \mu(n+1)p(n+1)$ is of course not a good answer.

s.0.2.100.

$$\frac{A(n,t)}{t} = \frac{A(t)}{t}\frac{A(n,t)}{A(t)}\frac{A(n,t)}{Y(n,t)}\frac{Y(n,t)}{t}.$$

Now take the limit $t \to \infty$ to get

$$\lambda\pi(n) = \lambda(n)p(n).$$

Hence, when $\lambda = \lambda(n)$ for all n, $\pi(n) = p(n)$, which is the PASTA property.

Some students said that $p(n) = \pi(n)$, but this is the meaning of PASTA, not the condition for PASTA. (For the curious, there is an Anti-PASTA theorem.)

If you would just have said something like 'Poisson arrivals', or 'exponential interarrival times', I gave one point. The answer is not complete.

s.0.2.101. The availability is $A = 9/10$. Hence, $\mathsf{E}[S_e] = 2/(9/10)) = 20/9$. We also have

$$C_e^2 = C_0^2 + 2A(1-A)\frac{m_r}{\mathsf{E}[S_0]} = 1 + 2\frac{9}{10}\frac{1}{10}\frac{2}{.}$$

Now,

$$\rho = \lambda\,\mathsf{E}[S_e] = \frac{1}{3}\frac{20}{9} = \frac{20}{27}$$

and

$$\mathsf{E}\left[W_Q\right] = \frac{1+C_e^2}{2}\frac{\rho}{1-\rho}\mathsf{E}[S] = \frac{1+C_e^2}{2}\frac{20/27}{7/27}\frac{20}{9}.$$

Finally, $EW = \mathsf{E}\left[W_Q\right] + \mathsf{E}[S_e]$.

It is essential that you should that both $\mathsf{E}[S]$ and C_e^2 are affected by failures. If you forgot to compensate in either of the two, I subtracted one point.

s.0.2.102.

$$\lambda\sum_{m=0}^{n} G(n-m)\pi(m) = \mu\pi(n+1).$$

s.0.2.103. See the book.

No points if you do not explicitly use the exponential distribution.

s.0.2.104. See the book. $\mathsf{P}(\min\{X,S\} > x) = \exp-(\mu+\lambda)x$.

A number of students don't seem to understand the difference between time and number. The time to the next event is exponentially distributed. The fact that the number of jobs arriving in a certain amount of time is Poisson distributed is not the same as that the time to the next event is Poisson distributed.

Stating that the time is $\min\{X,Y\}$ where X and Y are exp. distr. random variables gives 1/2 point.

s.0.2.105. See the book. $\mathsf{P}(\min\{X,S_1,\dots,S_4\} > x) = \exp-(4\mu+\lambda)x$.

s.0.2.106. See the book.

We want to show that

$$\mathsf{P}(N(t)=k) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}.$$

Now observe that $\mathsf{P}(N(t)=k) = \mathsf{P}(A_k \leq t) - \mathsf{P}(A_{k+1} \leq t)$. Using the density of A_{k+1} as obtained previously and applying partial integration leads to

$$\begin{aligned}
\mathsf{P}(A_{k+1} \leq t) &= \lambda \int_0^t \frac{(\lambda s)^k}{k!} e^{-\lambda s}\, \mathrm{d}s \\
&= \lambda \frac{(\lambda s)^k}{k!} \frac{e^{-\lambda s}}{-\lambda}\Big|_0^t + \lambda \int_0^t \frac{(\lambda s)^{k-1}}{(k-1)!} e^{-\lambda s}\, \mathrm{d}s \\
&= -\frac{(\lambda t)^k}{k!} e^{-\lambda t} + \mathsf{P}(A_k \leq t)
\end{aligned}$$

We are done.

Half point for stating that we are looking for $\mathsf{P}(A_k \leq t) - \mathsf{P}(A_{k+1} \leq t)$. Some students reversed these two probabilities. It is simple to see that that is wrong.

For a full point I also wanted to see the algebra that leads to the final answer.

s.0.2.107. First we need to implement the switching policy. For this we need an extra variable to keep track of the state of the server. Let $I_k = c_+$ if the machine is working fast in period k and $I_k = c_-$ if it is working slowly. Then $\{I_k\}$ must satisfy the relation

$$I_{k+1} = \begin{cases} c_+ & \text{if } Q_k \geq M, \\ c_- & \text{if } Q_k <= m, \\ I_k & \text{else.} \end{cases}$$

and assume that $I_0 = c_+$ at the start, i.e., the machine if off. Thus, we can write:

$$I_{k+1} = c_+ \mathbb{1}_{Q_k \geq M} + I_k \mathbb{1}_{m < Q_k < M} + c_- \mathbb{1}_{Q_k \leq m}.$$

With I_k it follows that $d_k = \min\{Q_{k-1}, I_k\}$, from which Q_k follows, and so on.

I want to see an update rule for the state of the machine. If that is missing or wrong, 1/2 point.

s.0.2.108.

$$\sum_{k=1}^{n} \mathbb{1}_{I_{k-1}=c_-, I_k=c_+}.$$

The expression

$$\sum_{k=1}^{n} \mathbb{1}_{Q_k>m}$$

counts too many periods.

