Lecture 27: Inspection Paradox and the M/G/1 queue

Announcements:

1. Congratulations to the CMU Programming Team, including Thomas Quisel from our class, for advancing to the International World Finals. We wish you the best of luck in the ACM-ICPC World Finals in Shanghai.

2. Last quick-quiz is Tuesday.

Contents

1  The Inspection Paradox  2

2  The M/G/1 queue and its analysis  2

3  Renewal Reward theory  6

4  Application of Renewal Reward theory to derive E[Excess]  9

5  Back to the Inspection Paradox  11

6  Back to the M/G/1 queue  13
1 The Inspection Paradox

I would like to motivate this lecture by asking a question. We will come back to this question repeatedly throughout the lecture.

**Question:** Suppose buses arrive at a bus stop every 10 minutes on average and the time between arrivals at the bus stop is *Exponentially*-distributed. I arrive at the bus stop at a random time. How long can I expect to wait for a bus?

![Figure 1: The Inspection Paradox.](image)

**Question:** While you’re thinking about the answer to the first question, I’d like you to also ask yourself whether your answer changes if I change the distribution of the time between buses (the mean time between buses is still 10 minutes).

While you’re staring at the picture of the time between buses, let me introduce a definition.

**Definition 1** Let $A$ denote the time between bus arrivals. Let’s suppose a person arrives at a random time. Then the time that person has to wait until the next bus is denoted by the random variable $A_e$ and is called the excess of $A$.

2 The M/G/1 queue and its analysis

An M/G/1 queue consists of a single server and queue with Poisson job arrivals, where the size or service time of a job has a *general* distribution. That is, the job service
time, denoted by the random variable $S$, may take any distribution, where $E[S] = 1/\mu$. First-come-first-serve (FCFS) service order is assumed unless otherwise stated.

![Figure 2: An M/G/1 queue](image-url)

Because the job size distribution is not exponential, it is not clear how we can immediately use a Markov chain model to analyze this system. One approach is to try to approximate the general distribution by a mixture of exponential distributions, and then create a Markov chain model of that system. This is an approach that you will get a glimpse of in your next homework, and is explored in much more depth in a graduate-level class.

In this lecture we will learn a different approach, called the “tagged-job” technique. In the “tagged-job” method, we “tag” an arbitrary arrival and reason about the time it spends in the queue.

We will need the following notation:

- $T_Q$: Time in queue
- $N_Q$: Number in queue
- $N_Q^A$: Number in queue seen by arrival. Recall that $E[N_Q^A] = E[N_Q]$ by PASTA.
- $S$: Service time of a job, where $E[S] = 1/\mu$
- $S_i$: Service time of the $i$th job in the queue
- $S_e = \text{EXCESS of } S$: Remaining service time of job in service as seen by a random arrival into the system, given that there is some job in service at that time.
We now have,

\[
E[T_Q] = E[\text{Unfinished work that an arrival witnesses in system}]
\]

\[
= E \left[ \sum_{i=1}^{N_Q^A} S_i \right] + E[\text{Remaining time for job (if any) at server}]
\]

\[
= E[N_Q^A] \cdot E[S] + P_r\{\text{Arrival sees a job in service}\} \cdot E[S_e]
\]

\[
= E[N_Q] \cdot E[S] + (\text{Time-average probability there is a job in service}) \cdot E[S_e]
\]

\[
= E[N_Q] \cdot E[S] + \rho \cdot E[S_e]
\]

\[
= E[T_Q] \cdot \lambda \cdot E[S] + \rho \cdot E[S_e]
\]

\[
= E[T_Q] \cdot \rho + \rho \cdot E[S_e]
\]

\[
E[T_Q](1 - \rho) = \rho \cdot E[S_e]
\]

\[
E[T_Q] = \frac{\rho}{1 - \rho} \cdot E[S_e]
\]

We have thus easily obtained a formula for the mean time in queue of an M/G/1 system, provided we can compute \( E[S_e] \) where \( S_e \) is the excess of the service time \( S \).

\[
E[T_Q] = \frac{\rho}{1 - \rho} \cdot E[S_e]
\]

We will soon derive a general formula for \( E[S_e] \), but first it is instructive to go through some examples:
Examples

- **Example 1: M/M/1 Queue**
  The service time, $S$, is exponentially distributed with rate $\mu$ and mean $1/\mu$. Since the service time distribution is memoryless, $E[Excess] = 1/\mu$. Therefore,

  \[
  E[T_Q] = \frac{\rho}{1 - \rho} \cdot \frac{1}{\mu}
  \]

  This agrees with our previous results concerning the M/M/1 queue.

