6.262: Discrete Stochastic Processes 3/16/11

Lecture 13 : Little, M/G/1, ensemble averages

Outline:

- Review Wald's equality
- The elementary renewal theorem
- Generalized stopping trials
- The G/G/1 queue
- Little's theorem
- Pollaczek-Khinchin result for M/G/1

Def: A stopping trial (or stopping time) J for a sequence $\{X_n;\, n \geq 1\}$ of rv's is a positive integer-valued rv such that for each $n \geq 1$, the indicator rv $\mathbb{I}_{\{J=n\}}$ is a function of $\{X_1, X_2, \dots, X_n\}$.

A possibly defective stopping trial is the same except that J might be a defective rv. For many applications of stopping trials, it is not initially obvious whether J is defective.

Theorem (Wald's equality) Let $\{X_n; n \geq 1\}$ be a sequence of IID rv's, each of mean $\overline{X}$. If J is a stopping trial for $\{X_n;\, n \geq 1\}$ and if $\mathsf{E}[J] < \infty$, then the sum $S_J = X_1 + X_2 + \cdots + X_J$ at the stopping trial J satisfies

$$\mathsf{E}[S_J] = \overline{X}\mathsf{E}[J]\,.$$

The elementary renewal theorem

Wald's equality is useful for determining $\mathsf{E}[N(t)]$ as a function of t for a renewal counting process. We have the strong and weak laws for $N(t)$ as $t \to \infty$, but often it is useful to be explicit for finite t.

For a given t, let J be the smallest n for which $S_n > t$. Then J is a stopping trial for the inter-arrivals $\{X_i;\, i \geq 1\}$. That is, we stop at trial n if $S_n > t$ and $S_{n-1} \leq t$, and this is determined by $X_1, \dots, X_n$.

Note $N(t)$ is the number of arrivals that have occured up to and including t, so $N(t) + 1 = J$ is the number of the first arrival after J. Since $\mathsf{E}[N(t)]$ is finite, $\mathsf{E}[J] < \infty$. From Wald,

$$\mathsf{E}\left[S_{N(t)+1}\right] = \overline{X}\mathsf{E}[J] = \overline{X}(\mathsf{E}[N(t)] + 1)$$

$$\mathsf{E}\left[S_{N(t)+1}\right] = \overline{X}\mathsf{E}[J] = \overline{X}(\mathsf{E}[N(t)] + 1)$$

Wald's equality relates two unknown quantities, $\mathsf{E}[N(t)]$ and $\mathsf{E}\left[S_{N(t)+1}\right]$. Since $\mathsf{E}\left[S_{N(t)+1}\right] > t$, we get a simple bound from this.

$$\mathsf{E}[N(t)] \;=\; \frac{\mathsf{E}\left[S_{N(t)+1}\right]}{\overline{X}} - 1 \;>\; \frac{t}{\overline{X}} - 1$$

Thm (Elementary renewal thm): Let $\overline{X}$ be mean inter-renewal of a renewal counting process $\{N(t);\, t > 0\}$. Then $\lim_{t\to\infty} \mathsf{E}[N(t)/t] = 1/\overline{X}$.

Pf: Need an upper bound on $\mathsf{E}[N(t)]$. Truncate X to $X \leq b$, carry out bound, let b grow with t.

We can view this as convergence in the mean (one more type of convergence).

Generalized stopping trials

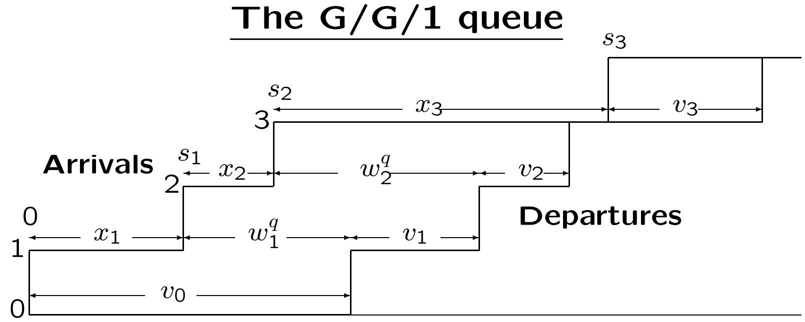
Def: A generalized stopping trial J for a sequence of pairs of rv's $(X_1, V_1), (X_2, V_2)\dots$, is a positive integer rv such that, for each $n \geq 1$, $\mathbb{I}_{\{J=n\}}$ is a function of $X_1, V_1, X_2, V_2, \dots, X_n, V_n$.