This,

$$\sum_{k=1}^{n} (\mathbb{1}_{I_{k-1}=c_-} - \mathbb{1}_{I_k=c_+})^2$$

counts all changes in speed.

s.0.2.109. The time T_1 to move from slow to fast satisfies $(a-c_-)T_1 = M-m$, and the time to move from fast to slow satisfies $(c_+-a)T_2 = M-m$. The total cycle time is T_1+T_2.

s.0.2.110. A machine that needs to be warm/hot to produce. When there are no jobs, it is better to switch it off ($m=0$ case); only switch it on when the queue length in front of it is longer than M. If $m>0$, assume that the machine can work at two speeds, but the higher speed requires more power.

s.0.2.111. For instance, let a_k be the number of arrivals in period k. Then take $a_1 = 1$, $a_2 = a_3 = 0$, $a_4 = a_5 = 1$, and then we have 3 period with no arrivals, and 3 periods with 1 arrivals and then 4 without and 4 with 1 arrival, and so on.

I accepted the following suggestion $A(t) = t^2$. However, formally the limit does exist, it is $\lim_{t\to\infty} t^2/t = \infty$.

s.0.2.112. By the PASTA property, arrivals of an $M/G/n$ queue see the time average stationary distribution. Thus, we can focus on the time-average probability that the system contains 3 jobs. This is $(1-\rho)\rho^3$. Here, $\rho = 3\cdot 10/60 = 1/2$. Clearly, $\rho < 1$, hence the system is stable.

s.0.2.113. See the previous problem: pasta + stability. Just mentioning PASTA is ok.

s.0.2.114. See the book. Note that the rate from state 1 to 0 is 5, not 10.

s.0.2.115. see the book. The rate at which customers are accepted is $\sum_n \lambda_n p_n$. Multiply this by 10 to get the rate at which the system makes money. Note that this is not $10\,\mathsf{E}[L]$. Customers pay to get served, not to stay in the line...

s.0.2.116. $A(n,t) \approx D(n,t)$. Hence

$$\frac{A(t)}{t}\frac{A(n,t)}{A(t)} \approx \frac{D(n,t)}{D(t)}\frac{D(t)}{t}.$$

Now take the limit $t\to\infty$ to get $\lambda\pi(n) = \delta(n)\delta$.

Using $D(n,t)/Y(n,t)\cdot Y(n,t)/t$ is plain wrong (check the book). $-1/2$.

s.0.2.117. Rate stability, i.e., $\lambda = \delta$.

s.0.2.118. Sure, to derive (0.2.6), we only used counting arguments, but nothing in particular about the interarrival times.

Some students say no, because it is not true in case of batch arrivals. Ok. but recall, by assumption we are dealing here with $G/G/1$ queue...

s.0.2.119. No, for PASTA we need more.

$$\frac{A(t)}{t}\frac{A(n,t)}{A(t)} = \frac{A(n,t)}{Y(n,t)}\frac{Y(n,t)}{t}$$

Taking limits gives $\lambda\pi(n) = \lambda(n)p(n)$. When $\lambda(n) = \lambda$, arrivals see time averages. In particular, when the arrival process is Poisson, $\lambda(n) = \lambda$.

Some students seem to think that if you use PASTA to prove that $\pi(n) = \delta(n)$ (which is a bit convoluted), that this $\pi(n) = \delta(n)$ then implies PASTA. This, however, is a simple logical failure: note that $p \implies q$ is not the same as $q \implies p$. (A dog is an animal with four legs, an animal with four legs is not necessarily a dog...)

s.0.2.120.

$$\mathsf{E}\left[W_q\right] = \frac{1+C_a^2}{2}\frac{\rho}{1-\rho}\mathsf{E}[S] = \frac{\rho}{1-\rho}\mathsf{E}[S],$$

since for the $M/M/1$ queue, $C_a^2 = 1$. Next, $\mathsf{E}[W] = \mathsf{E}\left[W_q\right] + \mathsf{E}[S]$. Thus, with Little's law,

$$\mathsf{E}[L] = \lambda\,\mathsf{E}[W] = \frac{\rho^2}{1-\rho} + \rho.$$

The result now follows.

Not using the PK-formula: no point.

s.0.2.121. The level-crossing equations are like this:

$$\begin{aligned}\lambda p(0) &= 2\mu p(2)\\ \lambda p(1) &= 2\mu p(2) + \mu p(3)\\ \lambda p(2) &= \mu p(3).\end{aligned}$$

Since $\lambda = \mu$, $p(0) = 2p(2)$, $p(2) = p(3)$. But then, from the second equation: $p(1) = 3p(2)$. Finally, since $\sum p(i) = 1$, we get that $p(2)(2+3+1+1) = 7$. Hence $p(2) = 1/7$.

s.0.2.122. First use PASTA to see that $\pi(n) = p(n)$. Then, conclude that $\pi(3) = p(3) = 1/7$.

You need to mention that you used PASTA. It's not evident that $p(n) = \pi(n)$.

s.0.2.123. Now,

$$\begin{aligned}p(2) &= \frac{2}{x}p(1) & p(3) &= xp(2) = \frac{x^2}{2}p(0)\\ p(1) &= \frac{2}{x}p(2) + \frac{1}{x}p(3) = p(0)(1+\frac{x}{2}).\end{aligned}$$

Hence

$$p(0)(1+1+\frac{x}{2}+\frac{x}{2}+\frac{x^2}{2}) = 1.$$

With this we have, for given x, found $p(0)$. Then we want x to be such that

$$p(3) = \frac{x^2/2}{2+x+x^2/2} > \alpha.$$

s.0.2.124. Yes, because we add an extra state that can be reached, while the transition rates between states $0,1,2$ and 3 do not change. Hence, since $p(4) > 0$, and $p(4) = p(3)$, the fraction of time spent in state 3 must be smaller in the presence of a state 4 than without this extra state.

