IEOR 4615: Service Engineering, Professor Whitt

SOLUTIONS to the Midterm Exam, March 31, 2015.

Please explain your reasoning; show your work.

1. Average Performance of a Hospital Emergency Room (25 points)

A hospital emergency room (ER) is organized so that all patients register through an initial check-in process. At his/her turn, each patient first registers, then is seen by an ER doctor and then exits the ER, either followed immediately by admission to the hospital or departing altogether. Currently, 40 people per hour arrive at the ER, 20% of whom are admitted to the hospital after seeing an ER doctor. On average, 20 people are waiting to be registered and 40 are registered and waiting to see an ER doctor. The registration process takes, on average, 3 minutes per patient. Among patients who ultimately are admitted to the hospital the average time spent with a doctor is 40 minutes. Among those not admitted to the hospital, the average time is 5 minutes.

(a) On average, how long does a patient stay in the ER?

The idea is to apply Little's law, $L = \lambda W$. This is a minor variant of Question 8 on Homework 2, which was discussed in Lecture 3. The first step is to observe that the system can be partitioned into 4 subsystems: (i) queue for registration, (ii) registration, (iii) queue for doctors and (iv) doctors. That is, the patient is either waiting for one of these two service processes or is being served by one of the two service processes. We thus can focus on the three components of Little's law: L_i, λ_i and W_i for $1 \le i \le 4$. First, all patients pass through all four subsystems in order. Thus,

$$\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 40 \quad \text{per hour} \quad = 40/60 = 2/3 \quad \text{per minute.}$$

We are told that $L_1 = 20$, $L_3 = 40$ and $W_2 = 3$ minutes. And we can easily calculate W_4, getting $W_4 = 0.2(40) + 0.8(5) = 12$ minutes. Hence, we can calculate the four missing terms, one for each of the four subsystems:

$$\begin{aligned} W_1 &= L_1/\lambda_1 = 20/(2/3) = 30 \quad \text{minutes} \\ L_2 &= \lambda_2 W_2 = (2/3) \times 3 = 2 \\ W_3 &= L_3/\lambda_3 = 40/(2/3) = 60 \quad \text{minutes} \\ L_4 &= \lambda_4 W_4 = (2/3) \times 12 = 8 \end{aligned}$$

Now we can answer all the questions. the answer to the first question is

$$W = W_1 + W_2 + W_3 + W_4 = 30 + 3 + 60 + 12 = 105 \quad \text{minutes.}$$

(b) On average, how many patients are being examined by doctors at any one time?

By above,

$$L_4 = \lambda_4 W_4 = (2/3) \times 12 = 8$$

(c) On average, how many patients are in the ER?

$$L = L_1 + L_2 + L_3 + L_4 = 20 + 2 + 40 + 8 = 70$$

or

$$L = \lambda W = 2/3 \times 105 = W = 70$$

from (a).

---

2. Staffing the Emergency Room (25 points)

Again consider the emergency Room (ER) introduced in Problem 1, with the parameters specified there, but for this problem we focus only on the doctors, because the doctors are reasonably judged to be the critical resource. In this problem we want to estimate how many doctors we need in the emergency room.

(a) We start with a stationary model, which has a constant patient arrival rate over each day, as specified in problem 1. We assume that each patient is seen by a single doctor. For this constant arrival rate model, what is the offered load for the doctors?

Here we draw on Lecture 6. The offered load is the expected number of busy doctors if there were an unlimited supply of doctors. We assume that the doctors are the only constraint, so we look at the subsystem 4 above containing only the doctors. For the stationary model, we are considering the $M/GI/\infty$ infinite-server queue, but the Poisson property of the arrival process is not actually required. Thus, **the offered load is simply the arrival rate multiplied by the expected time in system (with a doctor) per patient**. That is, the offered load for the doctors is precisely

$$L_4 = \lambda_4 \times W_4 = (2/3) \times 12 = 8,$$

as given for problem 1 (c). This represents the average number of busy doctors in the existing system, under the stated assumptions.

