MONTE CARLO SIMULATION OF SOME FINITE-STATE MARKOVIAN MODELS

P.K. Pollett

The University of Queensland

Abstract. We shall describe an acceptance-rejection method of sampling from the equilibrium distribution of a finite-state, continuous-time Markov process. The method applies to any process that can be viewed as a truncation of another Markov process which exhibits partial balance. These include, for example, reversible and dynamically reversible processes, and some quasireversible processes. We shall give particular attention to models of certain chemical processes, telecommunications networks and queueing systems.

1. Introduction

A wide variety of Markov processes which are used in modelling stochastic systems exhibit the property of partial balance. In most cases this simplifies their analysis. For example, one usually finds that the equilibrium distribution admits a product form (see, for example, Jansen and König [12]) and that this is insensitive to variations in the values of certain parameters of the system (see, for example, Whittle [36], [37]). The class of processes which exhibit partial balance include reversible processes (see, for example, Kelly [15]), dynamically reversible processes (see Whittle [35]), symmetric queues (see Kelly [14]) and networks of quasireversible nodes (see, for example, Kelly [16] or Henderson, et al. [10]). In the sequel we shall be concerned with finite-state processes. These are often used to model closed systems, for example queueing networks with no exogenous arrival or departure streams, or indeed any queueing system with restrictions of the numbers of customers (see, for example, Kelly [15]), models for simple chemical reactions (see, for example, McQuarrie [22]) and models for circuit-switched telecommunications systems (see, for example, Burman et al. [1]). Although the equilibrium distributions of the processes cited have a pleasingly simple product form, it is usually impossible to write down an explicit expression for the normalizing constant, and, thus, it is of little comfort to the practitioner to know that many quantities of interest, for example measures of performance, can be expressed in terms of this constant. However, a good deal of attention has been given to finding efficient numerical methods for evaluating the normalizing constant (see, for example, Buzen

1991 *Mathematics Subject Classification.* 62M05; 60J27 .

Key words and phrases. Markov processes, Partial balance, Monte Carlo simulation .

[2] and Reiser [30]) and these are used widely. Another approach has been to provide approximations to the equilibrium distribution (see, for example, McKenna and Mitra [21] in the context of queueing networks, Dunstan and Reynolds [8] in the context of chemical kinetics and Kelly [17] in the context of telecommunications networks); the accuracy of these methods often increases as the complexity or the size of the system grows. They are particularly useful in cases when the normalizing constant can be written down explicitly, but the state probabilities are difficult to handle from a computational point of view. For example, the equilibrium distribution for the numbers of molecules in simple chemical reactions are usually expressed in terms of hypergeometric functions (see Darvey, Ninham and Staff [6]), which are very difficult to handle, yet the Gaussian approximations thereof are simple to use and very accurate indeed (see Pollett and Vassallo [29]). Yet another approach is to use computer simulation. Of course simulation is an extremely useful tool for analysing models which are analytically intractable. However, even when explicit formulae are available, there are instances where it is quicker, and indeed far simpler, to obtain point estimates and confidence limits, for quantities of interest, using simulation.

One important simulation method that has emerged in recent times is the *regenerative method* (see, for example, Crane and Lemoine [5]). This is now used widely as an alternative to methods which involve the use of *warm-up time*, because usually only one simulation run is needed and measurements can be taken right from the beginning of the run. However, if regeneration points are scarce, one has to seek an alternative method. This is often the case when the system in question, or, more precisely, the number of states is very large, for example if one is simulating models for chemical processes, telecommunications systems, or queueing networks with moderately large traffic intensities.

We shall describe an alternative simulation method which can be used in these instances. It is ostensibly different from *real-time* simulation, of which the regenerative method provides an example, in that it involves sampling directly from the equilibrium distribution.

2. Markov processes and partial balance

Full balance and detail balance. We shall consider a continuous-time Markov process which takes values in a countable state space, S. We shall suppose that it has a stable, conservative and regular q-matrix of transition rates, $\mathbf{Q} = [q(x,y),\ x,y \in S]$, where, for convenience, $q(x,x)$ is set to zero for each $x \in S$, and that it admits a unique equilibrium distribution, that is, a collection, $\mathbf{m} = (m(x),\ x \in S)$, of positive numbers which sum to unity and satisfy the *full-balance* equations

$$m(x)q(x) = \sum_{y\in S} m(y)q(y,x), \qquad x \in S, \tag{1}$$

where

$$q(x) = \sum_{y\in S} q(x,y)$$

is the rate out of state x, for $x \in S$. If $\mathbf{m}$ satisfies the *detail-balance* equations

$$(2) \qquad m(x)q(x,y) = m(y)q(y,x), \qquad x,y \in S,$$

then the process is said to be *reversible*, for it then has the same finite dimensional distributions under time reversal (see Kelly [15]). Indeed, if one can find a collection of positive numbers, $\mathbf{m} = (m(x),\ x \in S)$, which sum to unity and which satisfy (2), then, on summing (2) over $y \in S$, we see that $\mathbf{m}$ must be the equilibrium distribution.