- **Example 2: M/D/1 Queue**
  The service time is deterministic (constant) and equal to $1/\mu$. $E[Excess] = \frac{1}{2\mu}$, since there’s an equal probability of an arrival at any time during a job’s service interval. Therefore,

  \[
  E[T_Q] = \frac{\rho}{1 - \rho} \cdot \frac{1}{2\mu}
  \]

  Note that expected time in queue is half that of the M/M/1 queue. Why??

- **Example 3: M/\(E_k\)/1 Queue**
  The service time has an Erlang-$k$ distribution, $E_k$. The $E_k$ distribution consists of $k$ stages in series, each with exponential service time with mean $1/k\mu$. To compute $E[Excess]$ for $E_k$, we argue that if there is a job in service at the time of an arrival, then it is equally likely that the job in service is at each of the $k$ stages. On average, the job in service will be at the middle stage (assume $k$ is odd), leaving $\frac{k+1}{2}$ stages left to be completed. We therefore should have

  \[
  E[Excess] = \frac{k+1}{2} \cdot \frac{1}{k\mu},
  \]

  a quantity which decreases as $k$ increases (this makes sense because the distribution tends to a deterministic distribution as $k$ increases). Mean time in queue is then given by

  \[
  E[T_Q] = \frac{\rho}{1 - \rho} \cdot \frac{k+1}{k} \cdot \frac{1}{2\mu}
  \]

  Observe that for $k = 1$ this is equal to the $M/M/1$ expression and for $k \to \infty$ this is equal to the $M/D/1$ expression.

- **Example 4: M/\(H_2\)/1 Queue**
  Imagine that the service time has a hyperexponential distribution. That is, the service time is one exponential distribution with probability $p$ and another with probability $1 - p$. A particular hyperexponential distribution is shown below:
In your homework, you will get a feeling for why a distribution of this form can represent a wide range of variability, depending on the parameter $p$. You should be able to express $E[\text{excess}]$ and $E[T_Q]$ for this hyperexponential service distribution. Please try it, and come see me if you’re stuck! (Note, it’s a little trickier, because if given that there is a job in service, that job must be in the lower branch of the hyperexponential.)

So far we’ve been lucky that we could argue what $E[\text{excess}]$ should be in each case. In order to compute $E[S_e]$ exactly, for any distribution $S$, we need to use the Renewal-Reward theorem, which we review in the next section.

### 3 Renewal Reward theory

Renewal-Reward theory is a powerful technique which will allow us to obtain time-averages of many quantities, by considering only the average over a single renewal cycle. This will allow us to compute the time-average excess, which is also the excess seen by a random observer.

**Definition 2** *Any process for which the time between events are i.i.d. RVs with a distribution $F$ is called a renewal process.*

An example of a renewal process is shown in Figure 4.

**Question:** Give an example of a renewal process.

**Answer:** Consider a Markov chain. Imagine an “event” occurring every time the MC returns to state $i$. The times between visits to state $i$ are i.i.d.
Figure 4: A Renewal process. $X_i \sim F$, for all $i$.

Let $N(t)$ denote the number of events by time $t$. Then, we have the following theorem:

**Theorem 1** For a renewal process, if $E[X]$ is the mean time between renewals, we have, with probability 1:

$$\lim_{t \to \infty} \frac{N(t)}{t} \to \frac{1}{E[X]}$$  \hspace{1cm} (1)

**Proof:** The basic idea in this proof is to apply the Strong Law of Large Numbers (SLLN). Let $S_n$ be the time of the $n^{th}$ event. Then, we have

$$S_{N(t)} \leq \frac{N(t)}{t} \leq \frac{S_{N(t)+1}}{N(t)} < S_{N(t)+1}$$

But,

$$\frac{S_{N(t)}}{N(t)} = \frac{\sum_{i=1}^{N(t)} X_i}{N(t)} \to E[X] \text{ as } t \to \infty, \text{ w.p.1, by SLLN}$$

Likewise,

$$\frac{S_{N(t)+1}}{N(t)} = \frac{S_{N(t)+1} N(t)+1}{N(t) N(t)} \to E[X] \text{ as } t \to \infty, \text{ w.p.1, by SLLN}$$
So, we get
\[
\frac{t}{N(t)} \to E[X], \text{ w.p.1}
\]

\[
\Rightarrow \frac{N(t)}{t} \to \frac{1}{E[X]} \text{ as } t \to \infty, \text{ w.p.1}
\]

(Here w.p.1 is shorthand for “with probability 1.”)