(b) In the setting of part (a), what is an approximate number of doctors required at each time to provide reasonable performance in the ER? (Explain your assumptions and reasoning.)

We assume that the doctors are the critical resource. We assume that there is enough capacity in the other three subsystems. We assume that the patients are each seen one at a time by a single doctor. By part (a), we need more than 8 doctors, on average, for the system to be stable (rate in to be strictly less than the maximum rate out). We need some extra slack to get reasonable performance. As discussed in Lecture 6, see slide 8, we can use the square root staffing formula. Let m be the offered load, with $m = 8$ here. Then the staffing should be approximately

$$s = m + \beta\sqrt{m} = 8 + \beta\sqrt{8} \approx 8 + 3 = 11,$$

where we have let $\beta = 1$ and we round up by approximating $\sqrt{8} \approx 3$. (We know that $2 = \sqrt{4} < \sqrt{8} < \sqrt{9} = 3$.)

If we assume Poisson arrival process, then the stochastic offered load is approximately Poisson with variance equal to the mean. (That is based on the $M/GI/\infty$ infinite-server model.) We

might use a QoS parameter of $\beta = 1$. That is based on $P(N(0,1) > 1) \approx 0.16$. Reasonable values of β range from about 0.5 to 2.0, with 2 designed to produce good quality of service or to meet unexpected surges of demand. That indicates that reasonable staffing levels might be up to 14 doctors. (exactly 8 would yield $\rho = 1$ and thus be unstable, while $9 - 10$ would probably lead to too much congestion. Numbers in the range $11 - 14$ would probably be reasonable.)

The biggest problem with this analysis is that the arrival rate may well not be constant over the day. If that is the case, then we would want the staffing to vary by the time of day also. Hence, the next part.

(c) Suppose that the arrival rate to the emergency room varies by time of day, as it does in practice. In particular, for simplicity, assume that, each day, the arrival rate is 10 per hour in the interval $[0, 8]$ (between midnight and 8 am), 60 per hour in the interval $[8, 16]$ (between 8 am and 4 pm) and 50 per hour in the interval $[16, 24] = [16, 0]$ (between 4 pm and midnight). Assume that the distribution of the length of time each patient spends with a doctor does not change (provided that there are adequately many doctors). What are the approximate offered loads at (i) 4 am and at (ii) 12 noon, and what are consistent approximately appropriate time-varying staffing levels for these two times?

Note that the two times are in the middle of two of the three 8-hour time periods. Since the average time a patient spends with a doctor is about $W_4 = 12$ minutes, and should be less than an hour even for patients admitted to the hospital, we can fairly judge that the system is stationary with the specified arrival rate at each of these times, in the middle of an 8-hour interval. (The situation is more complicated after a rate change.) Since $10 = (1/4) \times 40$ and $60 = (3/2) \times 40$, the offered loads are

$$m(4) = (1/4) \times 8 = 2 \quad \text{at 4 am} \quad \text{and} \quad m(12) = (3/2) \times 8 = 12 \quad \text{at 12 noon},$$

(adjusting the result from part (a)). If we use the square-root-staffing formula, then we would have associated staffing levels at the times 4 am and 12 (noon) as

$$s(4) = m(4) + \sqrt{m(4)} = 2 + 1.414 = 4 \quad \text{doctors, rounding up to the nearest integer}$$

and

$$s(12) = m(12) + \sqrt{m(12)} = 12 + 3.46 = 16 \quad \text{doctors, rounding up to the nearest integer.}$$

The values of the square root rounded up to the next integer are easy to compute without any calculating aids. For example, we can find the next integer greater than $\sqrt{12}$ by observing that the next perfect square greater than 12 is 16, whose square root we know.

(d) How is the general time-varying offered load defined for this system with the specified time-varying arrival rate (under the assumptions of part (c))?

The general time-varying offered load, denoted by $m(t)$, is defined as the time-varying expected number of busy servers in the $M_t/GI/\infty$ infinite-server model with the specified time-varying arrival rate function. That is all that is required. You are not being asked for a formula, but instead for the definition.