Probability flux and partial balance. The notion of partial balance (or local balance) lies, as it were, part way between (1) and (2). Although it is possible to provide a very general formulation of partial balance (see Pollett [27]), we shall restrict our attention to the very specialized form used by Whittle [36] in his study of insensitivity. We shall find it convenient to use Whittle's notation: if A and B are subsets of S, then let

$$[A,B] = \sum_{x \in A} \sum_{y \in B} m(x)q(x,y);$$

this quantity is known as the *probability flux* from A to B under $\mathbf{m}$. Using this notation, (1) can be rewritten as

$$(1') \qquad [S,x] = [x,S], \qquad x \in S,$$

and (2) can be rewritten as

$$(2') \qquad [x,y] = [y,x], \qquad x,y \in S.$$

It is sometimes said that the detail-balance equations stipulate that the probability fluxes between any two states be balanced, while the full-balance equations require only that the probability flux out of x be balanced with that into x, for each state x. If, for each state x, the probability flux from x to some *subset*, A, of S is balanced with that from A into x, that is,

$$(3) \qquad [A,x] = [x,A],$$

then we say that $\mathbf{m}$ is *partially balanced over (or with respect to) A*. One should note that, in view of (1′), equation (3) holds if and only if

$$[A^c,x] = [x,A^c], \qquad x \in A,$$

where A^c is used to denote the complement of A in S.

Truncation. A Markov process is said to be *truncated* to a set $A \subseteq S$ if $q(x,y)$ is changed to zero for each $x \in A$ and $y \in A^c$. The following result (Kelly [15], Exercise 1.6.2) provides the basis for the Monte Carlo technique which we shall describe in Section 3:

Lemma 1. *If the Markov process is truncated to a subset, A, of S, then the equilibrium distribution, $\mathbf{p} = (p(x),\ x \in A)$, of the truncated process is given by*

$$p(x) = \frac{m(x)}{\sum_{y \in A} m(y)}, \qquad x \in A, \tag{4}$$

if and only if $\mathbf{m}$ *is partially balanced over A.*

Remarks. (i) If $\mathbf{m}$ is partially balanced over A, then the equilibrium probability that the truncated process is in state $x(\in A)$ can be interpreted as the conditional probability that the original process is in state x, given that it is somewhere in A.

(ii) The condition that $\mathbf{m}$ be partially balanced over A amounts to a requirement that the measure $(m(x),\ x \in A)$ be invariant for $\mathbf{Q}_A$, the q-matrix defined to be the restriction of $\mathbf{Q}$ to the set A.

(iii) A sufficient condition for the truncated process to have the equilibrium distribution specified by (4) is that $\mathbf{m}$ be detail balanced, that is, the original process be reversible; clearly the truncated process must then be reversible also.

3. The method

The idea is to imagine that the process, $\mathcal{Y} = (Y(t),\ t \geq 0)$, which we wish to simulate, is a truncation of another process, $\mathcal{X} = (X(t),\ t \geq 0)$, the simulation of which can be done simply. Suppose that $\mathcal{Y}$ has the finite state space A and that its equilibrium distribution is $\mathbf{p} = (p(x),\ x \in A)$. If we were able to sample from $\mathbf{p}$, we could produce a sequence of independent observations, $y_1, y_2, \ldots, y_n$, of the state of the process and, then, this could be used to provide estimates of any quantities of interest. However, our premise is that one cannot easily sample from $\mathbf{p}$.

We shall assume (i) that there exists a process, $\mathcal{X} = (X(t),\ t \geq 0)$, taking values in $S \supseteq A$, of which $\mathcal{Y}$ is a truncation, (ii) that its equilibrium distribution, $\mathbf{m} = (m(x),\ x \in S)$, is partially balanced over A and (iii) that sampling from $\mathbf{m}$ is a straightforward matter; of necessity, $m(x)$ will be proportional to $p(x)$, for $x \in A$. If one repeatedly samples from $\mathbf{m}$, thus producing a sequence of independent observations, $x_1, x_2, \ldots,$ of $\mathcal{X}$, then, by virtue of Lemma 1, the first n observations, $y_1, y_2, \ldots, y_n$, that lie in A will comprise a random sample from $\mathbf{p}$. We therefore propose the following algorithm:

```
{obtain sample points, $y_1, y_2, \ldots, y_n$, from $\mathbf{p}$}
for $i := 1$ to $n$ do
    begin
        repeat
            obtain a sample point, $x$, from $\mathbf{m}$
        until $x \in A$;
        $y_i := x$
    end.
```

The algorithm we have described is one of the class of *acceptance-rejection* algorithms, known as such because, at each stage, sampling continues until a sample point is "accepted" according to some criterion (for an account of these methods, see Rubinstein [31]).