Now consider a renewal process \( \{N(t), t \geq 0\} \) having interarrival times \( X_n, n \geq 1 \) with distribution \( F \), and suppose that each time a renewal occurs we receive a reward. We denote by \( R_n \) the reward earned at the time of the \( n^{th} \) renewal. We shall assume that the \( R_n, n \geq 1 \), are i.i.d. However, we do allow for the possibility that the \( R_n \) may (and often will) depend on \( X_n \), the length of the \( n^{th} \) renewal interval. If we let

\[
R(t) = \sum_{n=1}^{N(t)} R_n,
\]

then \( R(t) \) represents the total reward earned by time \( t \). Let

\[
E[R] = E[R_n]
\]

\[
E[X] = E[X_n]
\]

**Theorem 2 (Renewal-Reward)** If \( E[R] < \infty \) and \( E[X] < \infty \), then with probability 1,

\[
\frac{R(t)}{t} \to \frac{E[R]}{E[X]} \text{ as } t \to \infty
\]

**Question:** Interpret the above theorem.

**Answer:** The Renewal-Reward theorem says that the average rate at which we earn reward is equal to the expected reward earned during a cycle, divided by the expected cycle length. This should make a lot of sense intuitively, since every cycle is probabilistically identical. The result is non-trivial, however, since normally it’s meaningless to just divide two expectations.
Proof:

\[ R(t) \frac{t}{t} = \sum_{n=1}^{N(t)} \frac{R_n}{N(t)} \frac{N(t)}{t} \]

By the strong law of large numbers, we have

\[ \frac{\sum_{n=1}^{N(t)} R_n}{N(t)} \rightarrow E[R] \text{ as } t \rightarrow \infty, \text{ w.p.1} \]

By Theorem 1,

\[ \frac{N(t)}{t} \rightarrow \frac{1}{E[X]} \text{ as } t \rightarrow \infty, \text{ w.p.1} \]

Hence the result. \[ \blacksquare \]

The above theorem says that the time average rate at which we earn reward is equal to the expected reward earned during a cycle divided by the average cycle length.

4 Application of Renewal Reward theory to derive \( E[Excess] \)

Consider a renewal process consisting of a sequence of service times, each an instance of the r.v. \( S \), as shown below:

\begin{center}
\begin{tabular}{c|c|c|c}
 & S & S & S
\end{tabular}
\end{center}

Figure 5: Renewal occurs at the end of each service time.

Here the server is assumed to be always busy, and we’re imagining a renewal occurring at the end of each service.

Now let’s imagine that we arrive at some random time \( t \) during which the server is busy. Let \( E(t) \) represent the excess service time at time \( t \), i.e., the time until the job in service completes, given that we witness the job in service at time \( t \).
An arriving job (by PASTA) witnesses the time-average excess, i.e. the average excess over all time.

Here is a picture of the function $E(t)$:

![Diagram of $E(t)$](image)

Figure 6: The function $E(t)$, representing the excess service time at time $t$.

**Question:** We need an expression for the time-average excess. How do we express this?

**Answer:**

$$\text{Time-average Excess} = \lim_{s \to \infty} \frac{\int_0^s E(t)dt}{s}$$

Our goal is to compute this time-average excess, since this is the average excess as seen by a random arrival. To compute the time-average excess, we need to phrase this as a long-run average award. Let $R(s)$ denote the “reward” earned by time $s$.

**Question:** What is $R(s)$ for our problem?

**Answer:**

$$R(s) = \int_0^s E(t)dt$$

**Question:** So what is the time-average reward?

**Answer:**
Time-average Reward \( = \lim_{s \to \infty} \frac{R(s)}{s} = \lim_{s \to \infty} \frac{\int_0^s E(t)dt}{s} \) = Time-avg Excess

Now, by renewal-reward theory, the time-average reward is equal to the reward earned during one cycle divided by the expected length of one cycle.

**Question:** What is a “cycle”?

**Answer:** A cycle is one service time.

\[
\text{Reward earned during a cycle} = \int_0^S (S-t)dt \\
= S^2 - \frac{S^2}{2} \\
= \frac{S^2}{2}
\]

\[
\mathbb{E}\{\text{Reward earned during a cycle}\} = \frac{\mathbb{E}[S^2]}{2}
\]

\[
\mathbb{E}\{\text{Cycle length}\} = \mathbb{E}[S]
\]

So, average value of excess \( = \frac{\mathbb{E}[S^2]}{2\mathbb{E}[S]} \)

5 Back to the Inspection Paradox

Let’s return to our buses question now.
**Question:** Suppose buses arrive at a bus stop every 10 minutes on average and the time between arrivals at the bus stop is Exponentially-distributed. I arrive at the bus stop at a random time. How long can I expect to wait for a bus?