Assuming that the system has been operating continuously in the past in this mode (which is a reasonable approximation), the formula for the offered load is

$$m(t) = \int_{-\infty}^{t} \lambda(s) G^c(t-s)\, ds,$$

where $G^c(x) \equiv 1 - G(x)$ with $G(x) \equiv P(S \le x)$ being the cdf of a service time, denoted by S. In this case, we could not actually calculate $m(t)$ because we do not know G, but we do know $E[S]$, in particular $E[S] = W_4 = 12$ minutes from problem 1. We would get a reasonable rough idea if we assumed that G were exponential with that mean. But we do not yet know enough to carry out the calculation. So you could not give a number.

(e) Suppose that we wanted to staff in a way to achieve roughly a constant level of performance for all times in the day. Consider the three times: (i) $0:05 = 12:05$ am (right after midnight), (ii) $8:05$ am (right after 8 am), and (iii) $16:05 = 4.05$ pm (right after 4 pm). How should the staffing levels at these three times compare to achieve our goal of stable performance? That is, how should the staffing levels at these three times be ranked (ordered); i.e., when should the staffing level be highest, next highest and then lowest? Justify your answers.

Since patients will be in the system a random time period after their arrival time, there is a time lag in performance after the arrival rate. (See slides 19-21 of Lecture 6.) Since the three mentioned times are less than 1 expected service time after the time of the rate change, the offered load at these times will be closer to the nearly constant value appropriate toward the end of the previous interval than the offered load later in the new interval. Thus, we can anticipate that

$$m(16:05) > m(0:05) > m(8:05)$$

even though

$$\lambda(8:05) = 60 > \lambda(16:05) = 50 > \lambda(0:05) = 10.$$

Since we would staff using the square-root-staffing formula with the offered load $m(t)$, we should have the staffing levels ordered by

$$s(16:05) > s(0:05) > s(8:05).$$

3. Using Data to Estimate the Expected Time Spent in the ER (25 points)

Suppose that you have collected data on the performance of the emergency room (ER) in problem 1 above. You observe the time each successive arriving patient spends in the ER for all arrivals in a common two-hour period on each of 10 Mondays. Let $X_{i,j}$ be the time spent in the ER on day i by patient j, with j indexing the order of arrival within the interval on day i. Let n_i be the number of patients observed on day i. (The ten numbers n_i all fall in the interval $[50, 120]$.) You compute the following statistics:

$$\begin{aligned} \bar{X}_i &\equiv \frac{1}{n_i}\sum_{j=1}^{n_i} X_{i,j} \quad \text{and} \quad \bar{S}_i^2 \equiv \sum_{j=1}^{n_i}(X_{i,j} - \bar{X}_i)^2, \quad 1 \le i \le 10, \\ \bar{X} &\equiv \frac{1}{10}\sum_{j=1}^{10} \bar{X}_i \quad \text{and} \quad \bar{S}^2 \equiv \sum_{j=1}^{10}(\bar{X}_i - \bar{X})^2. \end{aligned}$$

(a) Is it appropriate to assume, for each fixed i, that the n_i random variables $X_{i,j}$, $1 \le j \le n_i$, are independent random variables? Why or why not?

No, we should anticipate that these times will be positively correlated, because successive customers are likely to experience a similar level of congestion. If the number of arrivals is larger than usual on one day, then many patients arriving on that day are likely to spend longer in the ER. If the ER has an unusually high number of serious cases on one day, many patients arriving on that day are likely to spend longer in the ER. On a light day, many patients are likely to spend less time. Similarly, in a time of exceptionally high congestion, successive arrivals are likely all to have high delays.

(b) What is an appropriate estimate of the expected time for a patient to spend in the ER during this 2-hour period on a Monday based on the statistics above?

The obvious estimator is simply the sample mean of the daily means, $\bar{X}$ above.

(c) How is your answer to part (b) affected by your answer to part (a)? Explain.

It is unaffected. Since the expected value of a sum of random variables is always the sum of the expected values of the individual random variables, the answer to part (b) is unaffected by any independence assumptions.