The form chosen for $\mathbf{m}$ will depend on the particular process $\mathcal{Y}$ and the form of its equilibrium distribution. It turns out that little ingenuity is required to obtain $\mathbf{m}$ in any given instance, although it is not possible to provide a general method for determining $\mathbf{m}$ which applies in all situations. There is, however, one important criterion that should be met: $\mathbf{m}$ should be chosen so that the probability of "acceptance" (call it q) is as large as possible. Clearly q is given by $q = \sum_{x \in A} m(x)$ and the expected number of sample points from $\mathbf{m}$ needed to produce one sample point from $\mathbf{p}$ is q^{-1}; q is aptly called the *efficiency* of the algorithm.

In the most common case, where $\mathbf{p}$ admits a product form, say

$$p(\mathbf{x}) = B \prod_{r=1}^{N} p_r(x_r), \qquad \mathbf{x} \in A,$$

where $N \geq 2$, $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ and B is the normalizing constant, an obvious choice for $\mathbf{m}$ is given by

$$m(\mathbf{x}) = \prod_{r=1}^{N} m_r(x_r), \qquad \mathbf{x} \in S,$$

where S is a product space with $A \subseteq S$ and which can be written as $S = S_1 \times S_2 \times \cdots \times S_N$ and, for each r, $\mathbf{m}_r = (m_r(x_r),\ x_r \in S_r)$ is a probability distribution over S_r. Thus a sample point obtained from $\mathbf{m}$ will consist of *independent* observations from the distributions $\mathbf{m}_1, \mathbf{m}_2, \ldots \mathbf{m}_N$. In addition, it is commonly the case that there exists a set I and a sequence, $A_1, A_2, \ldots, A_N$, with $A_k \subseteq I^k$, $k = 1, 2, \ldots, N$, and $A_N = A$, such that $\mathbf{x} \in A$ if and only if $x_1 \in A_1$, $(x_1, x_2) \in A_2$, $\ldots$, $(x_1, x_2, \ldots, x_N) \in A_N$. In these cases, the basic acceptance-rejection algorithm can be made computationally more efficient:

```
{obtain sample points, $y_1, y_2, \ldots, y_n$, from $\mathbf{p}$}
for $i := 1$ to $n$ do
    begin
        for $j := 1$ to $N$ do
            repeat
                obtain a sample point, $x_j$, from $m_j$
            until $(x_1, x_2, \ldots, x_j) \in A_j$;
        $y_i := (x_1, x_2, \ldots, x_N)$
    end.
```

4. Applications

In this section we shall consider a number of applications of the method to stochastic models of telecommunications networks, migration processes and queueing systems, and chemical processes.

A telecommunications network. One important problem in teletraffic theory is the determination of grades of service or blocking probabilities, as these provide a good measure of performance of telecommunications networks. However, they are usually very difficult to calculate for networks of reasonable size. We shall explain how the method we have described can be used to estimate blocking probabilities in a circuit-switched network.

Let us begin with a brief description of the model. Calls emanate from various localities that are interconnected by groups of circuits, which we shall call links. If there are J links, then any route in the network can be expressed as a subset of $\{1,2,\ldots,J\}$. Let $\mathcal{R}=\{1,2,\ldots,N\}$ index the collection of all possible routes and suppose that calls using route $r\in\mathcal{R}$ require $a_{jr}(\geq 0)$ circuits from link j and that these are held for the duration of any such call; route r is then the collection of links, j, for which a_{jr} is positive. The blocking probability for route r is the probability that a request for transmission on that route cannot be met owing to the unavailability of sufficiently many circuits, that is, if there are fewer than a_{jr} circuits free on link j, for some j. We shall assume that requests for various routes form independent Poisson processes with rates $(c_r,\ r\in\mathcal{R})$ and that call lengths have an arbitrary distribution with mean 1. If we denote by $\mathbf{n}=(n_r,\ r\in\mathcal{R})$ a typical state of the network, where n_r is the number of calls in progress on route r, then the set of all states is given by