In simple terms, if there is a waiting room with 3 seats, and you add an extra seat, than the fraction of people that have to stand becomes less.

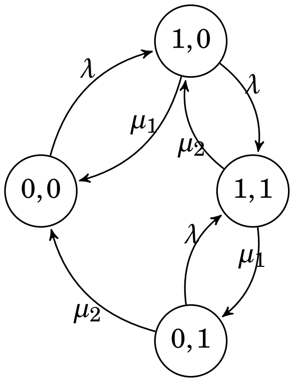
s.0.2.125. Jobs arrive at rate λ. The fraction accepted is $(1-\beta)$. Hence, the rate at which jobs enter is $\lambda(1-\beta)$. The average time jobs stay in the system is 1
mu. Thus, by Little's law, the average number of jobs in the system $\rho = \lambda(1-\beta)/\mu$.

s.0.2.126. The service times are not i.i.d., they depend on the number of jobs in the system, in other words, the service rate in state 2 is 2μ, while in state 1 it is 0 and in state 3 it is μ. It's simply not a $G/G/n/K$ queue, even though there is loss.

Interestingly, quite a number of students mess up the concepts of $G/G/1$ and so on. Note that the the $M/M/1$ queue is a special case of the $G/G/1$ queue, and this in turn is a special case of the $G/G/n$ queue. Check out the definitions in the book.

In the $G/G/n/K$ queue, it is not true that $\rho = \lambda/\mu$. This would imply with (0.2.7) that $\beta = 0$. But this ridiculous: the fraction of lost customers is not 0 in the $G/G/n/K$ queue in general, while (0.2.7) holds for any such queueing system.

s.0.2.127. $(0,0)$ both servers are empty, $(1,0)$ first server is busy, $(0,1)$ second server is busy, $(1,1)$ both servers are busy. As no jobs can be in queue, these are all possible states.



s.0.2.128.

There cannot be an arrow from $(1,1)$ to $(0,0)$, nor from $(0,0)$ to $(0,1)$ (Besides that this is not in line with the problem specification, why would you join the slow server if the fast is free?)

s.0.2.129. For ease, call $p(0,0) = a$, $p(1,0) = b$, $p(1,1) = c$, $p(0,1) = d$. Balance equations:

$$\begin{aligned}
\lambda a &= \mu_1 b + \mu_2 d \\
(\lambda + \mu_1) b &= \lambda a + \mu_2 c \\
(\mu_1 + \mu_2) c &= \lambda b + \lambda d \\
(\lambda + \mu_2) d &= \mu_1 c.
\end{aligned}$$

Thus, from (4) we have d as function of c. With (3) we get c as function of b. With (2) we get b as function of a. Then normalize to get a.

This is NOT a Jackson network...

s.0.2.130. Jobs arrive at rate λ. By PASTA $\lambda p(1,1)$ is the rate at which jobs are lost. Hence, the loss fraction is $p(1,1)$. Note, that $\lambda p(1,1) \neq p(1,1)$, but somehow, many students give the answer $\lambda p(1,1)$. Even by checking the units (number per unit time versus just number) it is apparent that $\lambda p(1,1)$ must be wrong. You need to mention that you used PASTA. It's not evident that it is $p(1,1)$.

s.0.2.131. If $\mu_2 = \mu_1$ and the queue can contain no job, then the above system reduces to the $M/M/2/2$ queue. It is not the $M/M/1$ or $M/M/2$ or $M/M/1/2$ queue.

s.0.2.132. Add a state (1) to the right of $(1,1)$. There is an arrow λ from $(1,1)$ to (1), and an arrow $\mu_1+\mu_2$ in the other direction.

Some students made two queues, one for each server. But why would you wait for a server when another server is free?

Also, this is not an $M/M/2/3$ queue, because the servers are not identical.

s.0.2.133. $p(1) = p(1,1)\lambda/(\mu_1+\mu_2)$.

s.0.2.134. $C_a^2 = 1$. $c = 2$. However, $C_s^2 \neq 1$ (you should realize this as soon as you read the problem.). As an approximation

$$\begin{aligned}\mathsf{E}[S] &= \frac{1}{3}\frac{1}{3} + \frac{2}{3}\frac{1}{6},\\ \mathsf{E}\left[S^2\right] &= \frac{1}{3}\frac{1}{3}^2 + \frac{2}{3}\frac{1}{6}^2.\end{aligned}$$

This gives $\mathsf{V}[S]$, and then C_s^2.

The real problem is more interesting than this. The fraction of orders served by the fast server should depend on λ. Think about this.

Some students think that $\mathsf{E}[S] = 1/(3+6)$.

s.0.2.135. See the section in the book.

Some students say that $A(m,n,t) \approx D(n,t)$. Others say that $\lambda G(n-m)p(m)$ is the rate at which state $n+1$ is entered from state m. Why is this wrong? Yet others just say that up-crossings are down-crossings, and leave it at that.

s.0.2.136. $B = 1$, hence $\mathsf{E}[B] = \mathsf{E}\left[B^2\right] = 1$. Substitute and simplify.

