**IEOR 3106: Professor Whitt**

**Topics for Discussion, Thursday, October 24, 2013**

# Infinite-Server Queues and NHPP's

We discussed the $M_t/GI/\infty$ infinite-server queueing model, with a nonhomogeneous Poisson arrival process (the $M_t$, $M$ for Markov and subscript $t$ for time-varying rate) and i.i.d. service times with a general cdf (the $GI$). Infinitely many servers means that each arrival enters service (and remains in the system) for a random length of time after the arrival.

This model is discussed in the 1993 "Physics" paper. About 2-3 pages of reading there are required. However, the model is also discussed in the class textbook. See Example 5.18 on p. 327 for a discussion of the $M/G/\infty$ infinite-server queue (with constant arrival rate). See §5.4.1 for a discussion of the NHPP. See Example 5.25 on p. 344 for a discussion of the departure (or output) process of an infinite-server queue. See Exercise 5.94 for a problem on a Poisson process on the plane. See Theorem 1 of the physics paper for more on the $M_t/G/\infty$ queueing model.

**1. Planning a Special Exhibit at an Art Museum**

We anticipate that arrivals will come to the museum randomly over time according to a **nonhomogeneous Poisson process (NHPP)** over the time interval $[0, 8]$, corresponding to an 8-hour day. (See §5.4.1 of the book.) We close the doors for new arrivals at time $t = 8$, but we let the visitors remain for up to two hours more. We anticipate that the arrival rate will start low and rise to a peak in the middle of the day and then drop down again toward the end of the day. Specifically, we estimate that the arrival rate function will be approximately quadratic, according to the function

$$\lambda(t) = 10(8t - t^2), 0 \leq t \leq 8. \tag{1}$$

For simplicity, we assume that the visitors arrive one at a time. From a quick look, we see that $\lambda(0) = \lambda(8) = 0$ and it peaks at time 4, where $\lambda(4) = 160$. Let $A(t)$ count the number of arrivals in $[0, t]$. The arrival process $\{A(t) : t \geq 0\}$ is a counting process, specifically an NHPP.

We assume that successive visitors will each stay a random time, distributed as a random variable $S$, with these times being independent and identically distributed (i.i.d.) with mean and variance

$$E[S] = 1 \quad \text{and} \quad Var(S) = 3. \tag{2}$$

We plan to allow all these customers to enter the museum. Except possibly at the end of the day, visitors can stay as long as they want. We want to estimate how many people will be in the museum at each time. We plan to stop admitting new arrivals at time 8, but we will let customers remain until time 10.

Let $N(t)$ be the number of people in the museum at time $t$. Let $m(t) = E[N(t)]$ be the mean. We ask several questions:

1. What is the distribution of the total number of visitors to come to the museum on one day (according to the model)?

2. At what time $t$ is the arrival rate $\lambda(t)$ highest?

3. What is $m(t) = EN(t)$, the mean number of visitors in the museum as a function of time?
4. What is distribution of the number of visitors in the museum as a function of time?
5. At what time $t$ is the mean $m(t)$ highest?
6. Approximately when would the mean $m(t)$ drop to 0?
7. Is there a time $t$ such that $P(Q(t) > 130) > 0.1$? (Assume that $Var(S_e) = 6$; see §3.)

### 1. Events Happening "At Random:" the Poisson distribution

The arrival process is an NHPP, so it can be viewed as a Poisson random measure (or Poisson process) on the real line $\mathbb{R}$ with time-varying intensity function. In analyzing the associated stochastic process $\{N(t) : t \geq 0\}$, we will use the fact that the arrivals together with their service times can be viewed as a random measure (or Poisson process) on the plane $\mathbb{R}^2$ with an associated intensity function on the plane. That is discussed in the proof of Theorem 1 of the physics paper.

There are two key properties of an NHPP and a Poisson random measure:

1. The number of points in each subset has a Poisson distribution, with the mean being the integral of the intensity function over the set.
2. The numbers of points in $k$ disjoint subsets are $k$ mutually independent random variables.