$$A_{\mathbf{c}}=\{\mathbf{n}\in\{0,1,\ldots\}^{\mathcal{R}}:\mathbf{A}\mathbf{n}\leq\mathbf{c}\},$$

where $\mathbf{A}$ is the matrix $[a_{jr},\ j=1,2,\ldots,J,\ r\in\mathcal{R}]$ and $\mathbf{c}$ is the vector $(c_j,\ j=1,2,\ldots,J)$, where c_j specifies the total number of circuits of link j. The transition rates of the process $(\mathbf{n}(t),\ t\geq 0)$ have an extraordinarily simple form, since the only possible transitions involve either an upward or a downward jump of size 1 in one, and only one, component of $\mathbf{n}$. It can be shown (see Burman et al. [1]) that the unique equilibrium distribution, $\mathbf{p}=(p(\mathbf{n}),\ \mathbf{n}\in A_{\mathbf{c}})$, is given by

$$p(\mathbf{n})=B_{\mathbf{c}}\prod_{r=1}^{N}\frac{c_r^{n_r}}{n_r!},\qquad \mathbf{n}\in A_{\mathbf{c}},$$

where $B_{\mathbf{c}}$ is a normalizing constant, and that $\mathbf{p}$ is detail balanced. It is easy to see that the process is a truncation of a process for which there is no upper limit on the number of circuits on each of the links. This process has the property that the equilibrium numbers of calls, $n_r,\ r\in\mathcal{R}$, are independent Poisson random variables with n_r having mean c_r. Thus, in order to generate a random state of the network, one can simply generate values, $n_r,\ r\in\mathcal{R}$, of these random variables until $\mathbf{n}=(n_r,\ r\in R)$ lies in $A_{\mathbf{c}}$. We can therefore use the following algorithm to obtain a random sample of the state of the network:

```
begin {obtain sample points, $\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_n$, from $\mathbf{p}$}
    $i := 0$;
    repeat
        $i := i + 1$;
        repeat
            {generate a value of $\mathbf{n} := (n_r,\ r = 1, 2, \ldots, N)$}
            $r := 0$;
            repeat
                $r := r + 1$;
                {generate the value of a Poisson random variable with mean $c_r$;
                 see Devroy [7] for details}
                $n_r := Poisson(c_r)$
            until $r = N$
        until $\mathbf{n} \in A_{\mathbf{c}}$;
        $\mathbf{n}_i := \mathbf{n}$
    until $i = n$
end.
```

If, as is usually the case, M, given by

$$M = \max_{1 \le j \le J} c_j$$

is quite small, the method can be made somewhat more efficient by replacing the Poisson random variables by ones with the *truncated* Poisson distribution, $(p_r(n_r),\ n_r = 0, 1, \ldots, M)$, given by

$$p_r(n_r) = \frac{c_r^{n_r}}{n_r!} \left(\sum_{l=0}^{M} \frac{c_r^l}{l!} \right)^{-1}, \qquad n_r = 0, 1, \ldots M.$$

It is known as the *individual state selection method* (see Harvey and Hills in [9] or Pollett [26], [28]).

Once we have obtained the sample $\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_n$, we can estimate the blocking probability, g_r, for route r, using the usual sample-mean estimator

$$\hat{g}_r = \frac{1}{n} \sum_{i=1}^{n} 1_{ri},$$

where

$$1_{ri} = \begin{cases} 0, & \text{if } \mathbf{A}\mathbf{n}_i \le \mathbf{c} - \mathbf{A}\mathbf{e}_r, \\ 1, & \text{otherwise;} \end{cases}$$

here $\mathbf{e}_r$ is the unit vector with a 1 in the r^{th} position. Confidence intervals for g_r can then be obtained in the usual way (for details see Pollett [28]).

Migration processes and queueing networks. The sampling method we have described can also be used to estimate quantities of interest in a variety of systems which involve the flow and the interaction of units and where there are constraints on the numbers of those units. These include, for example, models for describing the migration of birds (see Whittle [33] and [34]), machine interference (Cox and Smith [4]) and road traffic (see Kelly [15]), models for computer networks (see Kelly [15]) and message/packet-switched communications networks (see, for example, Pollett [25]), and compartmental models (see Matis and Hartley [20]). The model we shall consider encapsulates those cited. We shall provide only a brief description and refer the reader to Kingman [18] or Pollett [24] for details.

