Lecture 5: Beyond Little's Law: $\overline{H} = \lambda \overline{G}$ & RCL

1 Announcements

1. Midterm: March 27, 2007, in class.
2. In case you need help with the homework problems from lecture, I will be available Monday morning: 11 a.m.- 12 noon, in Wean 8119, but you can also send email: harchol@cs.cmu.edu, to set up a meeting, or just drop by any time the door is open.

2 Organization for next 5-6 classes

Mor will teach the next 5-6 classes. These lectures will focus on *queueing analysis*, primarily involving the M/GI/1 queue and its many variants, where by variants we mean situations where the server may go on vacation, or jobs may be prioritized, or there may be a startup cost for the first job starting a busy period, etc.

We will be starting where we left off in 15-857. Unlike 15-857 where we focused on mean response time as our metric, our emphasis now will be on *higher moments* of response time. Specifically we will derive the Laplace transform of all the quantities that we study.

We will assign you a couple *homework problems* during each class, due at start of next class. We will then go over the solutions to the homework problems at the start of the next class. Problems are based directly on class material. Assuming that you have reviewed the class material carefully, the problems should not take too long.

3 Review of notation

We will be using the following notation, summarized in Table 1 to help you keep it straight. (Note that sometimes there is more than one way of representing the same thing. Sorry about that ...)

T_S or just T	Time in System or Sojourn time or Response Time
T_Q or D	Time in Queue or Delay
S or X	Service time or Job size
N_S or just N	Number in system
N_Q	Number in queue
V	Work in system
λ	Average arrival rate $= \lim_{t\to\infty} \frac{A(t)}{t}$, where $A(t) = \#$ arrivals by t
X	Throughput $= \lim_{t\to\infty} \frac{C(t)}{t}$, where $C(t) = \#$ completions by t
ρ	Average load $= \lambda E[S]$. Always assume $\rho < 1$ and further that our system is ergodic.

Table 1: Table of Notation

4 Review of Little's Law: $\overline{N} = \lambda \overline{T}$

We start by quickly reviewing Little's Law.

Theorem 1 *Given any queueing system. Let*

$$\lambda = \lim_{t\to\infty} \frac{A(t)}{t} \qquad \text{and} \qquad X = \lim_{t\to\infty} \frac{C(t)}{t}$$

where $A(t)$ is the number of arrivals by time t and $C(t)$ is the number of system completions (departures) by time t. Assume that $\lambda = X$ and that the limits: $\overline{N}_S^{TimeAvg}$, $\overline{T}_S^{TimeAvg}$, λ and X all exist. Then

$$\begin{aligned} \lim_{t\to\infty} \frac{1}{t} \int_0^{\infty} N(s)ds &= \lambda \cdot \lim_{t\to\infty} \frac{1}{A(t)} \sum_{i=1}^{A(t)} T_i \\ (\overline{N}_S^{TimeAvg} &= \lambda \overline{T}_S^{TimeAvg}) \end{aligned}$$

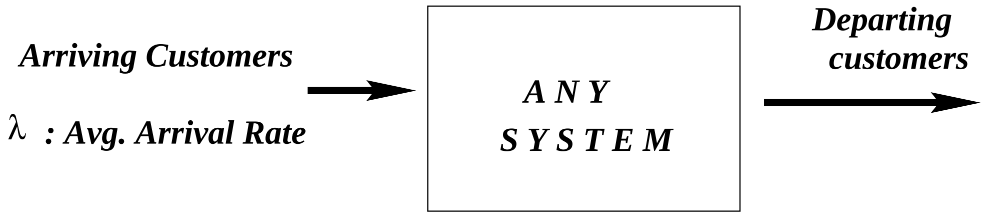

Figure 1: Setup for Little's Theorem

It is important to notice that Little's theorem makes no assumptions about the arrival process, the service time distributions at the servers, the network topology, the service order, nothing!

