Proceedings of the 1983
Winter Simulation Conference
S. Roberts, J. Banks, B. Schmeiser (eds.)

ON THE ROLE OF GENERALIZED SEMI-MARKOV PROCESSES IN SIMULATION OUTPUT ANALYSIS

Peter W. Glynn
Industrial Engineering
1513 University Avenue
Madison, WI 53706

A generalized semi-Markov process (GSMP) is a stochastic process description of a large class of discrete-event simulations. GSMP's are defined, and some basic properties of GSMP's are described. It is argued that GSMP's provide a convenient frame-work in which to analyze many questions of interest to practitioners.

1. INTRODUCTION

Consider the discrete-event simulation of a single-server first-come first-serve GI/G/I queueing system. Assuming that the customer inter-arrival and service times form sequences of independent and identically distributed (i.i.d.) continuous random variables (r.v.'s), a sample realization of the queue is generated by the following algorithm:

1. Set QUEUE = 0, ACLK = 0, SCLK = 0, Q = 0, T = 0.
2. Generate a r.v. A with the inter-arrival distribution. Set T = A.
3. If QUEUE > 0, go to 5.
4. Generate r.v.'s A,S with the inter-arrival and service time distributions, respectively. Set ACLK = A, SCLK = S, QUEUE = 1.
5. If ACLK > SCLK, go to 7.
6. Put T = T + ACLK, Q = Q + QUEUE * ACLK, SCLK = SCLK - ACLK, QUEUE = QUEUE + 1. Generate A from the inter-arrival distribution. Set ACLK = A. Go to 5.
7. Put T = T + SCLK, Q = Q + QUEUE * SCLK, ACLK = ACLK - SCLK, QUEUE = QUEUE - 1. Generate S from the service distribution. Set SCLK = S. Go to 3.

Note that the steady-state queue-length of the GI/G/I queue under study can be estimated by Q/T. The key to the above simulation is the use of "clocks" (ACLK, SCLK) which record the time to the next event (either an arrival or departure). When an event occurs, the state changes (QUEUE is incremented or decremented by 1), and the appropriate clock is re-set.

As is well known in the simulation community, a wide variety of discrete-event simulations can be handled in a similar manner. One specifies a set of states and a set of events. The process of interest is then simulated by running clocks corresponding to all possible events in the current state. When a clock runs down to zero, a state transition occurs and the clock associated with the trigger event is re-set; the remaining clocks continue to run down. This procedure is then repeated indefinitely.

In this paper, we shall discuss the class of stochastic processes corresponding to discrete-event simulations of the above type--the probability literature refers to such simulations as generalized semi-Markov processes (GSMP's). In Section 2, we shall give a formal definition of a GSMP; Section 3 discusses certain ergodic properties of GSMP's. Section 4 is devoted to a class of structural theorems for GSMP's known as insensitivity results--these theorems show that GSMP's possess a surprising amount of structure. In Section 5, we present some concluding remarks.

2. DESCRIPTION OF A GENERALIZED SEMI-MARKOV PROCESS

As indicated above, a GSMP is a probabilist's terminology for a discrete-event simulation of the type described in Section 1. It should therefore be no surprise that the basic components of a GSMP are a set S of states and E of events, together with a family of clocks corresponding to the events. We proceed now to formally define a GSMP; we will then show how the GI/G/I queue can be viewed as a special case.

We start by identifying the sets S and E with non-empty subsets of the positive integers; elements $s \in S$ are regarded as states, and elements $i \in E$ as possible events. For each state $s \in S$, E(s) is that subset of E consisting of events possible in s. For each state

CH1953-9/83/0000-0039 $01.00

$s \in S$, let $C(s)$ be the set of all possible clock readings in s:

$$C(s) = \{c \in \mathbb{R}^{|E|} : c_i > 0 \text{ iff } i \in E(s)\}$$

(a component of c is positive if and only if the clock corresponding to that component is "active" in state s).

