Generalized Semi-Markov Processes (GSMP’s)

Ref: Section 1.4 in Shedler or Section 4.1 in Haas

1. Motivation

The Markov and semi-Markov models that we have discussed previously do not have sufficient modeling power to capture many of the complex discrete-event stochastic systems that arise in practice. The exponential distributional assumptions of the CTMC model often do not hold; neither does the implicit assumption in the semi-Markov process model that only a single “clock” is running in each state. The GSMP model avoids these restrictive assumptions. See the textbook by Gerald Shedler for a thorough treatment of GSMP’s.

Heuristically, a GSMP \{X(t): t \geq 0\} makes stochastic state transitions when one or more events associated with the occupied state occur. (Associated events = events that can possibly occur in the state = events that are scheduled in the state.)

- events associated with a state “compete” to trigger the next state transition
- each event has its own distribution for determining the next state
- new events can be scheduled at each state transition—a new event with respect to a state transition is an event that is associated with the new state and either (a) is not associated with the old state or (b) is associated with the old state and also triggers the state transition
- for each new event, a clock is set with a reading that indicates the time until the event is scheduled to occur; when the clock runs down to 0 the event occurs (unless it is cancelled in the interim)

\[
\text{clock reading} \quad \rightarrow \quad \text{time}
\]

- an old event with respect to a state transition is an event, associated with the old state, that does not trigger the state transition and is associated with the next state; its clock continues to run down
- a cancelled event with respect to a state transition doesn’t trigger the state transition and is not associated with the next state; its clock reading is discarded
- clocks can run down at state-dependent speeds

2. GSMP Building Blocks

- \textbf{S}: a (finite or countably infinite) set of states
- \textbf{E} = \{e_1, e_2, \ldots, e_M\}: a finite set of events
- \textbf{E}(s): the set of events scheduled to occur in state \textbf{s} \in \textbf{S}. Of course, \textbf{E}(s) \subseteq \textbf{E}. We say that event \textbf{e} is \textit{active} in \textbf{s} if \textbf{e} \in \textbf{E}(s).
• $p(s'; s, E^*)$: the probability that the new state is $s'$ given that the events in $E^*$ simultaneously occur in $s$. If $E^* = \{e^*\}$ for some $e^* \in E(s)$, then we simply write $p(s'; s, e^*)$.

• $r(s, e)$: the nonnegative finite speed at which clock for $e$ runs down in state $s$; typically $r(s, e) = 1$, but can be set to other values in order to model “processor sharing” or “preempt resume” service discipline. (For modeling the latter we allow $r(s, e) = 0$.)

• $F(\cdot; s', e', s, E^*)$: the distribution function used to set the clock for the new event $e'$ when the simultaneous occurrence of the events in $E^*$ triggers a state transition from $s$ to $s'$.

• $\mu$: the initial distribution function for the state and clock readings. We will always assume that $\mu$ is such that the initial state $s$ is chosen according to a distribution $\nu$ and for each event $e \in E(s)$ the clock is set independently according to $F_0(\cdot; e, s)$.

### Sets of new and old events:

- New transitions
- Old transitions
- Newly disabled transitions

$$\begin{align*}
(E(s) - E(s')) - E^* &\quad N(s'; s, E^*) = E(s') - (E(s) - E^*) \\
(E(s) - E(s')) \cap E^* &\quad O(s'; s, e) = E(s') \cap (E(s) - E^*)
\end{align*}$$

### Example: GI/G/1 Queue

Assume that the interarrival-time distribution $F_a$ and the service-time distribution $F_s$ are continuous, so that an arrival and service completion never occur simultaneously. Also assume that a job arrives to an empty system at time 0.

Let $X(t) = \text{number of jobs in service or waiting in queue at time } t$

Then $\{X(t): t \geq 0\}$ can be specified as a GSMP with

- $S = \{0, 1, 2, \ldots\}$
- $E = \{e_1, e_2\}$, where $e_1 = \text{“arrival”}$ and $e_2 = \text{“completion of service”}$
- $E(s) = \{e_1\}$ if $s = 0$ and $E(s) = \{e_1, e_2\}$ if $s > 0$
- $p(s+1; s, e_1) = 1$ and $p(s-1; s, e_2) = 1$
- $F(x; s', e', s, e) = F_a(x)$ if $e' = e_1$ and $F_s(x)$ if $e' = e_2$
- $r(s, e) = 1$ for all $s$ and $e$
- $\nu(1) = 1$, $F_0(\cdot; e_1, s) = F_a(\cdot)$, and $F_0(\cdot; e_2, s) = F_s(\cdot)$
3. GSMP’s and GSSMC’s