Some students show that $\mathsf{E}[L] = \sum_n np(n) = \rho/(1-\rho)$, but that is not the answer to the question.

s.0.2.137. Formulas of general model should reduce to formulas for special cases of the general model. If not, the formula for the general model is wrong.

Some students say that such checks prove that the general formula is correct. In other words, they reverse the logic. Others say that it is important to 'connect the systems'. This is of course not the answer to the question.

0.3 OLD EXAM QUESTIONS

0.3.1 *Multiple-choice Questions*

0.3.1 (201807)**.** *Sakasegawa's approximation for the waiting time in a G/G/c queue is*

$$\mathsf{E}[W_Q] = \frac{C_a^2 + C_s^2}{2} \frac{\rho^{\sqrt{2(c+1)}-1}}{c(1-\rho)} \mathsf{E}[S].$$

We claim that it is exact for the $M/M/1$ *queue.*

0.3.2 (201807)**.** *A production system consists of 2 stations in tandem. The first station has one machine, the second has two identical machines. Machines never fail and service times are deterministic. Jobs arrive at rate 1 per hour. The machine at first station has a service time of 45 minutes per job, a machine at the second station has a service time of 80 minutes. We claim that the second station is the bottleneck.*

0.3.3 (201904)**.** *A job's normal service time, without interruptions, is given by* S_0*. The durations of interruptions are given by the i.i.d. random variables* $\{R_i\}$ *and have common mean* $\mathsf{E}[R]$ *and variance* $\mathsf{V}[R]$*. If* N *interruptions occur, the effective service time will then be*

$$S = S_0 + \sum_{i=1}^{N} R_i.$$

Then all steps in the computation below are correct:

$$\mathsf{E}\left[\sum_{i=1}^{N} R_i\right] = \mathsf{E}\left[\sum_{n=0}^{\infty} \mathbb{1}_{N=n}\right] \mathsf{E}\left[\sum_{i=1}^{n} R_i\right] = \mathsf{E}[N]\,\mathsf{E}[R]$$

0.3.4 (201907)**.** *When a bus reaches the end of its line, it undergoes a series of inspections. The entire inspection takes 5 minutes on average, with a standard deviation of 2 minutes. Buses arrive with inter-arrival times uniformly distributed on* $[3,9]$ *minutes. We model the number of buses waiting to be inspected as an* $G/G/1$ *queue. The code below correctly computes the waiting time in queue.*

```
>>> a = 3.
>>> b = 9.
>>> EX = (b+a)/2. # expected inter-arrival time
>>> EX
6.0
>>> labda = 1./EX # per minute
>>> VA = (b-a)*(b-a)/12.
>>> CA2 = VA/(EX*EX)
>>> ES = 5.
>>> sigma = 2
>>> VS = sigma*sigma
>>> CS2 = VS/(ES*ES)
>>> rho = labda*ES
>>> Wq = (CA2+CS2)/2. * rho/(1.-rho) * ES
```

0.3.5 (201907)**.** *A station contains 3 identical machines in parallel. Jobs arrive as a Poisson process with rate* 3 *per hour. Service times are exponential with a mean duration of 1/2 hour. This queueing system cannot be modeled with level-crossing arguments, hence we need Sakasegawa's formula to compute the average time in queue.*

0.3.2 *Open Questions*

0.3.6 (201706)**.** *We have a machine that fails regularly, but we can control the way it fails. Option 1: Clean and tune the machine for precisely 15 minutes at the start of every hour. Option 2: Don't clean the machine, but wait until it fails and then do a repair. In the latter situation, the time between two failures is exponentially distributed with a mean of 6 hours. Repair times are i.i.d. and exponentially distributed with a mean of 1 hour. To simplify the analysis, you are allowed for Option 1 to model the time between the cleaning actions and the cleaning actions themselves as exponentially distributed.*

Jobs arrive as a Poisson process with rate 2 *per hour, regular service times are exponential with a mean of* 10 *minutes. Which of the two options would you prefer?*

A station contains 3 identical machines in parallel. Jobs arrive as a Poisson process with rate 3 per hour. Service times are exponential with a mean duration of 1/2 hour.

0.3.7 (201807)**.** *Is there a model by which we can obtain an exact answer for the average time the system contains n jobs?*

0.3.8 (201807)**.** *Use Sakasegawa's approximation to compute the average time a job spends in queue. (You do not have to use your calculator to compute the final answer;. I just want to see the numerical values of each component of the formula, you don't have to compute the end result.)*

0.3.9 (201807)**.** *What is the average time a job spends in the system?*

Some students add three times the average service time, but servers in parallel are not servers in tandem. . .