We shall define the GSMP by first constructing a Markov chain (M.C.) $X_n = (S_n, C_n)$, which records the values of the state variable and clock readings at successive transition epochs. Clearly, the state space of X_n is given by

$$\Sigma = \bigcup_{s \in S} (\{s\} \times C(s)).$$

For a given element $(s,c) \in \Sigma$, we let

$$t^* = t^*(s,c) = \min\{c_i : i \in E(s)\}$$

$$c_i^* = c_i^*(s,c) = c_i - t^*$$

$$i^* = i^*(s,c) = \min\{i \in E(s) : c_i^* = 0\};$$

hence, t^* is the amount of time until the first clock i^* in $E(s)$ runs down. When the clock i^* runs down to zero, a state transition occurs; state s' is chosen with probability $p(s';s,i^*)$. Some of the clocks in s, denoted by the set $O(s',s,i^*)$ (the "old" clocks) continue to run down in state s'; the rest of the clocks $j \in E(s')$, denoted $N(s',s,i^*)$ (the "new" clocks) will have their clock values generated independently from distributions $F(\cdot;s',j,s,i^*)$. It is assumed that $F(0;s',j,s,i^*) = 0$ and that $i^* \notin O(s',s,i^*)$. The transition kernel P of $\{X_n : n \geq 0\}$ can now be defined:

$$P((s,c),A) \triangleq P\{X_{n+1} \in A \mid X_n = (s,c)\}$$

$$= p(s';s,i^*) \prod_{j \in N_{s'}} F(a_j;s',j,s,i^*) \prod_{i \in O_{s'}} I_{[0,a_i]}(c_i^*)$$

where $A = \{s'\} \times \{c' \in C(s') : c'_i \leq a_i, i \in E(s')\}$, and I_B is the indicator function of the set B. The M.C. $\{X_n : n \geq 0\}$ just constructed is called the generalized semi-Markov ordered pair (GSMOP). Finally, the GSMP $\{X(t) : t \geq 0\}$ associated with $\{X_n : n \geq 0\}$ is obtained by setting

$$X(t) = \sum_{n=0}^{\infty} S_n I_{[T_n, T_{n+1}]}(t)$$

where $T_0 = 0$, $T_n = \sum_{k=0}^{n-1} t^*(S_k, C_k)$.

(2.1) Example (GI/G/1 queue): $S = \{0,1,2,\ldots\}$ (states = queue - length), $E = \{0,1\}$ (0=arrival, 1=departure), $E(0) = \{0\}$, $E(n) = \{0,1\}$ (for $n \geq 1$), $p(n+1;n,0) = 1$, $p(n-1;n,1) = 1$ (for $n \geq 1$), and

$$F(x;s',0,s,i^*) = P\{A \leq x\}$$

$$F(x;s',1,s,i^*) = P\{D \leq x\}$$

where A and D have the inter-arrival and service time distributions, respectively.

As Example 2.1 indicates, GSMP's arise naturally in the modelling of queueing systems. At the expense of some additional notation, Example 2.1 can easily be extended to include networks of queues involving complicated priority schemes. For certain purposes, however, it turns out to be convenient to allow the clocks to run down at different speeds in different states; this is necessary, for example, in systems possessing interruptible components. For a discussion of GSMP's with speeds, see (Burman, 1981), (Hordijk, 1980), (Schassberger, 1978), (Whitt, 1980). Nevertheless, our above class of GSMP's is sufficient for most modelling purposes. Furthermore, much of our discussion in Sections 3 and 4 can be extended to include GSMP's with speeds.

3. ERGODIC THEORY FOR GSMP'S

Although GSMP's appear to possess little of the structure common to elementary stochastic processes such as finite-state M.C.'s, it turns out that much of the ergodic theory for finite-state M.C.'s carries over to GSMP's. This suggests that the theoretical development of output analysis for GSMP's should be mathematically tractable, and that many of the results currently known for finite-state M.C.'s should possess GSMP analogues.

We first need to review the basic theory for continuous time M.C.'s. We recall that when the state space is countably infinite, such processes may not have limiting distributions; in fact, they may even explode in finite time. Thus, in order to guarantee the existence of limiting distributions, it is necessary to impose a finite state space assumption. For a finite-state M.C., limiting distributions always exist--however, the distribution may depend on the initial state.