A GSMP is formally defined in terms of a GSSMC \( \{(S_n, C_n): n \geq 0\} \), where

- \( S_n = \) state just after \( n^{th} \) transition
- \( C_n = (C_{n,1}, C_{n,2}, \ldots, C_{n,M}) = \) clock-reading vector after \( n^{th} \) transition (by convention, \( C_{n,i} = 0 \) if event \( e_i \) is not active in state \( S_n \))

The transition kernel is given by a somewhat complicated-looking expression, namely

\[
P((s,c), A) = p(s'; s, E^*) \prod_{x \in N(s')} F(a_i; s', s, E^*) \prod_{x \in O(s')} I_{[0, a_i]}(c_i^*)
\]

for sets \( A = \{s'\} \times \{(c_1', \ldots, c_M') \in C(s): 0 \leq c_i' \leq a_i \text{ for } 1 \leq i \leq M\} \), where \( E^* = E^*(s,c) \), \( N(s') = N(s'; s, E^*) \), \( O(s') = O(s'; s, E^*) \), and \( c_i^* = c_i^*(s,c) = c_i - t^*(s,c) r(s,e_i) \).

The initial distribution \( \mu(A) \) is of the form

\[
\mu(A) = v(s') \prod_{x \in E(s')} F_0(a_i; e_i, s').
\]

for a set \( A \) as above.

To construct the process \( \{X(t): t \geq 0\} \) from \( \{(S_n, C_n): n \geq 0\} \), let \( \zeta_n \) be the time of the \( n^{th} \) state transition:

\[
\zeta_n = \sum_{k=0}^{n-1} t^*(S_k, C_k)
\]

where \( t^*(s,c) \) is the holding time in state \( s \) starting with clock-reading vector \( c = (c_1, c_2, \ldots, c_M) \):

\[
t^*(s, c) = \min_{[i \in E(s)]: r(s,e_i)} \{c_i / r(s,e_i)\}
\]

and \( N(t) \) the number of state transitions in \( [0, t] \):

\[
N(t) = \max\{n \geq 0: \zeta_n \leq t\}
\]

Then let \( \Delta \not\in S \) and set

\[
X(t) = \begin{cases} S_{N(t)} & \text{if } N(t) < \infty \\ \Delta & \text{if } N(t) = \infty \end{cases}
\]

(If \( S \) is finite then \( N(t) \) is finite for each \( t < \infty \).)

for \( t \geq 0 \). The process \( \{X(t): t \geq 0\} \) is the GSMP. Thus, we can use results from GSSMC theory to study GSMP’s.
4. Sample Path Generation

The GSMP definition leads directly to an algorithm for sample-path generation:

Sample-Path Generation Algorithm

1. (Initialization) Select \( s \in S \) according to \( \nu \). For each \( e_i \in E(s) \) generate a clock reading \( c_i \) according to \( F_i(\cdot; e_i, s) \). Set \( c_i = 0 \) for \( e_i \not\in E(s) \).
2. Determine the time \( t^*(s,c) \) until the next state transition and determine the set of events \( E^* = E^*(s,c) = \{ e_i: c_i / r(s, e_i) = t^*(s,c) \} \) that will trigger the next state transition.
3. Generate the next state \( s' \) according to the probability mass function \( p(\cdot; s, E^*) \).
4. For each \( e_i \in N(s'; s, E^*) \), generate \( c_i' \) according to \( F(\cdot; s', e_i, s, E^*) \).
5. For each \( e_i \in O(s'; s, E^*) \), set \( c_i' = c_i - t^*(s,c)r(s,e_i) \).
6. For each \( e_i \in (E(s) - E^*) \cap E(s') \), set \( c_i' = 0 \) (i.e., cancel event \( e_i \)).
7. set \( s = s' \) and \( c = c' \), and go to Step 2. (Here \( c = (c_1, c_2, \ldots, c_M) \) and similarly for \( c' \)).

This algorithm generates a sequence of states \( \{S_n: n \geq 0\} \), clock-reading vectors \( \{C_n: n \geq 0\} \), holding times \( \{t^*(S_n, C_n): n \geq 0\} \). The state-transition times \( \{\zeta_n: n \geq 0\} \) are then computed as

\[
\zeta_n = \sum_{k=0}^{n-1} t^*(S_k, C_k).
\]

The algorithm is an example of a variable time-advance procedure (and is given at a high level of abstraction). We will discuss some more detailed computational issues for such procedures shortly. (Exercise: What other kinds of time advance procedures can you think of?)
If each clock-setting distribution function $F(\cdot; s', e', s, e^*)$ is exponential with rate $\lambda(s', e', s, e^*)$, is \{X(t): t \geq 0\} a CTMC? Not in general. (Exercise: find a counterexample.) But we have the following result, which gives an alternative way to specify (and to view) a CTMC.