A Poisson distribution naturally models "events occurring at random." That is naturally modeled directly by Bernoulli random variables. We assume that the space can be divided into small regions, with at most one point in each region, and with a point occurring in each region with probability $p$. Thus if there are $n$ regions in the space, then the total number of points has a binomial distribution with mean $np$. If we let $n$ get large and $p$ get small, so that $np = \lambda$, then the binomial distribution approaches the Poisson distribution. That explains the Poisson approximation for the binomial distribution; See p. 30 of the textbook. This approximation applies directly where the intensity function can be regarded as constant or approximately constant.

### 2. The Tail Integral Formula for the Mean

A convenient formula for computing the mean of a nonnegative random variable $X$ is the tail integral formula,

$$E[X] = \int_0^\infty P(X > t)\, dt$$

A derivation is given on the bottom of page 610 in the textbook. It involves constructing a two-dimensional integral and changing the order of integration.

### 3. The Stationary-Excess Random Variable $S_e$

Given a nonnegative random variable $S$, the associated stationary-excess random variable $S_e$ has cdf

$$P(S_e \leq x) = \frac{1}{E[S]} \int_0^x P(S > u)\, du$$

We use the tail integral formula to see that this is a bonafide cdf. It turns out that the moments of $S_e$ can be expressed simply in terms of the moments of $S$ by

$$E[S_e^k] = \frac{E[S^{k+1}]}{(k+1)E[S]} \quad \text{for} \quad k \geq 1,$$

so that

$$E[S_e] = \frac{E[S^2]}{(2)E[S]} = \frac{(c_S^2+1)E[S]}{2} \quad \text{for} \quad k \geq 1,$$

where $c_S^2 = Var(S)/E[S]^2$ is the squared coefficient of variation.

It is not difficult to see that $S_e$ is distributed the same as $S$ if $S$ has an exponential cdf. That turns out to be the only case.

As an aside, we mention that the stationary-excess distribution plays an important role in renewal theory in Chapter 7. See Examples 7.16, 7.17, 7.23 and 7.24. We will be considering it again later.

In our example, the mean and variance of the random variable have been given above. Thus $E[S^2] = 4$ and $E[S_e] = 2$. From the formula for $E[S_e]$, we see that $E[S_e] > E[S]$ whenever $c_S^2 > 1$.

**4. The Infinite-Server Queue**

The number of visitors in the museum at time $t$ can be represented by the number of busy servers in an infinite-server queue, specifically in the $M_t/G/\infty$ model, in which the $M_t$ denotes an NHPP arrival process with time-varying arrival rate function, specifically given in (1) above, and i.i.d. service times distributed as $S$. Since there are infinitely many servers, each arrival enters service immediately upon arrival, and thus remains in the system for a random time distributed as $S$. If $S$ is not constant, this allows arrivals to depart in a different order than they arrive.

Key facts: The distribution of $N(t)$ is Poisson for each $t$, with mean given by

$$m(t) = \int_0^\infty [\int_s^\infty \lambda(t-s)g(x)\,dx]\,ds = \int_0^\infty \lambda(t-s)G^c(s)\,ds, \tag{3}$$

where $G^c(t) \equiv P(S > t)$ or $G^c(t) = 1 - G(t)$, where $G(t) \equiv P(S \leq t$. From above it follows that we can express the mean as

$$m(t) = \int_0^\infty \lambda(t-s)E[S]g_e(s)\,ds = E[S]\int_0^\infty \lambda(t-s)g_e(s)\,ds = E[\lambda(t-S_e)]E[S], \tag{4}$$

where $S_e$ has the stationary-excess distribution as specified above, with $G_e$ and $g_e$ being the cdf and pdf of $S_e$, respectively. (Just observe that $G^c(s) = E[S]g_e(s)$ by the definition of $S_e$ and $g_e$.) See Theorem 1 of the 1993 physics paper.

Explanation: If we put points in the plane at locations $(t, x)$ where $t$ is an arrival time of a visitor and $x$ is the time that visitor remains in the museum, then the points are distributed as a Poisson random measure on the plane ($\mathbb{R}^2$) or, equivalently, a Poisson process on the plane with the intensity of a point occurring at $(t, x)$ being $\lambda(t)g_S(x)$ where $g_S(x)$ is the probability density function (pdf) of the random variable $S$. That occurs because the arrival process is a NHPP and the number of points in a rectangle in the plane can be constructed by independent thinning of the arrivals in the appropriate subinterval of the real line.