Proof: Let T_i denote the "time that the ith arrival to the system spends in the system," as shown in Figure 2:

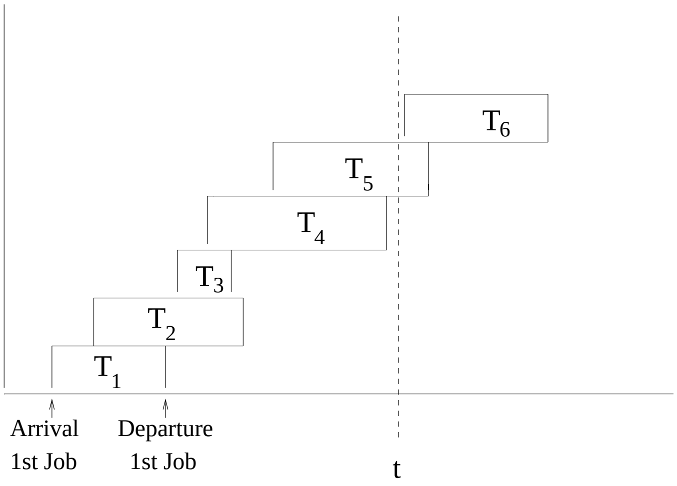

Figure 2: *Graph of Arrivals in Open System*

Then, for any time t,

$$\sum_{i=1}^{C(t)} T_i \leq \int_0^t N(s)ds \leq \sum_{i=1}^{A(t)} T_i$$

$$\frac{\sum_{i=1}^{C(t)} T_i}{t} \leq \frac{\int_0^t N(s)ds}{t} \leq \frac{\sum_{i=1}^{A(t)} T_i}{t}$$

$$\frac{\sum_{i=1}^{C(t)} T_i}{C(t)} \cdot \frac{C(t)}{t} \leq \frac{\int_0^t N(s)ds}{t} \leq \frac{\sum_{i=1}^{A(t)} T_i}{A(t)} \cdot \frac{A(t)}{t}$$

Taking limits as $t \to \infty$:

$$\lim_{t\to\infty} \frac{\sum_{i=1}^{C(t)} T_i}{C(t)} \cdot \lim_{t\to\infty} \frac{C(t)}{t} \leq \overline{N}_S \leq \lim_{t\to\infty} \frac{\sum_{i=1}^{A(t)} T_i}{A(t)} \cdot \lim_{t\to\infty} \frac{A(t)}{t}$$

$$\Rightarrow \overline{T}_s^{TimeAvg} \cdot X \leq \overline{N}_S^{TimeAvg} \leq \lambda \cdot \overline{T}_s^{TimeAvg}$$

But we assumed that X and λ are equal. Therefore

$$\overline{N}_S^{TimeAvg} = \lambda \cdot \overline{T}_S^{TimeAvg}$$

■

5 Some Applications of Little's Law

We now list some immediate applications of Little's Law:

Corollary 1

$$\overline{N}_Q^{TimeAvg} = \lambda \cdot \overline{T}_Q^{TimeAvg}$$

Proof: Same proof as Little's Law except that now T_i depicts the time the ith arrival to the system spends in queues (wasted time). Note that T_i may not be a solid rectangle. It may be made up of several rectangles because the ith job might be in queue for a while, then in service then in some other queue, then in service then in some other queue, etc.. ■

Corollary 2 *In a single queue with average arrival rate* λ *and average service rate* μ*:*

$$\rho = \frac{\lambda}{\mu}$$

Proof: The proof follows from Little's Law. Let the "system" consist of just the service facility *without* the associated queue, as shown in Figure 3. Now the number of jobs in the "system" is always just 0 or 1, but the *expected number of jobs in the system* is ρ.

So

$$\begin{aligned}
\rho &= \text{Expected number jobs in service facility for device.} \\
&= \text{Arrival rate into service facility} \cdot \text{Mean time in service facility} \\
&= \lambda \cdot \mathbf{E}\{\text{Service time at device}\} \\
&= \lambda \cdot \frac{1}{\mu}
\end{aligned}$$

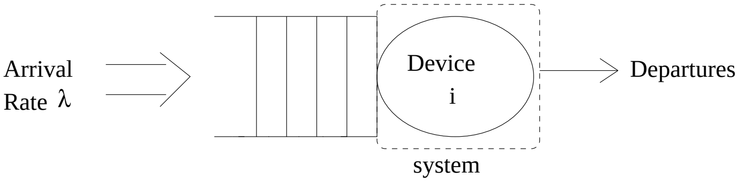

Figure 3: *Using Little's Law to prove the Utilization Law*

■

6 More powerful generalization: $\overline{H} = \lambda \overline{G}$

We now introduce a more powerful generalization of Little's Law, which is commonly referred to as:

$$\overline{H} = \lambda \overline{G}$$

and again applies to any queueing system.