(d) Explain how you would use the statistics above to estimate a 95% two-sided confidence interval for the expected value estimated in part (b). Give a representation of the estimated confidence interval that is as explicit as possible (given the information above).

The estimated confidence interval of the "true" mean is $[\bar{X} - \bar{H}, \bar{X} + \bar{H}]$, where $\bar{H}$ is the estimated half-width. As indicated in part (e), we assume that $\bar{X}_i$ are 10 i.i.d. random variables with a normal distribution having unknown mean μ and variance σ^2. The standard estimator of the variance is

$$\hat{\sigma}^2 = \frac{\bar{S}^2}{n-1} = \frac{\bar{S}^2}{9}$$

where here $n = 10$. (The division is *not* given in the specification above, so we need to include it here.) The key theory states that

$$T \equiv \frac{\bar{X} - \mu}{\sqrt{\hat{\sigma}^2/10}} = \frac{\bar{X} - \mu}{\sqrt{\bar{S}^2/90}}$$

has the Student-t distribution with $n - 1 = 9$ degrees of freedom. Thus, the halfwidth of the confidence interval is

$$H = \frac{t_{0.025,9}\sqrt{\bar{S}^2}}{\sqrt{90}}.$$

You are not expected to know that $t_{0.025,9} = 2.262$; it is easily looked up in a table, but you should know that you should use the Student-t distribution and, specifically $t_{0.025,9}$, where 9 is the "degrees of freedom," because the sample size is 10. The critical value $t_{0.025,9}$ has the obvious meaning that $P(T < -t_{0.025,9}) = P(T > t_{0.025,9}) = 0.025$, so we get a two-sided 95% confidence interval.

(e) What assumptions justify your answer in part (d)?

We assume that the daily averages $\bar{X}_i$, $1 \leq i \leq 10$, are approximately 10 independent and identically distributed random variables, each normally distributed with an unknown mean μ and variance σ^2. The normality is justified as an approximation for each day by the central limit theorem for dependent random variables. On each day we have a sample mean of approximately $40 \times 2 = 80$ observations, which we note in part (a) should be dependent. Since the sample size 10 is quite small, we need to use the Student-t distribution.

4. A Small Medical Clinic (25 points)

Consider a small clinic run by a single doctor. Patients arrive at this clinic according to a Poisson process at rate $\lambda = 1.6$ per hour. The expected time the doctor spends with each patient is 30 minutes.

(a) If the times patients spend with the doctor are i.i.d. exponential random variables, what is the steady-state distribution of the number of patients at the clinic?

The assumptions make this the $M/M/1/\infty$ single-server birth-and-death queue. The steady-state number in the system, say Q, has a **geometric distribution:**

$$P(Q = k) = (1 - \rho)\rho^k, \quad k \geq 0,$$

where $\rho \equiv \lambda/\mu = 1.6/2.0 = 0.8$, as can be seen from slides 10 and 13 of Lecture 5. This can easily be derived from the BD formulas on the CTMC formula sheet. In that context, simply observe that $r_j = \rho^j$, where $\rho \equiv \lambda/\mu$. The steady-state distribution is then seen to be the familiar geometric distribution.

(b) What is the expected time each patient spends at the clinic?

The expected value of that geometric distribution in part (a) is

$$E[Q] = \rho/(1 - \rho).$$

(The explicit formula is given on the basic probability formula sheet.) The mean time spent in system, say $E[W]$, can be found by Little's law:

$$E[W] = E[Q]/\lambda = E[S]/(1 - \rho) = 30/(1 - 0.8) = 150\,\text{minutes} = 2.5\,\text{hours}$$

(c) Suppose, instead, that the clinic serves two kinds of patients: Type 1 patients have mean service times of 60 minutes, whereas Type 2 patients have mean service times of 15 minutes. Each arrival is type 1 with probability 1/3. What is the mean and squared coefficient of variation (scv, variance divided by the square of the mean) of the service time of an arbitrary patient?