## 5. The Quadratic Approximation

If we made the simplifying assumption that the arrival rate function is approximately quadratic, of the form $\lambda(t) = a + bt - ct^2$, as in (1), throughout all time, then we have a simple approximation formula, given in Theorem 9 and (14) of the physics paper, in particular,

$$m(t) \approx \lambda(t - E[S_e])E[S] - cVar(S_e)E[S]. \tag{5}$$

Such a quadratic function can arise by taking a Taylor series approximation of the arrival rate function. (This approximation does not make sense for all time, because the quadratic arrival rate function with a positive maximum necessarily assumes negative values in the past and in the future. We are thus assuming that the past where the arrival rate function is negative does not influence the answer much at the time of interest.)

This approximation formula (5) should be compared to the exact formula at the end of (4) above. Note that the random variable $S_E$ appears inside the nonlinear function $\lambda$ in (4), so that we have an expectation of a nonlinear function of a random variable, but instead we have the mean $E[S_e]$ appearing inside the deterministic function $\lambda$ in (5).

Thus, $m(t) \approx \lambda(t)E[S]$ except for a deterministic time shift and a deterministic space shift. First notice that, if $\lambda(t) = \lambda$, a constant, then $m(t) = m = \lambda E[S]$ is the long-run mean.

In the example above, the time shift is by $E[S_e] = 2$. We have not yet specified $Var(S_e)$, which depends on the first three moments of $S$, but suppose that $Var(S_e) = 6$, as in the questions. Then, for our example,

$$m(t) \approx 10(8(t-2) - (t-2)^2) - 60. \tag{6}$$

The peak of $m(t)$ in (6) occurs at time $t^* = 6$. The peak value of the mean is approximately

$$m(t^*) \approx 10(32 - 16) - 60 = 160 - 60 = 100.$$

Since the distribution of $N(t)$ is Poisson, it is approximately normal, and since the variance of a Poisson random variable equals its mean, we have

$$P(N(t^*) > 130) \approx P\left(\frac{N(t^*) - m(t^*)}{\sqrt{m(t^*)}} > \frac{130 - m(t^*)}{\sqrt{m(t^*)}}\right) \approx P(N(0,1) > 3).$$

## 6. Answers to the Questions

1. What is the distribution of the total number of visitors to come to the museum on one day (according to the model)?

---

For each $t$, the number of arrivals, $A(t)$, has a Poisson distribution with mean

$$E[A(t)] = \int_0^t \lambda(s)\, ds, \quad t \geq 0.$$

Hence,

$$E[A(8)] = \int_0^8 10(8t - t^2)\, dt = 10(8(32) - 512/3) = 853.33$$

---

2. At what time $t$ is the arrival rate $\lambda(t)$ highest?

---

Set the derivative equal to 0, i.e., solve $\dot{\lambda}(t) = 0$. Get $t^* = 4$. The arrival rate function is symmetric, with $\lambda(0) = \lambda(8) = 0$.

---

3. What is $m(t) = EN(t)$, the mean number of visitors in the museum as a function of time?

---

Here we use the infinite-server formula (3) above.

$$m(t) = \int_0^\infty \lambda(t-s)G^c(s)\,ds \tag{7}$$

We do not calculate here, because we did not actually specify the cdf $G$. However, we can approximate by using the quadratic approximation. That produces (6) above.

---

4. What is distribution of the number of visitors in the museum as a function of time?

---

The distribution is Poisson for each $t$:

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

where $m(t)$ is the mean, discussed in the previous question.

---

5. At what time $t$ is the mean $m(t)$ highest?

---

We use the approximation in (6). It tells us that there is a time lag of the peak in the arrival rate to the peak of $m(t)$ of $E[S_e] = 2$. So the time of the peak in $m(t)$ is 6.

---

6. Approximately when would the mean $m(t)$ drop to 0?

---

Apply (6). Set

$$m(t) \approx 10(8(t-2) - (t-2)^2) - 60 = 0 \tag{8}$$

so that $t = x + 2$, where

$$8x - x^2 - 6 = 0,$$

so that we have the solution of a quadratic equation, $x = 4 + \sqrt{10} \approx 7.16$ and $m(t) \approx 0$ at time $t = 9.16$. Hence closing at $t = 10$ seems reasonable.