It follows that $\mathbb{I}_{\{J<n\}} = 1 - \mathbb{I}_{\{J\geq n\}}$ is a function of $X_1, V_1, X_2, V_2, \dots, X_{n-1}, V_{n-1}$.

Wald's equality, $\mathsf{E}[S_n] = \overline{X}\mathsf{E}[J]$, where $S_n = X_1 + \cdots + X_n$ still holds (by the same proof) if the X_i are IID and each X_i is independent of $(X_1, V_1, \dots, X_{i-1}, V_{i-1})$.

Also each V_i can be replaced by a vector of rv's.

The G/G/1 queue



Consider the first arrival (s_3 above) that starts a new busy period as a generalized stopping trial.

The sequence of paired rv's is $(X_1, V_0), (X_2, V_1), \dots$ 'Stopping' at $J = 3$ is $f(X_1, V_0, X_2, V_1, X_3, V_2)$.

Wald's equality holds. Also new arrivals $(X_{J+1}, X_{J+2}, \dots$ and services $(V_J, V_{J+1}, \dots$ are independent of the old.

The stopping rule J here is the index of the first arrival in a new busy period. The arrivals and departures in the new busy period are independent and identically distributed to those in the old.

Thus the intervals between new busy periods form a renewal process.

We then have one renewal process embedded in another. Call one the arrival process and the other the renewal process. The renewal process embodies both arrivals and services.

This analysis applies also to G/G/m and to many other queuing systems.

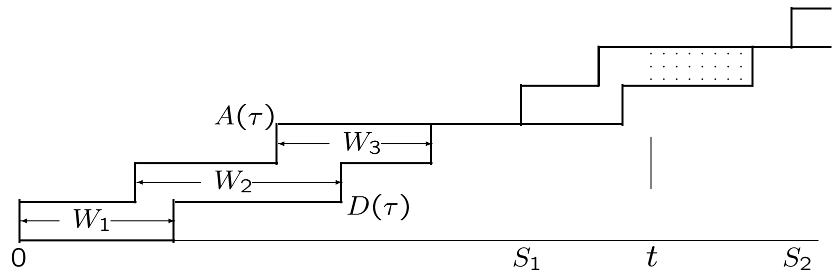
Little's theorem

Consider a queueing system where the arrival process is a renewal process. Assume an arrival at time 0.

The service process can be almost anything, but assume a G/G/1 queue to be specific.

Assume the system empties out eventually WP1 and that it restarts on the next arrival.

We have seen that intervals between restartings form a renewal process for the G/G/1 queue, and for an even broader class of queues.



Let $L(\tau) = A(\tau) - D(\tau)$. This depends on the departure process, but its value is a function of the inter-arrivals and service times within the current inter-renewal period.

Thus it can be viewed as a generalized renewal-reward function.

The total reward within an inter-renewal period is then the integral of $L(\tau)$ over that period (i.e., R_n).

In each inter-renewal period,

$$R_n = \int L(\tau)\, d\tau = \sum_i W_i,$$

where the sum is over the arrivals in that inter-renewal period. The time averages are then

$$\begin{aligned} L_{ta} &= \lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{A(t)} W_i \\ &= \lim_{t\to\infty} \frac{\sum_{i=1}^{A(t)} W_i}{A(t)} \lim_{t\to\infty} \frac{A(t)}{t} \\ &= W_{ta}\,\lambda \end{aligned}$$

where λ is arrival rate.

This is Little's theorem. The time-average number in the system equals λ times the time-average customer wait WP1.

In essence, Little's theorem is an accounting identity. Over a busy period and the following idle period (i.e., an inter-renewal period), $\int L(\tau)\,d\tau = \sum_i W_i$.

To turn this simple result into mathematics, we need renewal theory, which essentially allows us to go to the limit of many renewals.