When looking at this, it will help a lot to think of $\overline{H}$ as a generalization of $\overline{N}$, and think of $\overline{G}$ as a generalization of $\overline{T}$. The λ term is the same mean arrival rate as before.

We will need a function: $g_n(t)$ which is non-negative, and vanishes outside the interval $[a_n, a_n + \ell_n]$, where a_n denotes the time of arrival of the nth customer, and ℓ_n is any non-negative time (length) associated with the nth customer.

We define the following limits (when they exist):

$$\begin{aligned}
\lambda &\equiv \lim_{t\to\infty} \frac{A(t)}{t} \\
G_n &\equiv \int_{a_n}^{a_n+\ell_n} g_n(s)ds \\
\overline{G} &\equiv \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} G_j \\
H(t) &\equiv \sum_{n=0}^{\infty} g_n(t) \\
\overline{H} &\equiv \lim_{t\to\infty} \frac{1}{t} \int_0^t H(s)ds
\end{aligned}$$

Question: Consider the case where $g_n(t) = 1$ and where ℓ_n is T_n, the time that the nth customer spends in the system. What does $\overline{H} = \lambda \overline{G}$ say?

Answer: In this case: $\overline{G} = \overline{T}$, the mean time a customer spends in the system, and $\overline{H} = \overline{N}$, the mean number of customers in the system. So we simply get Little's Law.

More generally, $\overline{H} = \lambda \overline{G}$ can be interpreted as follows: Imagine a queueing system where the nth customer C_n arrives at time a_n and pays money to the system at rate $g_n(s)$ during time $a_n < s < a_n + \ell_n$. Then $H(s)$ represents the total money that the system takes in at time s, from all customers combined. Thus

$$\begin{aligned} \overline{H} &= \text{Average rate system earns money} \\ \lambda &= \text{Average rate customers arrive} \\ \overline{G} &= \text{Average money paid per customer} \end{aligned}$$

Question: So how would we prove $\overline{H} = \lambda \overline{G}$?

Answer: Basically same proof as for Little's Law, but the nth "rectangle" now has length ℓ_n, rather than T_n, and the top of the rectangle is a curve, rather than a line segment, since $g(s)$ fluctuates during $[a_n, a_n + \ell_n]$.

7 Applications of $\overline{H} = \lambda \overline{G}$

We show two examples of using $\overline{H} = \lambda \overline{G}$.

7.1 Deriving Brumelle's formula and the P-K mean delay formula

We will now show how Brumelle's formula for mean work, that Alan showed you, and the P-K formula for mean delay immediately follow from $\overline{H} = \lambda \overline{G}$.

Let $V(t)$ denote the *work* at time t in a GI/GI/1 queue, including the remaining time on the guy in service and the sizes of the jobs in the queue. Let $E[V]$ denote the the average work in the system, where:

$$E[V] = \lim_{t \to \infty} \frac{1}{t} \int_0^t V(s) ds$$

Let $E[D]$ denote the mean delay (time-in-queue) of jobs, where D_n will denote the delay of the nth job.

Our goal is to express $E[V]$ in terms of $E[D]$ and $E[S^2]$.

Question: What should $H(t)$ be?

Answer: $H(t) = V(t)$, because we want to look at $E[V]$.

So the total contributions by all customers at time t is $V(t)$.

Question: What is an individual customer contributing at time t?

Answer: Customer C_n should contribute its remaining size at time t.

Question: During what time period is customer C_n contributing?