0.3.10 (201807)**.** *Suppose now that each machine can fail between any two jobs with constant probability $p = 1/3$. The repair time is constant, and takes 1 hour. What is the average time a job spends in queue? (Again, it suffices if you show the numerical answer for each component in the formula.)*

0.3.3 *simulation*

In this section, we will deal with the code below.

```python
from heapq import heappop, heappush
import numpy as np
from scipy.stats import expon

np.random.seed(3)

ARRIVAL = 0
DEPARTURE = 1

stack = [] # this is the event stack
queue = []
served_jobs = [] # used for statistics

num_jobs = 10
```

```
labda = 2.0
mu = 3.0
rho = labda / mu
F = expon(scale=1.0 / labda) # interarrival time distributon
G = expon(scale=1.0 / mu) # service time distributon


class Server:
 def __init__(self):
 self.busy = False


server = Server()


class Job:
 def __init__(self):
 self.arrival_time = 0
 self.service_time = 0
 self.departure_time = 0
 self.queue_length_at_arrival = 0

 def sojourn_time(self):
 pass # write your code below.


 def __repr__(self):
 return f"{self.arrival_time}, {self.service_time}, {self.departure_time}\n"


def start_service(time, job):
 server.busy = True
 job.departure_time = time + job.service_time
 heappush(stack, (job.departure_time, job, DEPARTURE))


def handle_arrival(time, job):
 job.queue_length_at_arrival = len(queue)
 if server.busy:
 heappush(queue, (job.arrival_time, job))
 else:
 start_service(time, job)


def handle_departure(time, job):
 server.busy = False
 if queue: # queue is not empty
```

```
62   time, next_job = heappop(queue)
63   start_service(time, next_job)
64
65
66  time = 0
67  for i in range(num_jobs):
68   job = Job()
69   time += F.rvs()
70   job.arrival_time = time
71   job.service_time = G.rvs()
72   heappush(stack, (job.arrival_time, job, ARRIVAL))
73
74
75  while stack:
76   time, job, typ = heappop(stack)
77   if typ == ARRIVAL:
78   handle_arrival(time, job)
79   else:
80   handle_departure(time, job)
81   served_jobs.append(job)
```

0.3.11 (201907)**.** *The above code simulates the G/G/1 queue. Complete the code to compute the sojourn time. (You can write your code at the correct place.)*

0.3.12 (201907)**.** *Explain the code from lines 45 to 48.*

0.3.13 (201907)**.** *Suppose we want to implement the LIFO queue. Which line of the above code has to change, and what should the change be? (It is ok if your answer is not perfect python code; however, your answer should show your correct understanding.)*

0.3.14 (201907)**.** *Suppose we want to implement the shortest-processing-time-first scheduling rule. Which line of the above code has to change, and what should the change be?*

0.3.15 (201907)**.** *What is the aim of test-driven coding?*

0.3.16 (201907)**.** *Why is an event stack an essential concept for the discrete-time simulation of complicated systems?*

Solutions

s.0.3.1. Answer = A.

s.0.3.2. Answer = B. The second station has a utilization of $80/(2*60) = 8/12 = 2/3$, while the first has a utilization of $45/60 = 3/4$, which is higher.

s.0.3.3. Answer = B, **3.4.6**

s.0.3.4. Answer = A, see **3.1.2**.

s.0.3.5. Answer = B, see **0.3.7**.

s.0.3.6. Note, the time between two failures is not the same as the time to failure.

Option 1. The availability is $A = 45/60 = 3/4$. Hence,

$$\mathsf{E}[S_e] = 10 \cdot 4/3 = 40/3 \approx 13 \text{ minutes}$$

We also have

$$C_e^2 = C_0^2 + 2A(1-A)\frac{m_r}{\mathsf{E}[S_0]} = 1 + 2\frac{3}{4}\frac{1}{4}\frac{15}{10}.$$

Now,

$$\rho = \lambda\,\mathsf{E}[S_e] = \frac{2}{60}\frac{40}{3}.$$

Now we can fill in

$$\mathsf{E}\left[W_Q\right] = \frac{1+C_e^2}{2}\frac{\rho}{1-\rho}\mathsf{E}[S_e].$$

Finally, $EW = \mathsf{E}\left[W_Q\right] + \mathsf{E}[S_e]$.

Option 2. The availability is $A = 50/60 = 5/6$. Hence,

$$\mathsf{E}[S_e] = 10 \cdot 6/5 = 12 \text{ minutes}$$

We also have

$$C_e^2 = C_0^2 + 2A(1-A)\frac{m_r}{\mathsf{E}[S_0]} = 1 + 2\frac{5}{6}\frac{1}{6}\frac{60}{10}.$$

and

$$\rho = \lambda\,\mathsf{E}[S_e] = \frac{2}{60}12$$

It is essential that you realize that both $\mathsf{E}[S]$ and C_e^2 are affected by failures. If you forgot to compensate in either of the two, I subtracted one point.

s.0.3.7. Yes, use the $M/M(n)/1$ queueing model, or the $M/M/3$ model (which is a subset of the $M/M(n)/1$ queue).

Note the $M/M/1$ queue with a fast server is not the same as the $M/M/3$ server. The service process is different.

s.0.3.8. $\rho = \lambda\,\mathsf{E}[S]/c = 3 * 0.5/3 = 0.5$.

$$\begin{aligned}\mathsf{E}\left[W_Q\right] &= \frac{C_a^2 + C_s^2}{2}\frac{\rho^{\sqrt{2(c+1)}-1}}{c(1-\rho)}\mathsf{E}[S] \\ &= \frac{1+1}{2}\frac{0.5^{\sqrt{7}}}{3*0.5} * 0.5 \\ &= \frac{0.5^{\sqrt{7}}}{3}.\end{aligned}$$

s.0.3.9. This is $\mathsf{E}\left[W_Q\right] + \mathsf{E}[S] = \mathsf{E}\left[W_Q\right] + 0.5$

s.0.3.10. The expected service time of a job now becomes

$$\mathsf{E}[S] = \mathsf{E}[S_0] + \mathsf{E}[T]/B = 0.5 + 1/3 = 5/6$$

Next, with the formula sheet:

$$\mathsf{V}[S] = 0.5 * 2 + \frac{1}{3} + \frac{3-1}{3^2}1 = 1.074074074074074.$$

With this,

$$C_s^{=}\frac{\mathsf{V}[S]}{(\mathsf{E}[S])^2} = \mathsf{V}[S]\frac{36}{25} = 1.3748148148148147.$$

Finally, $\rho = 3 * \mathsf{E}[S]/3 = 5/6$, and fill in Sakasegawa's formula with these values.