---

7. Is there a time $t$ such that $P(N(t) > 130) > 0.1$?

———————————————————————————————-

Look at the time $t^*$ that makes $m(t)$ largest. At $t^* = 6$, $m(t^*) = 100$ (done above in the quadratic approximation. Use a normal approximation for the Poisson:

$$P(N(t) > 130) > 0.1 \approx P(N(0,1) > 3) \approx 0.0013 < 0.1$$

So the answer is no. This result is useful to know that, under the assumptions, the number in the museum is unlikely ever to exceed 130.

———————————————————————————————-

## 7. Extra Observations (first three mentioned in class)

1. **not a Poisson process**

   It is important to point out and emphasize that $\{N(t) : t \geq 0\}$ is not a Poisson process or an NHPP. These processes are counting processes, which have nondecreasing sample paths. In addition, a Poisson process has independent increments. Neither is true for the process $\{N(t) : t \geq 0\}$ here.

2. **starting in the infinite past**

   It is important to point out and emphasize that the simple formulas depend on starting in the infinite past. Starting at time 0 is covered as the special case in which we set $\lambda(t) = 0$ for $t < 0$. But the formulas get more complicated.

3. **the departure process and rate**

   Theorem 1 in the physics paper includes a description of the departure process and the departure rate. The departure process is an NHPP. It is easy to see that it has independent increments and that each increment has a Poisson distribution. Note that the departure rate has a formula closely related to the mean $m(t) \equiv E[Q(t)]$.

4. **ODE with M service**

   Theorem 6 and Corollary 4 show that the mean $m(t)$ satisfies an ordinary differential equation (ODE) when the service-time distribution is exponential. That reveals how the peaks of $m$ and $\lambda$ are related. In particular, that explains why the curve for $m(t)$ crosses the curve for $\lambda(t)E[S]$ where the derivative $\dot{m}(t) = 0$, e.g., where $m(t)$ assumes its maximum.

5. **relaxation time: approach to steady state**

   For a stationary model, it is important to understand how the system approaches steady state as time evolves, starting with various typical special initial conditions, such as starting empty. A very simple revealing formula exists for the $M_t/GI/\infty$ model; formula (20).

6. **the covariance**

   Theorem 2 describes $Cov(Q(t), Q(t + u))$. The idea is to exploit the random measure representation and the picture, here Figure 3. We see that $Q(t) = X+Y$, while $Q(t+u) = Y + Z$, where $X$, $Y$ and $Z$ are independent. Hence

   $$Cov(Q(t), Q(t + u)) = Cov(X + Y, Y + Z) = Cov(Y, Y) = Var(Y) = E[Y],$$

where $E[Y]$ has a simple integral formula, like the mean $m(t) = E[Q(t)]$. (We use the facts that (i) the covariance function is linear in both its arguments, so that $Cov(X + Y, Y + Z) = Cov(X, Y) + Cov(X, Z) + Cov(Y, Y) + Cov(Y, Z)$ and (ii) $Cov(X, Y) = 0$ if $X$ and $Y$ are independent.

7. **sinusoidal and other periodic arrival rates**

   The sine paper describes results for periodic arrival rate functions. The key fact is that $m$ inherits the sinusoidal structure from $\lambda$. Hence revealing formulas are available.

8. **staffing: the 1996 paper**

   Application of this model to set staffing levels in service systems (which are not themselves modeled as infinite-server queues) is discussed the 1996 staffing paper. The infinite-server (IS) approximation, or offered-load approximation, we have been discussing is contrasted with the pointwise-stationary approximation (PSA) and the simple stationary approximation (SSA) there, for the case of a sinusoidal arrival-rate function. We noted that an explicit formula for $m(t)$ when $\lambda(t)$ is sinusoidal is given in the 1993 sine paper. The function $m(t)$ is also sinusoidal with the same frequency, but there is a time lag and space shift there too. An important concept and method is the modified offered load (MOL) approximation (not discussed in class).