Answer: During $[a_n, a_n + T_n]$.

$$g_n(t) = \begin{cases} S_n & \text{if } t \in [a_n, a_n + D_n]; \\ S_n - (t - (a_n + D_n)) & \text{if } t \in [a_n + D_n, a_n + D_n + S_n]; \\ 0 & \text{otherwise} \end{cases}$$

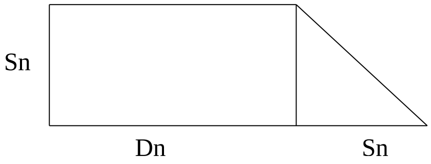

Figure 4: *Illustration of the cost function* $g_n(\cdot)$.

With the above definition, we have satisfied that:

$$V(t) = \sum_{n=1}^{\infty} g_n(t)$$

Looking at Figure 4 above, we see that G_n, the total contribution of customer C_n is:

$$G_n = \int_{a_n}^{a_n+D_n+S_n} g_n(t)dt = S_n D_n + \frac{S_n^2}{2}$$

Hence $\overline{G}$, the average customer contribution, is:

$$\begin{aligned} \overline{G} &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} G_i \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(S_i D_i + \frac{S_i^2}{2} \right) \end{aligned}$$

$$= \quad E[SD] + \frac{E[S^2]}{2}$$

Finally, applying $\overline{H} = \lambda \overline{G}$, we have:

$$E[V] = \lambda E[SD] + \lambda \frac{E[S^2]}{2}$$

which is known as *Brumelle's formula*.

Going a step further, since S_n and D_n are independent random variables, we have that:

$$\begin{aligned} E[V] &= \lambda E[S]E[D] + \lambda \frac{E[S^2]}{2} \\ &= \rho E[D] + \lambda \frac{E[S^2]}{2} \end{aligned}$$

Question: From here, it's almost immediate how we get to the P-K formula for mean delay in an M/GI/1. Can you see it?

Answer: Given PASTA, $E[V]$ represents the mean work as seen by an arrival, which is also $E[D]$. Hence we have:

$$\begin{aligned} E[D] &= \rho E[D] + \lambda \frac{E[S^2]}{2} \\ &= \frac{\lambda E[S^2]}{2(1-\rho)} \end{aligned}$$

which is the P-K formula for mean delay.

7.2 Deriving the mean attained sojourn time

To see another example of using $\overline{H} = \lambda \overline{G}$, consider the problem of deriving the mean *attained sojourn time* for the GI/GI/1 queue.

Let $W_a(t)$ denote the length of time that the customer who is currently in service at time t has already been in the system. ($W_a(t) = 0$ if the system is empty). This is called the *attained sojourn time* of the customer in service at time t. We are interested in $\overline{W_a}$, namely the average time that the customer in service has already been in the system.

Several papers (e.g., [Sakasegawa, Wolff 90], [Sengupta 89]) have looked at (more complex) approaches for how to derive $\overline{W_a}$. We will see that $\overline{W_a}$ follows very easily from $\overline{H} = \lambda \overline{G}$.

We are interested in:

$$\overline{H} = E[W_a(t)] = \lim_{t \to \infty} \frac{1}{t} \int_{s=0}^{\infty} W_a(s) dt$$

Question: Who is contributing to $W_a(t)$?

Answer: Only the customer currently in service at time t, if there is one, and no other customer.

Question: During what period of time does C_n contribute to W_a?

Answer: Only during the period: $[a_n + D_n, a_n + D_n + S_n]$.

Question: What does $g_n(t)$ look like:

$$g_n(t) = \begin{cases} D_n + (t - (a_n + D_n)) & \text{if } t \in [a_n + D_n, a_n + D_n + S_n] \\ 0 & \text{otherwise} \end{cases}$$

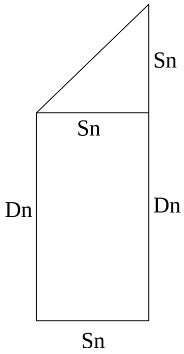

Figure 5: *Illustration of the cost function* $g_n(\cdot)$.

Question: So what is G_n, the contribution of the nth customer?

Answer:

$$G_n = S_n D_n + \frac{E[S_n^2]}{2}$$

Question: So what does $\overline{H} = \lambda \overline{G}$ tell us?

$$
\begin{aligned}
E[W_a] &= \lambda E[SD] + \lambda \frac{E[S^2]}{2} \\
&= \rho E[D] + \rho \frac{E[S^2]}{2E[S]}
\end{aligned}
$$

Thus, interestingly, we have proven that the expected attained sojourn time is the same as the average workload, for a GI/GI/1 system.