Mind that failures affect $\mathsf{E}[S]$ *and* C_s^2. If you forgot to compute C_s^2: -1/2.

Some students use the wrong failure model, the one with preemptive failures: -1/2.

s.0.3.11. This is the code:

```
def sojourn_time(self):
return self.departure_time - self.arrival_time
```

s.0.3.12. First we set a flag to indicate that the server is occupied. Then we compute the departure time. Finally, we push the job on the event stack such that it will popped by the simulator at its departure time.

s.0.3.13. `heappush(queue, (-job.arrival_time, job))`

s.0.3.14. `heappush(queue, (job.service_time, job))`

s.0.3.15. Write some small function. Test it on some known data. Once the function works, build a new function, test this, and so on. Then combine the tested functions to make something more difficult, test, etc.

s.0.3.16. The event stack keeps all events in the simulator ordered in time.

0.4 OLD EXAM QUESTIONS

0.4.1 *Multiple-choice Questions*

0.4.1 (201804)**.** *Consider a network with n stations in tandem. At station i, the service times S_i for all machines at that station are the same and constant; station i contains N_i machines. The number of jobs required to keep all machines busy is $N = \sum_{i=1}^n N_i$, and the raw processing time $T_0 = \sum_{i=1}^n S_i$. Thus, if the number w of allowed jobs in the system is larger than N, the number of jobs waiting somewhere in queue is $w - N$.*

0.4.2 (201804)**.** *Consider an M-station Jackson network with λ_i the total arrival rate of jobs at station i. The average waiting time jobs spend in the system is*

$$\mathsf{E}[W] = \sum_{i=1}^{M} \lambda_i \,\mathsf{E}[W_i]$$

0.4.3 (201804)**.** *In a Jackson network the time-average probability that station i contains n_i jobs (either in queue or in service) is given $(1-\rho_i)\rho_i^{n_i}$.*

0.4.4 (201804)**.** *For the G/G/1 we can approximate the SCV of the inter-departures by the formula $(1-\rho^2)C_a^2 + \rho^2 C_s^2$. For the M/M/1 queue this reduces to $C_d^2 = 1$.*

0.4.5 (201804)**.** *Suppose in a tandem network of G/G/c queues we can reduce C_s^2 of just one station by a factor 2. To improve the average waiting time in the entire chain, it is best to reduce $C_{s,1}^2$.*

0.4.6 (201804). *We have a two-station Jackson network. If the routing matrix is* $P = \begin{pmatrix} 0 & 1 \\ r & 0 \end{pmatrix}$ *the total arrival rate* λ_1 *at station 1 is finite only if* $r = 1$.

0.4.7 (201807). *A production network consists of 3 single-machine stations in tandem. The processing times constant and such that* $t_1 = 2$ *hours,* $t_2 = 3$ *hours and* $t_3 = 2$ *hours. We claim that the critical WIP* $W_0 = 7/3$*, the bottleneck capacity* $r_b = 1/3$ *and the raw processing time* $T_0 = 7$.

0.4.8 (201904). *We have a two-station single-server open network. Jobs enter the network at the first station with rate* γ. *A fraction* α *returns from station 1 to itself; the rest moves to station 2. At station 2, a fraction* β_2 *returns to station 2 again, a fraction* β_1 *goes to station 1. Then,*

$$P = \begin{pmatrix} \alpha & 1-\alpha \\ 1-\beta_1-\beta_2 & \beta_2 \end{pmatrix}.$$

0.4.9 (201904). *Jobs arrive at rate* λ *and are assembled into batches of size* B. *The average time a job waits until the batch is complete is* $\mathsf{E}[W] = \frac{B-1}{2\lambda}$.

0.4.10 (201907). *We have a queueing network with* M *exponential servers and Poisson arrival processes Let* P_{ij} *be the routing matrix. For stability it is necessary that* $\sum_{j=1}^{M} P_{ij} < 1$ *for at least one* i.

0.4.11 (201907). *We have two* $M/M/1$ *stations in tandem. The average queueing time for the network is given by*

$$\mathsf{E}[W_Q] = \frac{\rho_1}{1-\rho_1} + \frac{\rho_2}{1-\rho_2}. \tag{0.4.1}$$

0.4.12 (201907). *Consider a two-station Jackson network with* $P = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ *and* $\gamma = (1,0)$. *Let* $p(i,j)$ *be the stationary probability that the first (second) station contains* i (j) *jobs. Then*

$$\lambda p(0,0) = \mu_2 p(0,1).$$

0.4.2 *Open Questions*

We have a network with two single-server stations in tandem. Jobs arrive at the first station as a Poisson process with rate λ, the service times at stations 1 and 2 are i.i.d. and exponentially distributed with mean μ_i^{-1} for station i, $i = 1,2$. The entire network can contain at most one job, hence, when there is a job anywhere in the network, any new arrival is lost.