Turning now to GSMP's, assume that A.) $|S| < \infty$, $|E| < \infty$; B.) there exists $\infty > K(s',j,s,i) > 0$ such that $F(K(s',j,s,i);s',j,s,i) = 1$ for every 4-tuple (s',j,s,i).

(3.1) Theorem: i.) Under A.) and B.), $\{X_n : n \geq 0\}$ has a limiting distribution. ii.) If, in addition, the distributions $F(\cdot;s',j,s,i)$ are all continuous, then $\{X_n : n \geq 0\}$ has an invariant distribution.

Proof: We metrize Σ as follows:

$$p((s',c'),(s^2,c^2)) = \begin{cases} 1 & \text{if } s' \neq s^2, \text{ or if} \\ & \text{there exists } i,j \in E \text{ such that} \\ & (c_i^1 - c_j^1)(c_i^2 - c_j^2) < 0 \\ & \sum_{i=1}^{|E|} |c_i^1 - c_i^2|; \text{ else} \end{cases}$$

Thus, two points $(s^1,c^1),(s^2,c^2)$ are at unit distance unless they share the same state component, and all clocks are in the same order. Now, observe that under B.), X_n is restricted to a compact subset of Σ, and, hence, by Prohorov's theorem (Billingsley, 1968, p. 37), the measures

$$\left(\sum_{k=1}^{n} P\{X_k \in \cdot \mid X_0 = x\}\right)/n$$

are relatively compact; that is, there exists a probability π such that

$$(3.2)\quad \frac{1}{n_k}\sum_{j=1}^{n_k} P\{X_j \in \cdot \mid X_0 = x\} \Rightarrow \pi(\cdot)$$

($\Rightarrow$ denotes weak convergence)--in other words, X_n has a limiting distribution.

For ii.), observe that with our metric on Σ, $P(\cdot,x_n) \Rightarrow P(\cdot,x)$ whenever $x_n \to x$, by the continuity of the F's. Thus, for any real-valued bounded continuous function f,

$$(3.3)\quad (Pf)(x_n) \triangleq E\{f(X_1) \mid X_0 = x_n\} \to (Pf)(x),$$

when $x_n \to x$ ie. $(Pf)(\cdot)$ is continuous. Now, by the definition of weak convergence, (3.2) implies that

$$(3.4)\quad \frac{1}{n_k}\sum_{j=1}^{n_k} (P^j f)(x) \to \int f(y)\pi(dy)$$

($(P^j k)(x) \triangleq E\{f(X_j) \mid X_0 = x\}$) for any bounded continuous f. But by (3.3), Pf is also continuous so (3.4) also holds with Pf substituted for f; thus

$$\int f(y)\pi(dy) = \int (Pf)(y)\pi(dy)$$

ie. π is invariant.||

(3.5) Corollary: Let $g:S \to \mathbb{R}$, and suppose that $P\{X(0)\in\cdot\} = \pi(\cdot)$. Then, under the conditions of Theorem 3.1 ii.), there exists a r.v. $\mathcal{Z}$ s.t.

$$(3.6)\quad \frac{1}{t}\int_0^t g(X(s))ds \to \mathcal{Z} \text{ a.s.}$$

Proof: By Birkhoff's ergodic theorem (see Lamperti, 1977, p. 92), there exist $\mathcal{Z}_1$, $\mathcal{Z}_2$ such that

$$\frac{1}{n}\sum_{k=0}^{n-1} t^*(S_k,C_k) = T_n \to \mathcal{Z}_1 \text{ a.s.}$$

$$\frac{1}{n}\sum_{k=0}^{n-1} g(S_k)\cdot t^*(S_k,C_k) \to \mathcal{Z}_2 \text{ a.s.}$$

from which it follows that

$$(3.7)\quad \frac{1}{T_n}\int_0^{T_n} g(X(s))ds \to \mathcal{Z}_2/\mathcal{Z}_1 = \mathcal{Z} \text{ a.s.};$$

(3.6) is obtained from (3.7) by an approximation argument.||

Theorem 3.1 and its corollary indicate that one need impose only mild conditions in order to obtain existence of limiting and invariant distributions (in fact, B.) can be relaxed to requiring finite means for the F's; see (Whitt, 1980). A.) is satisfied by most GSMP models for closed queues). However, we emphasize that, in general, the limit $\mathcal{Z}$ in (3.6) is a r.v. which generally depends on the sample realization through X(0)--this means, in analogy with the M.C. case, that the state space Σ is not irreducible.