8 Rate Conservation Law (RCL)

Like $\overline{H} = \lambda \overline{G}$, our next law, the rate conservation law (RCL) also relates time and customer averages. However, whereas in $\overline{H} = \lambda \overline{G}$, we were drawing rectangles and figuring out areas, RCL is much more abstract and analytical, and doesn't usually have a nice geometrical interpretation.

We will begin by deriving RCL:

Consider a function $\{x(t), t \geq 0\}$ which has jumps (discontinuities) at certain points: $a_1, a_2, a_3, \ldots$, where $N(t)$ denotes the number of jumps by time t, as shown in Figure 6. Other from these discontinuities, the function is continuous, and

$$x'(t) \equiv \lim_{h \to 0} \frac{x(t+h) - x(t)}{h}$$

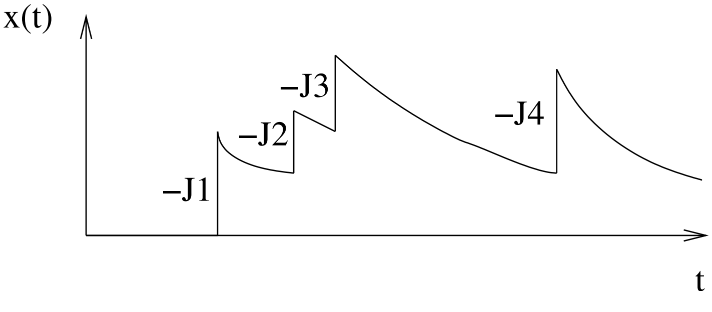

Figure 6: *Illustration of what $x(t)$ might look like.*

We denote the nth jump size by $-J_n$. That is:

$$-J_n = x(a_n+) - x(a_n-)$$

Note, importantly, that the jump size is defined to be negative what you might imagine it to be. Hence:

$$x(t) - x(0) \quad = \quad \int_0^t x'(s)ds - \sum_{n=0}^{N(t)} J_n \tag{1}$$

We now make the following definitions (assuming that the limits exist):

$$\begin{aligned}
\lambda &\equiv \lim_{t\to\infty} \frac{N(t)}{t} \\
E[J] &\equiv \lim_{n\to\infty} \sum_{j=1}^{n} J_n \\
E[x'] &\equiv \lim_{t\to\infty} \frac{1}{t} \int_0^t x'(s)ds
\end{aligned}$$

Even through no stochastic setup is assumed, we us the expectation notation, $E[x']$, for convenience.

Theorem 2 (RCL theorem) *Assuming λ and $E[J]$ exist and are finite, and assuming that $\frac{x(t)}{t} \to 0$ as $t \to \infty$, then $E[x']$ exists and*

$$E[x'] = \lambda E[J]$$

Proof: Starting from Equation (1) we have that:

$$\begin{aligned}
\int_0^t x'(s)ds &= x(t) - x(0) + \sum_{n=0}^{N(t)} J_n \\
\lim_{t\to\infty} \frac{1}{t} \int_0^t x'(s)ds &= \lim_{t\to\infty} \frac{x(t) - x(0)}{t} + \lim_{t\to\infty} \frac{1}{t} \sum_{n=0}^{N(t)} J_n \\
E[x'] &= 0 + \lim_{t\to\infty} \frac{\sum_{n=0}^{N(t)} J_n}{N(t)} \cdot \frac{N(t)}{t} \\
E[x'] &= E[J] \cdot \lambda
\end{aligned}$$

■

RCL can be extended in many ways. Jumps could occur both at arrival instants of jobs and at departure times. We can have m different classes of jobs, with a sequence of arrival times for each type, where

$$E[x'] = \lambda_1 E_1[J] + \lambda_2 E_2[j] + \cdots + \lambda_m E_m[j],$$

where $-E_i[j]$ denotes the average jump size for jobs of class i.