0.4.13 (201704). *Can the first station be characterized as an* $M/M/1/1$ *queue? why, or why not?*

0.4.14 (201704). *Make a sketch of the states and the transition rates.*

0.4.15 (201704). *What are the balance equations for this queueing network?*

0.4.16 (201704). *Find the stationary distribution of the number of jobs at the first and second station in terms of* λ, μ_1 *and* μ_2.

0.4.17 (201704). *Compute* $\mathsf{E}[L_1]$ *and* $\mathsf{E}[L_2]$.

0.4.18 (201704). *Compute* $\mathsf{V}[L_1]$.

0.4.19 (201704). *Suppose $\mu_1 = \mu_2 = \mu$. How much larger than λ should μ minimally be such that the loss probability is less than 5%?*

We have a network with two single-server stations in tandem. Jobs arrive at the first station as a Poisson process with rate λ, the service times at stations 1 and 2 are i.i.d. and exponentially distributed with mean μ_i^{-1} for station i, $i = 1,2$. The waiting room at the first station is unlimited; the second station can contain at most one job. When the server at the second station is occupied, the server at first station blocks in the sense that it does not start service when the second station is busy.

0.4.20 (201706). *The assumed blocking policy at the first station is equivalent to the* preemptive repeat with without resampling *discipline, which means that the interrupted customer starts again with the original service time. Why is this so?*

0.4.21 (201706). *Can the second station be characterized as an $M/M/1/1$ queue?*

0.4.22 (201706). *Make a sketch of the state space, the transitions and the transition rates.*

0.4.23 (201706). *What is the stability criterion for this queueing network?*

0.4.24 (201706). *Suppose you have some resources (money) available to increase the processing rate of* just *one of the servers or invest in queueing space between stations 1 and 2.. Which of these options should you suggest to analyze first?*

Henceforth, assume that also the first station cannot contain more than one job.

0.4.25 (201706). *Find the stationary joint distribution of the number of jobs at the first and second station in terms of λ, μ_1 and μ_2.*

0.4.26 (201706). *What is the fraction of time the network is empty?*

0.4.27 (201706). *What is the throughput of this queueing network, i.e., the departure rate?*

0.4.28 (201706). *What is throughput of this queueing system in the limit $\mu_2 \to \infty$?*

Solutions

s.0.4.1. Answer = B.

In general the number of jobs in queue is much higher than $w - N$. Consider the example $S_1 = 10$ and $N_1 = 10$, and $S_2 = 1, N_2 = 20$, and $n = 2$. Clearly, 19 machines at station 2 are always empty.

s.0.4.2. Answer = B. A simple reason is that the units left and right don't match: left time, right number per time times time.

s.0.4.3. Answer = A.

s.0.4.4. Answer = A.

s.0.4.5. Answer = A.

s.0.4.6. Answer = B.

When $r = 1$, the matrix P has an eigenvalue 1. Hence, the equation $\lambda = \gamma + \lambda P$ has no inverse in this case.

s.0.4.7. Answer = A.

s.0.4.8. Answer = B, **6.3.11**

s.0.4.9. Answer = A, **3.2.1**

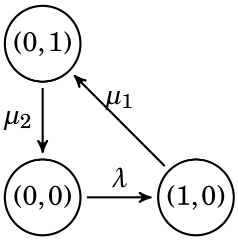
s.0.4.10. Answer = A, **??**.

s.0.4.11. Answer = B, see **3.5.6**. It is evidently wrong: the units don't check.

s.0.4.12. Answer = A.

s.0.4.13. No, if the server at the first station is free but the second is occupied, the first server still has to reject any arriving job. This is not the case for the $M/M/1/1$ queue.

s.0.4.14.



s.0.4.15.

$$\mu_1 p(1,0) = \lambda p(0,0)$$
$$\mu_2 p(0,1) = \mu_1 p(1,0) = \lambda p(0,0).$$

s.0.4.16. Define $\rho_i = \lambda/\mu_i$.

$$p(1,0) = \rho_1 p(0,0)$$
$$p(0,1) = \rho_2 p(0,0).$$

With the normalization requirement $p(0,0) + p(1,0) + p(0,1) = 1$ we get $p(0,0)(1 + \rho_1 + \rho_2) = 1$, hence

$$p(0,0) = \frac{1}{1 + \rho_1 + \rho_2}.$$

s.0.4.17.

$$\mathsf{E}[L_1] = 0(p(0,0) + p(0,1)) + 1p(1,0) = \frac{\rho_1}{1 + \rho_1 + \rho_2}$$

and

$$\mathsf{E}[L_2] = \frac{\rho_2}{1 + \rho_1 + \rho_2}$$

s.0.4.18.

$$\mathsf{V}[L_1] = \mathsf{E}\left[L_1^2\right] - (\mathsf{E}[L_1])^2 = \frac{\rho_1}{1 + \rho_1 + \rho_2} - \left(\frac{\rho_1}{1 + \rho_1 + \rho_2}\right)^2.$$