**Definition:** An event $e'$ is simple if

$$F(\cdot; s', e', s, E^*) \equiv F(\cdot; s')$$

and

$$F_0(\cdot; e', s) = F(\cdot; e')$$

for some function $F(\cdot; s')$ and all $s'$, $s$, and $E^*$.

**Theorem:** Let \{X(t): t \geq 0\} be a GSMP in which each event $e$ is simple, having an exponential clock-setting distribution with intensity $\lambda(e)$. Then \{X(t): t \geq 0\} is a CTMC with rate matrix $Q$ given by

$$Q(s, s') = \sum_{e \in E(s)} \lambda(e) r(s, e) p(s'; s, e) \quad \text{for } s' \neq s.$$ 

Moreover, at any time point the clock reading for an active event $e$ is exponentially distributed with intensity $\lambda(e)$.

Note that, summing over $s' \neq s$, we have

$$q(s) = \sum_{s' \neq s} Q(s, s') = \sum_{e \in E(s)} \lambda(e) r(s, e) [1 - p(s; s, e)].$$

This result generalizes our earlier result on CTMC’s by allowing jumps back from a state to itself.

**Theorem:** Let \{X(t): t \geq 0\} be a GSMP in which each clock-setting distribution is simple and $|E(s)| = 1$ for $s \in S$. Then \{X(t): t \geq 0\} is a semi-Markov process.

There are more elaborate versions of the above result, in which there is at most one non-exponential event is active in a state. (Exercise: give a precise statement of such a result.)

6. A More Complicated Example

**Patrolling Repairman:**

- N machines
- single repairman visits machines in order $1 \rightarrow 2 \rightarrow \cdots \rightarrow N \rightarrow 1 \rightarrow 2 \rightarrow \cdots$
- repairs stopped machine, passes running machine
- repair times for machine $j$ are i.i.d. as a random variable $R_j$
- lifetimes for machine $j$ are i.i.d. as a continuous random variable $L_j$
- walking time from machine $j$ to next machine is a constant $W_j > 0$
- at time 0, the repairman has just finished repairing machine 1 and all other machines are broken.
Suppose we wish to estimate $\mu_r$, the expected fraction of time in $[0, t]$ that the repairman spends repairing machines. If we define our system state by $X(t) = A(t)$, where

$$A(t) = \begin{cases} 
1 & \text{if repairman is repairing a machine} \\
0 & \text{otherwise}
\end{cases}$$

then $\mu_r = \mathbb{E}\left[\frac{1}{t} \int_0^t A(u) \, du\right]$. We might also want to estimate $\mu_s$, the expected number of stopped machines at time $t$, or $\mu_w$, the long-run average wait for repair for machine 1.

Problems:

- can’t determine number of stopped machines from observing $A(t)$
- not clear how to generate sample paths of $\{A(t): t \geq 0\}$

⇒ need to put more information into state definition

Here’s another attempt at a state definition:

$$X(t) = (Z_1(t), Z_2(t), \ldots, Z_N(t), M(t), N(t)),$$

where

$$Z_j(t) = \begin{cases} 
1 & \text{if machine } j \text{ is waiting for repair at time } t \\
0 & \text{otherwise}
\end{cases}$$

$$M(t) = \begin{cases} 
j & \text{if machine } j \text{ is under repair at time } t \\
0 & \text{if no machine is under repair at time } t
\end{cases}$$

$$N(t) = j \text{ if at time } t \text{ the repairman will next arrive at machine } j$$

Then we can generate sample paths of $\{X(t): t \geq 0\}$ (because this process is a well-defined GSMP as shown below and, as was shown earlier, there is a well-defined algorithm for generating sample paths of a GSMP). Also, all of the system characteristics of interest can be precisely expressed in terms of $\{X(t): t \geq 0\}$:

$$\mu_r = \mathbb{E}\left[\frac{1}{t} \int_0^t f_r(X(u)) \, du\right] \quad \text{and} \quad \mu_s = \mathbb{E}[f_s(X(t))]$$

where

$$f_r(z_1, \ldots, z_N, m, n) = 1_{\{1,2,\ldots,N\}}(m)$$
\[ f(z_1, \ldots, z_N, m, n) = z_1 + z_2 + \ldots + z_N + 1_{\{1,2,\ldots,N\}}(m) \]

(Recall that \(1_A(x) = 1\) if \(x \in A\) and \(1_A(x) = 0\) otherwise.)