This is like Question 3 (e) on Homework 4 part 2. Let S be the service time. Then

$$E[S] = (1/3)60 + (2/3)15 = 20 + 10 = 30\,\text{minutes}$$

just as before. To get the scv, we use the fact that the variance of an exponential is the square of the mean, and the second moment of a mixture is the mixture of the second moments:

$$E[S^2] = (1/3)(2 \times (60)^2) + (2/3)(2 \times (15)^2) = 2400 + 300 = 2700 \,\text{minutes}$$

Hence the variance is $Var(S) = E[S^2] - (E[S])^2 = 2700 - (30)^2 = 1800$, and the scv is

$$c_S^2 \equiv \frac{Var(S)}{E[S]^2} = \frac{1800}{900} = 2.0$$

(d) Use a heavy-traffic approximation to estimate how much the answer in part (b) changes under the assumption of part (c).

By the heavy-traffic approximation, the mean waiting time is approximately proportional to $(c_a^2 + c_s^2) = (1 + c_S^2)$. That sum of variability parameters increased from $(1 + 1) = 2$ to $(1 + 2) = 3$. Thus, the mean steady-state waiting time should be multiplied by a factor of 1.5, i.e. a 50% increase. See slides 8 and 9 of Lecture 11.

However, in this case the exact value can be computed. The model under part (c) is the $M/GI/1/\infty$ model for which the mean waiting time in queue is known to have the classic Pollaczek-Khintchine formula (which you are not supposed to know). Let W_q be the waiting time in queue and let W be the overall waiting time. Then

$$E[W_q] = \frac{\rho E[S](1 + c_S^2)}{2(1 - \rho)}$$

while

$$E[W] = E[S] + E[W_q] = E[S]\left(1 + \frac{\rho(1 + c_S^2)}{2(1 - \rho)}\right).$$

Hence, actually $E[W]$ goes from 150 minutes or 2.5 hours in part (b) to 210 minutes or 3.5 hours. So the exact ratio is 1.40. Clearly, for practical engineering purposes, 1.50 is a good practical approximation for 1.40, which can be found immediately, without applying any explicit formulas or doing any calculations.

(e) Suppose now that the clinic admits patients for 8 hours per day, starting at 8 am, but no new patients are admitted after the closing time of 4 pm. However, patients in the clinic at 4 pm will be seen by the doctor after 4 pm. The doctor also takes a lunch break from noon until 1 pm. Finally, now the arrival rate of patients between 8 am and 4 pm is 3.0 per hour, instead of the previously assumed 1.6 per hour. Use a fluid approximation to estimate when the doctor can leave the clinic.

Following the hint, do a simple deterministic fluid analysis. According to the fluid analysis, the number in the system grows at rate $3.0 - 2.0 = 1.0$ per hour from 8 am until 12 noon. At noon, there should be roughly 4 patients in the system. During the one hour break, patients arrive at rate 3 per hour. Hence there should be approximately 7 patients waiting at 1 pm. But after 1, the patients still arrive faster than they can be served. In the next three hours, until the closing time at 4 pm, the net input rate is again 1 per hour Hence there should be approximately 10 patients in the clinic at 4 pm. At that time there will be no arrivals. The doctor serves these remaining

patients at 2 per hour. So it takes the doctor 5 hours to see all the patients. And the randomness actually makes matters worse, as you could see by performing a detailed stochastic simulation. But the quick analysis can be done on the back of an envelope to quickly understand that the situation is untenable.

(f) How would the answer in part (e) change if a second doctor were hired to work part time in the clinic, so that the schedule before 1 pm is unchanged, but two doctors both work in the clinic after 1 pm (until all patients have been seen)?

Do a modification of the simple deterministic fluid analysis in part (e). If there were two doctors after 1 pm, then the maximum service rate would increase from 2 to 4. Hence, from 1 to 4, there is a net output rate of 1. Thus, by 4 pm, the number of patients left in the clinic should be approximately $7 - 3 = 4$. The two doctors together would see patients at rate 4 per hour. So the remaining time the doctors must remain is $4/4 = 1$ hour. The two doctors could leave at approximately $5:00$ pm. A little more reasonable. But maybe not good enough.