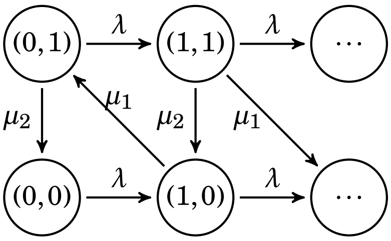
s.0.4.19. The acceptance probability is $p(0,0)$. Let $\rho = \lambda/\mu$. Then

$$0.95 = \frac{19}{20} \le \frac{1}{1+2\rho} \iff 2\rho \le 1/19 \iff \rho \le 1/38. \iff \lambda < \frac{\mu}{38}.$$

s.0.4.20. Because the service times are exponentially distributed, hence memoryless.

s.0.4.21. No, due to blocking, the departure process of the first station depends on the state of the next server. Hence, the how jobs arrive at the second station depends on the state of the second server. For the $M/M/1/1$ the arrival process does not depend on the state of the server; only the process of accepting jobs depends on the state of the server.

s.0.4.22.



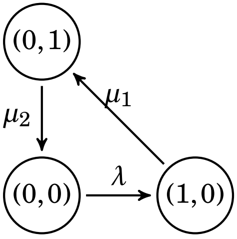
s.0.4.23. For step to the right, i.e., arrival, there must be a service at the first and second station. Hence $\lambda(\mathsf{E}[S_1] + \mathsf{E}[S_2]) < 1$.

The wording of the actual exam was wrong; I included the solution in the question... For this reason I removed the question. However, I still gave a point for a reasonable answer.

s.0.4.24. Due to blocking the first server and second server are hardly occupied. If we have unlimited buffer space between the two stations, the stability condition is $\lambda \min\{ES_1, \mathsf{E}[S_2]\} < 1$. Thus, if possible, expanding the queueing space is an easy solution.

However, any reasonable answer would do here.

s.0.4.25. The state space plus transitions becomes like this now:



Define $\rho_i = \lambda/\mu_i$. Write $p = p(0,0), q = p(1,0), r = p(0,1), s = p(1,1)$. Then,

$$\begin{aligned}\lambda p &= \mu_2 r \\ \lambda r &= \mu_2 s, \\ \mu_1 q &= \lambda p + \mu_2 s.\end{aligned}$$

We have four unknowns and three equations. With normalization we have four equations, so the above should suffice. Expression everything in terms of p:

$$\begin{aligned}&r = \rho_2 p \\ &s = \rho_2 r = \rho_2^2 p, \\ &q = \rho_1 p + \mu_2 s/\mu_1 = \rho_1 p + \rho_1 \rho_2 p = p\rho_1(1+\rho_2).\end{aligned}$$

The normalization condition gives

$$p(1 + \rho_1(1+\rho_2) + \rho_2 + \rho_2^2) = 1.$$

s.0.4.26. This is p.

s.0.4.27. Jobs arrive at rate λ. Only the jobs that arrive when the first server is free are accepted. Hence $\lambda(p+r)$.

s.0.4.28. In the limit $\mu_2 \to \infty$, there is no job at the second station. Thus, in this case, $r = s = 0$. Moreover

$$1 + \rho_1(1+\rho_2) + \rho_2 + \rho_2^2 \to 1 + \rho_1.$$

Hence $p = (1+\rho_1)^{-1}$. And this is what we get for a single server station with blocking.

It is actually interesting to also consider the limit $\mu_1 \to \infty$, rather than taking the limit $\mu_2 \to \infty$. Why are the answers so different?

FORMULA SHEET

$$\rho = \lambda \frac{\mathsf{E}[S]}{c}$$

$$\mathsf{E}\left[W_Q\right] = \frac{C_a^2 + C_s^2}{2} \frac{\rho^{\sqrt{2(c+1)}-1}}{c(1-\rho)} \mathsf{E}[S]$$

Batching: $C_{sB}^2 = \dfrac{B\,\mathsf{V}[S_0] + \mathsf{V}[T]}{(B\,\mathsf{E}[S_0] + \mathsf{E}[T])^2}$

Nonpreemptive: $\mathsf{V}[S] = \mathsf{V}[S_0] + \dfrac{\mathsf{V}[T]}{B} + \dfrac{B-1}{B^2}(\mathsf{E}[T])^2$

Preemptive: $A = \dfrac{m_f}{m_r + m_f}, C_s^2 = C_0^2 + 2A(1-A)\dfrac{m_r}{\mathsf{E}[S_0]}$

$$C_{di}^2 = 1 + (1-\rho_i^2)(C_{ai}^2 - 1) + \frac{\rho_i^2}{\sqrt{c_i}}(C_{si}^2 - 1)$$

$$f_i(n_i) = \frac{(c_i\rho_i)^{n_i}}{n_i!G(i)} \mathbb{1}_{n_i < c_i} + \frac{c_i^{c_i}\rho_i^{n_i}}{c_i!G(i)} \mathbb{1}_{n_i \geq c_i},$$

$$\text{with } G(i) = \sum_{n=0}^{c_i-1} \frac{(c_i\rho_i)^n}{n!} + \frac{(c_i\rho_i)^{c_i}}{c_i!} \frac{1}{1-\rho_i}$$

$$\mathsf{E}[L_i] = \frac{(c_i\rho_i)^{c_i}}{c_i!G(i)} \frac{\rho_i}{(1-\rho_i)^2} + c_i\rho_i$$