9 Applications of RCL

Thus far, RCL is very abstract. We will now use RCL to derive some M/GI/1 queueing formulas. You will see that the difficulty in using RCL is that, because of its abstract nature, it is often not at all clear how to define $x(t)$.

9.1 Deriving $\rho = \lambda E[S]$ for an M/GI/1

This is the simplest example of RCL that I could find.

We want to derive ρ, the long-run proportion of time that the server is busy. We want to show that $\rho = \lambda E[S]$, which, remember we originally derived by applying Little's Law to just the server component.

Question: How do we start? What's $x'(t)$? What's $x(t)$?

Answer: Before you get to $x(t)$, look for a clue. To think about how to apply RCL to derive this, you should first note that the jump size is likely to be related to S, or, more specifically, $-S$.

Question: So ask yourself what might have a jump of S?

Answer:

Let

$$x(t) = V(t),$$

the remaining work at time t. This has a jump of S every time that there is an arrival.

Question: So what is $x'(t)$?

Answer:

Since the server is assumed to be working at rate 1, we have:

$$x'(t) = V'(t) = \begin{cases} 0 & \text{if } V(t) = 0 \\ -1 & \text{if } V(t) > 0 \end{cases}$$

Hence,

$$E[x'] = -\mathbf{P}\{V > 0\} = -\rho$$

Together with:

$$E[J] = -E[S]$$

We have:

$$\begin{aligned} E[x'] &= \lambda E[J] \\ -\rho &= \lambda(-E[S]) \\ \rho &= \lambda E[S] \end{aligned}$$

and we're done.

9.2 Deriving the P-K formula for $E[D]$ for an M/GI/1

We next show how to derive the mean delay in an M/GI/1. We will do this by deriving $E[V]$, the mean work, which we know to be the same as the mean delay for a FCFS queue.

Question: If our goal is $E[V]$, what should $x(t)$ be?

Answer: Well, we know that we want to have $E[x']$ eventually come out to be $E[V]$. So $x(t)$ is likely to be the integral of V.

Let's try:

$$x(t) = V(t)^2$$

Then:

$$x'(t) = 2V(t)\frac{dV(t)}{dt} = -2V(t)$$

(Note that the value $-2V(t)$ is clearly correct when $V(t) > 0$. However, when $V(t) = 0$, the correct answer is $x'(t) = 0$, so $-2V(t)$ is also correct in this case as well.).

Question: A jump in $x(t)$ occurs at each arrival point. But what is the jump size?

Answer:

$$\begin{aligned}
-J_n &= x(a_n+) - x(a_n-) \\
&= (V(a_n-) + S_n)^2 - V(a_n-)^2 \\
&= (D_n + S_n)^2 - D_n^2 \\
&= 2S_nD_n + S_n^2
\end{aligned}$$

So

$$-E[J] = 2E[SD] + E[S^2]$$

Now applying RCL, we have:

$$\begin{aligned}
E[x'] &= \lambda E[J] \\
2E[V] &= \lambda(2E[SD] + E[S^2])
\end{aligned}$$

Finally, observing, as before that S and D are indpt, and that $E[D] = E[V]$, we have:

$$\begin{aligned}
E[V] &= \lambda(E[S]E[D] + \frac{E[S^2]}{2}) \\
&= \lambda E[S]E[V] + \lambda\frac{E[S^2]}{2} \\
&= \rho E[V] + \lambda\frac{E[S^2]}{2} \\
&= \frac{\lambda E[S^2]}{2(1-\rho)}
\end{aligned}$$

yielding the desired result.

9.3 Deriving $E[D^2]$ for an M/GI/1

Homework problem #1 Use RCL to derive $E[D^2]$ for the M/GI/1 by using RCL.

9.4 Deriving $\overline{N} = \lambda\overline{T}$ from RCL

Homework problem #2 – Extra Credit Show that you can obtain Little's Law from RCL. This is more challenging, so I'm making it extra-credit. The proof is short (only 5-10 lines), but it's hard to figure out how to define $x(t)$.