It is important to realize that this is not merely a mathematical problem--lack of irreducibility is a very important topic from a practitioner's viewpoint. It is important for a simulator to be aware of the fact that the system's long-run behavior may depend critically on the initial condition used, and to take appropriate action. For example, irreducibility would be violated in a simulation model of a queue which becomes "deadlocked" under some initial conditions, but not others. For such a model, questions of irreducibility are clearly of practical import.

We conclude this section with a summary; we showed, via Theorem 3.1 and its corollary, that existence of invariant distributions is rather simple to analyze. Our subsequent discussion showed that the critical question, from a simulation view-point, concerns the irreducibility of the system.

4. INSENSITIVITY THEORY FOR GSMP'S

It is well known that the limiting distribution of the M/G/∞ queue-length process depends on the service time distribution only through its mean; in other words, the limit distribution is "insensitive" to the detailed form of the service times. Over the last ten years, a number of papers have investigated the extent to which this insensitivity extends to discrete-event simulations of the GSMP type; see (Burman, 1981),(Helm, 1982),(Schassberger, 1977),(Schassberger, 1978a), and (Schassberger, 1978b).

These papers show that insensitivity holds for a broad class of GSMP's--for such GSMP's, it is clear that one should estimate the steady-state distribution by replacing the lifetime distributions F with distributions F', having the same mean, but better simulation properties (eg. exponentials or constant r.v.'s). Broadly speaking, the general form of the results is that insensitivity holds for a GSMP provided certain mass-balance relationships are satisfied.

5. CONCLUDING REMARKS

We saw, in Section 1 and 2, that GSMP's are a probabilistic description of a certain class of discrete-event simulations. In Section 3, it was shown that GSMP's often settle down to a steady-state; however, the steady-state may depend on the initial condition. As was indicated there, much work remains to be done on the ergodic theory for GSMP's particularly in terms of conditions guaranteeing irreducibility. In Section 4, the insensitivity theory for GSMP's was briefly described--these results suggest that GSMP's enjoy a certain degree of mathematical tractability.

The central theme of this paper, however, is that GSMP's provide a basic framework in which to analyze an enormous class of simulations. We conclude with a partial list of advantages deriving from such a framework:

1) the possibility of a comprehensive ergodic theory for discrete-event simulations (partially illustrated in Section 3)

2) development of a coherent framework in which to consider statistical analysis issues for discrete-event simulations

3) the possibility of developing a comprehensive and unified methodology for variance reduction

4) the possibility of systematic integration of sophisticated statistical analysis tools into simulation software packages.

ACKNOWLEDGEMENT

The author wishes to gratefully acknowledge support of the Graduate School of the University of Wisconsin at Madison, and of the United States Army under Contract No. DAAG29-80-C-0041.

REFERENCES

Billingsley, P. (1968), Convergence of Probability Measures, John Wiley and Sons, New York.

Burman, D.Y. (1981), Insensitivity in queueing systems, Advances in Appl. Probability.

Helm, W.E. and Schassberger, R. (1982), Insensitive generalized semi-Markov schemes with point process input, Math. Oper. Res., 7, 129-138.

Hordijk, A. and Schassberger, R. (1982), Weak convergence for generalized semi-Markov processes, Stochastic Processes Appl., 12, 271-292.

Lamperti, J. (1977), Stochastic Processes, Springer-Verlag, New York.

Schassberger, R. (1977), Insensitivity of steady-state distributions of generalized semi-Markov processes, I, Ann. Probability, 5, 87-99.

Schassberger, R. (1978a), Insensitivity of steady-state distributions of generalized semi-Markov processes, II, Ann. Probability, 6, 85-93.

Schassberger, R. (1978b), Insensitivity of steady-state distributions of generalized semi-Markov processes with speeds, Advances in Appl. Probability, 10, 836-851.

Whitt, W. (1980), Continuity of generalized semi-Markov processes, Math. Oper. Res., 5, 494-501.