The question is not whether $L = \lambda W$, but whether these quantities exist as sensible time averages WP1 or as limiting ensemble averages.

The Pollaczek-Khinchin formula for M/G/1 queues

Let $X_1, X_2, \ldots$ be IID exponential arrivals at rate λ and let $V_1, V_2, \ldots$ be IID service times with first and second moments $\overline{V} = \mathsf{E}[V]$ and $\overline{V^2} = \mathsf{E}\left[V^2\right]$. Then the Pollaczek-Khinchin formula gives the expected delay in queue (between arrival and entering service) as

$$\overline{W}^{\,q} = \frac{\lambda \overline{V^2}}{2(1-\rho)} \qquad \text{where } \rho = \lambda \overline{V}$$

The expected total delay, total number in system and number in queue are then

$$\overline{W} = \frac{\lambda \overline{V^2}}{2(1-\rho)} + \overline{V}; \quad \overline{N} = \frac{\lambda^2 \overline{V^2}}{2(1-\rho)} + \rho; \quad \overline{N}^q = \frac{\lambda^2 \overline{V^2}}{2(1-\rho)}$$

Examples: For deterministic service, $\overline{V^2} = (\overline{V})^2$, so

$$\overline{W^q} = \frac{\rho\overline{V}}{2(1-\rho)} \qquad \text{for M/D/1}$$

For exponential inter-arrivals M/M/1, $\overline{V^2} = 2(\overline{V})^2$ so

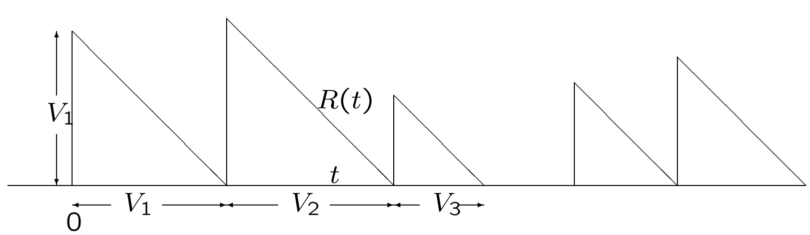
$$\overline{W^q} = \frac{\rho\overline{V}}{1-\rho} \qquad \text{for M/M/1.}$$

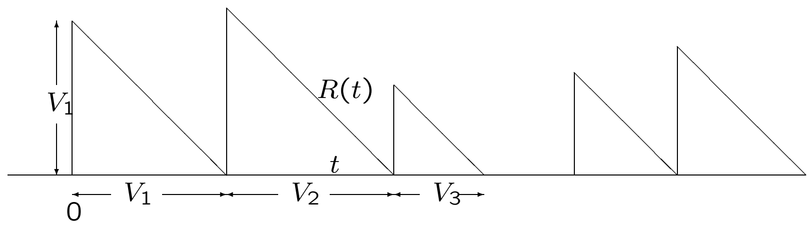
For $\mathsf{p}_V(\epsilon) = 1-\epsilon$, $\mathsf{p}_V(1/\epsilon) = \epsilon$,

$$\overline{W^q} \approx \frac{\rho}{\epsilon(1-\rho)}$$

Why does $\overline{W^q}$ go up with $\overline{V}^2$? Look at the time-average wait, $\mathsf{E}[R(t)]$, for the customer in service to finish service.

Have you ever noticed, when entering a line for service that the customer being served often takes much longer than anyone else?





$$
\begin{aligned}
\mathsf{E}[R] &= \lim_{\tau\to\infty}\frac{1}{\tau}\int_0^\tau R(t)\,dt = \lim_{\tau\to\infty}\frac{1}{\tau}\sum_{i=1}^{A(\tau)}\frac{V_i^2}{2} \\
&= \frac{\lambda\overline{V^2}}{2} \qquad \textbf{(time until end of current service)} \\
\overline{W}^q &= \mathsf{E}[R] + \overline{N}^q\overline{V} \qquad \textbf{(add service of queue)} \\
&= \frac{\lambda\overline{V^2}}{2(1-\lambda\overline{V})} \qquad \textbf{(a little Little)}
\end{aligned}
$$

6.262 Discrete Stochastic Processes
Spring 2011