10 Where we're going ...

RCL is extremely powerful for generating *simple* proofs. Don't worry, we're not done with RCL. We'll come back to it, after we introduce transforms. We will spend the next couple lectures on transforms. Here are a few "bits" to get you started thinking about these:

Definition 1 *The* **Laplace Transform** $L_f(s)$ *of a continuous function (on positive real axis)* $f(t)$ *is defined as*

$$L_f(s) = \int_0^\infty e^{-st} f(t) dt$$

When we speak of the Laplace Transform of a continuous R.V. X, we are referring to the Laplace transform $L_f(s)$, of the p.d.f. f associated with X. Often we will write $\tilde{X}(s)$ to denote the Laplace transform of X.

Observe that if X is a continuous R.V. and f is the p.d.f. of X, then

$$\tilde{X}(s) = L_f(s) = E[e^{-sX}]$$

Example: Consider the Laplace Transform of $X = Exp(\lambda)$:

$$\begin{aligned} \tilde{X}(s) = L_f(s) &= \int_0^\infty e^{-st} \lambda e^{-\lambda t} dt \\ &= \lambda \int_0^\infty e^{-(\lambda+s)t} dt \\ &= \frac{\lambda}{\lambda + s} \end{aligned}$$

One of the really nice things about transforms, is that once you have the transform of a random variable, you can immediately get all moments of that random variable. Thus, we can think of the transform as a "compact suitcase" that holds all sorts of valuable stuff, such as the moments.

Theorem 3 *Let* X *be a continuous R.V. with p.d.f.* $f(t)$. *Then*

$$E[X^n] = (-1)^n \frac{d^n L_f(s)}{ds} \bigg|_{s=0}$$

Proof:

$$
\begin{aligned}
e^{-st} &= 1 - (st) + \frac{(st)^2}{2!} - \frac{(st)^3}{3!} + \dots \\
e^{-st} f(t) &= f(t) - (st)f(t) + \frac{(st)^2}{2!} f(t) - \frac{(st)^3}{3!} f(t) + \dots \\
L_f(s) = \int_0^\infty e^{-st} f(t)dt &= \int_0^\infty f(t)dt - \int_0^\infty (st)f(t)dt + \int_0^\infty \frac{(st)^2}{2!} f(t)dt - \int_0^\infty \frac{(st)^3}{3!} f(t)dt + \dots \\
&= 1 - sE[X] + \frac{s^2}{2!}E[X^2] - \frac{s^3}{3!}E[X^3] + \dots \\
\frac{dL_f(s)}{ds} &= -E[X] + sE[X^2] - 3\frac{s^2}{3!}E[X^3] + \dots \\
\left.\frac{dL_f(s)}{ds}\right|_{s=0} &= -E[X] \\
\frac{d^2 L_f(s)}{ds} &= E[X^2] - sE[X^3] + \dots \\
\left.\frac{d^2 L_f(s)}{ds}\right|_{s=0} &= E[X^2] \\
& \quad etc.
\end{aligned}
$$

■

Question: How do we know that the Laplace Transform as defined necessarily converges?

Answer (partial): It does provided f(t) is a p.d.f.. To see this observe that

$$e^{-t} \leq 1$$

for all non-negative values of t. Thus

$$e^{-st} = \left(e^{-t}\right)^s \leq 1$$

assuming that s is non-negative. Thus:

$$L_f(s) = \int_0^\infty e^{-st} f(t)dt \leq \int_0^\infty 1 \cdot f(t)dt = 1$$

assuming that s is non-negative.

Observation 1 *Observe that in Theorem 3, $f(t)$ did not have to be a p.d.f.. In particular, for any function $f(t)$,*

$$\int_{t=0}^{\infty} t^n \cdot f(t)dt = (-1)^n \frac{d^n L_f(s)}{ds}\bigg|_{s=0}$$

11 Acknowledgments/Sources

In writing this lecture, I made use of:

1. The Ph.D. thesis of Gautam Jaim, "A Rate Conservation Analysis of Queues and Networks with Work Removal." Columbia University, IEOR.
2. The following paper: "Rate Conservation Laws: A Survey," by Masakiyo Miyazawa, *Queueing Systems*, vol. 15, No.1-4, pp. 1-58, 1994.
3. Lecture notes from Karl Sigman, Columbia University.

Thank You!