Also, we can express \(\mu_w\) in terms of \(\{X(t) : t \geq 0\}\). To see this, set \(B_0 = 0\) and recursively define the start and termination of the \(n\)th waiting time for machine 1 by

\[ A_n = \min\{\zeta_k > B_{n-1} : Z_1(\zeta_{k-1}) = 0 \text{ and } Z_1(\zeta_k) = 1\} \]

and

\[ B_n = \min\{\zeta_k > A_n : M(\zeta_{k-1}) \neq 1 \text{ and } M(\zeta_k) = 1\} \]

where \(\zeta_n\) is the time of the \(n\)th state transition.

We can then write the \(n\)th waiting time as \(D_n = B_n - A_n\), and hence

\[ \mu_w = \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} D_k \] (assuming that it exists)

The process \(\{X(t) : t \geq 0\}\) can be specified as a GSMP as follows:

- \(S\) consists of all \((z_1, \ldots, z_N, m, n) \in \{0,1\}^N \times \{0, 1, \ldots, N\} \times \{1, 2, \ldots, N\}\) such that
  - \(n = m + 1\) if \(0 < m < N\)
  - \(n = 1\) if \(m = N\)
  - \(m = j\) only if \(z_j = 0\) (\(1 \leq j \leq N\))
- \(E = \{e_1, e_2, \ldots, e_{N+2}\}\), where
  - \(e_j = \text{"stoppage of machine } j\" (1 \leq j \leq N)
  - \(e_{N+1} = \text{"completion of repair"}\)
  - \(e_{N+2} = \text{"arrival of repairman"}\)
- \(E(s)\) is defined as follows for \(s = (z_1, \ldots, z_N, m, n)\):
  - \(e_j \in E(s)\) (\(1 \leq j \leq N\)) iff \(z_j = 0\) and \(m \neq j\)
  - \(e_{N+1} \in E(s)\) iff \(m > 0\)
  - \(e_{N+2} \in E(s)\) iff \(m = 0\)
- \(p(s' ; s, e^*)\) is defined as follows:
  - if \(e^* = e_j (1 \leq j \leq N)\), then \(p(s' ; s, e^*) = 1\)
    when \(s = (z_1, \ldots, z_{j-1}, 0, z_{j+1}, \ldots, z_N, m, n)\) and \(s' = (z_1, \ldots, z_{j-1}, 1, z_{j+1}, \ldots, z_N, m, n)\)
  - if \(e^* = e_{N+2}\), then \(p(s' ; s, e^*) = 1\)
    when \(s = (z_1, \ldots, z_{j-1}, 1, z_{j+1}, \ldots, z_N, 0, j)\) with \(j < N\) and \(s' = (z_1, \ldots, z_{j-1}, 0, z_{j+1}, \ldots, z_N, j, j+1)\);
    when \(s = (z_1, z_2, \ldots, z_{N-1}, 1, 0, N)\) and \(s' = (z_1, z_2, \ldots, z_{N-1}, 0, 0, N)\);
    when \(s = (z_1, \ldots, z_{j-1}, 0, z_{j+1}, \ldots, z_N, 0, j)\) with \(j < N\) and \(s' = (z_1, \ldots, z_{j-1}, 0, z_{j+1}, \ldots, z_N, 0, j+1)\);
    and when \(s = (z_1, z_2, \ldots, z_{N-1}, 0, 0, N)\) and \(s' = (z_1, z_2, \ldots, z_{N-1}, 0, 0, 1)\)
  - exercise: do the case \(e^* = e_{N+1}\)
- \(F(x; s', e^*, s, e^*)\) is defined as follows
  - if \(e^* = e_j (1 \leq j \leq N)\), then \(F(x; s', e^*, s, e^*) = P\{L_j \leq x\}\)
if $e' = e_{N+1}$ and $s' = (z_1, \ldots, z_N, m, n)$ then $F(x; s', e', s, e^*) = P\{R_m \leq x\}$

if $e' = e_{N+2}$ and $s' = (z_1, \ldots, z_N, 0, n)$ then $F(x; s', e', s, e^*) = \mathbb{1}_{[0,x]}(W_{n-1})$ if $n > 1$ and $\mathbb{1}_{[0,x]}(W_N)$ if $n = 1$

- $r(s, e) = 1$ for all $s$ and $e$
- initial dist' n: $v(s) = 1$, where $s = (0,1,1,\ldots,1,0,2)$, $F_0(x; e_1, s) = P\{L_1 \leq x\}$ and $F_0(x; e_{N+2}, s) = \mathbb{1}_{[0,x]}(W_1)$