**Answer:** You can expect to wait 10 minutes. If the time between bus arrivals had a deterministic (constant) distribution, then you could expect to wait 5 minutes. If the distribution of the time between bus arrivals had a more variable distribution than the exponential, you would expect to wait longer than 10 minutes.

Let r.v. $S$ denote the time between arrivals of buses. Then the average time until the next bus is just the average excess, $E[S_e]$, namely

$$\text{Time-average Excess} = \frac{E[S^2]}{2E[S]}$$

Observe that if $A$ has a deterministic (constant) distribution, then the above quantity is just $E[S]/2$. However if $S$ has an exponential distribution, then the above quantity is equal to $E[S]$. If $S$ has very high variability, as in the Pareto distribution, then the expected excess will be much higher than $E[S]$.

Thus far we have been interested in the “time until the next bus arrives.” We could also have asked about the time since the last bus arrived.

**Question:** What would you guess is the distribution of the time since the last bus arrived?

**Answer:** This is called the “age of $S$,” and has the same distribution as the “excess of $S$”, namely $S_e$. You can prove this using the same type of renewal-reward argument. Please try this and come see me if you get stuck!

Adding the expected age and the expected excess, we see that a random arrival is likely to land in an interval where the time between buses is

$$\text{Time between buses} = \frac{E[S^2]}{2E[S]} + \frac{E[S^2]}{2E[S]} = \frac{E[S^2]}{E[S]}$$

If the time between buses is highly variable, this quantity can be way higher than $E[S]$. 
**Question:** Explain intuitively why a random arrival is likely to experience a bigger-than-average time between buses.

**Answer:** If you think about the renewal process represented by the interarrival times between buses, you will see that some renewals are short and some are long. It more likely that a random arrival lands in a long interval as compared with a short interval.

### 6 Back to the $M/G/1$ queue

![M/G/1 queue diagram](image)

Figure 7: An $M/G/1$ queue

Recall that when analyzing the $M/G/1$/FCFS queue (Figure 7) we proved that:

$$E[T_Q] = \frac{\rho}{1-\rho}E[S_e], \quad (2)$$

where $E[S_e]$ was defined as the expected remaining service time on the job in service at the time of an arrival, given that there is a job in service.

Observe that by PASTA, $E[excess]$ as seen by an arrival is equivalent to the (time) average remaining service time for the job in service, given that there is a job in service.

From the previous section, we know that

$$E[S_e] = \frac{E[S^2]}{2E[S]} \quad (3)$$
Sustituting (3) in (2), we get the final result

\[ E(T_Q) = \frac{\rho}{1 - \rho} \frac{E[S^2]}{2E[S]} \]  

(4)

Equation (4) is commonly referred to as the Pollaczek-Khinchin formula or the P-K formula.

Another equivalent way of writing the P-K formula is:

\[ E[T_Q] = \frac{\lambda E[S^2]}{2(1 - \rho)} \]

**Definition 3** The squared coefficient of variation, \( C^2 \), of a random variable \( S \) is defined as

\[ C^2_S = \frac{\text{var}(S)}{E[S]^2} \]

This can be thought of as a “normalized” variance.

If we observe that:

\[
E[\text{Excess}] = \frac{E[S^2]}{2E[S]} = \frac{E[S]}{2} \cdot \frac{E[S^2]}{(E[S])^2} = \frac{E[S]}{2} \cdot \frac{\text{Var}(S) + (E[S])^2}{(E[S])^2} = \frac{E[S]}{2} \cdot (C^2 + 1)
\]

we see that yet another equivalent way of writing the P-K formula is:

\[ E[T_Q] = \frac{\rho}{1 - \rho} \cdot \frac{C^2 + 1}{2} \cdot E[S] \]

**Question:** Why does \( C^2 \) play such a key role in determining delay?

**Answer:** What causes delays is “bunching up of jobs”

- For the D/D/1 queue, with arrival rate \( \lambda \) and service rate \( \mu > \lambda \), there are no delays.
• For the M/D/1 queue, delays occur because arrivals sometimes “bunch up”.

• For the M/M/1 queue, “bunching up” is also created by occasional long service times. Thus the expected delay in M/M/1 is greater than that in M/D/1.

So what creates delays is the occasional extra-long service time or extra-large number of arrivals in some service period. That is, delay is proportional to the variance in the arrival and service processes. Of course once one packet is delayed, that affects all packets in the queue behind it.

**Very Important Observation:** Expected waiting time in an M/G/1 queue can be huge, even under very low utilization \( \rho \), if \( C^2 \) is huge.

Thank you!