We shall suppose that there are N colonies (queues) labelled $1, 2, \ldots, N$. For simplicity we shall suppose that a typical state of the process is $\mathbf{n} = (n_1, n_2, \ldots, n_N)$, where n_r is the number of individuals in colony (queue) r, although a richer definition of the state of the process is sometimes required. The state space, A, of the process will depend on the precise functional relationship between the numbers of individuals in each of the colonies. When the state of the process is $\mathbf{n}$, the rate at which a migration occurs from colony r to colony s is $\lambda_{rs}\phi_r(n_r)\psi_s(n_s)$, where $\lambda_{rs} \geq 0$, $\phi_r(n_r) > 0$ if $n_r > 0$ and $\phi_r(0) = 0$, and $\psi_s(n_s) > 0$ for all $n_s \geq 0$; the rate of immigration to colony s from outside the system is $\nu_s\psi_s(n_s)$, where $\nu_s \geq 0$, and the rate of emigration from colony r to the outside is $\mu_r\phi_r(n_r)$, where $\mu_r \geq 0$. It is usually convenient to assume that the parameters λ_{rs}, μ_r and ν_s, $r, s = 1, 2, \ldots, N$ are chosen so that A is irreducible.

If, for all s, $\psi_s(n_s) = 1$, $n_s > 0$, then the migration rates depend only on the state of the colony from which migration occurs, and immigration from outside the system occurs as independent Poisson processes; thus we obtain the migration process of Whittle [34] (see also Jackson [11]). The process always has an invariant measure, $\mathbf{m} = (m(\mathbf{n}),\ \mathbf{n} \in A)$, and, provided that there exist positive quantities $c_1, c_2, \ldots, c_N$, such that

$$c_r \left(\mu_r + \sum_{s=1}^{N} \lambda_{rs} \right) = \nu_r + \sum_{s=1}^{N} c_s \lambda_{sr}, \qquad r = 1, 2, \ldots, N,$$

$m(\mathbf{n})$ is proportional to

$$\prod_{r=1}^{N} \frac{c_r^{n_r}}{\prod_{l=1}^{n_r} \phi_r(l)},$$

for $\mathbf{n} \in A$. If there is no restriction on the numbers of individuals in the various colonies, then an equilibrium distribution exists if and only if, for each r,

$$b_r^{-1} = \sum_{n=0}^{\infty} \frac{c_r^{n}}{\prod_{l=1}^{n} \phi_r(l)}$$

is finite. When this condition is satisfied, the equilibrium distribution, $\mathbf{m} = (m(\mathbf{n}),\ \mathbf{n} \in$

S), is given by

$$(5) \qquad m(\mathbf{n}) = \prod_{r=1}^{N} m_r(n_r), \qquad \mathbf{n} \in S,$$

where $S = \{0, 1, \dots\}^N$ and $\mathbf{m}_r = (m_r(n_r),\ n_r = 0, 1, \dots)$ is the probability distribution given by

$$m_r(n) = b_r \frac{c_r^n}{\prod_{l=1}^{n} \phi_r(l)}, \qquad n = 0, 1, \dots$$

Thus, in equilibrium, the numbers, $n_1, n_2, \dots, n_N$, in the various colonies are independent random variables.

If we were to impose a constraint on the total number of individuals in the system, for example, if the process were truncated to the set A_M, given by

$$A_M = \left\{ \mathbf{n} = (n_1, n_2, \dots, n_N) : \sum_{r=1}^{N} n_r \le M \right\},$$

thus stipulating an upper limit, M, on the total number of individuals, then the equilibrium numbers would fail to be independent. However, $\mathbf{m}$ is partially balanced over A_M and so $\mathbf{p} = (p(\mathbf{n}),\ \mathbf{n} \in A_M)$, the equilibrium distribution of the truncated process, is obtained simply by renormalizing (5) over A_M. It follows that the acceptance-rejection method described above can be used to sample from $\mathbf{p}$. If $\phi_r(n_r) = n_r,\ r = 1, 2, \dots, N$, then the algorithm specified above can be used without any alteration, except that the "acceptance" condition, $\mathbf{n} \in A_{\mathbf{c}}$, should be replaced by $\mathbf{n} \in A_N$. If, as is commonly the case,

$$\phi_r(n_r) = \begin{cases} 0, & \text{if } n_r = 0, \\ 1, & \text{if } n_r > 0, \end{cases}$$

then the Poisson/truncated Poisson sampling method, which forms the core of the algorithm, should be replaced by a method of sampling from a geometric/truncated geometric distribution with parameter c_r, (see, for example, Devroy [7]).

The procedure which we have outlined can be adapted to deal with a variety of networks consisting of quasireversible nodes, for example, networks of symmetric queues. Indeed, in the latter case, the method can be used without any alteration, provided that the quantities to be estimated are functions only of the numbers of individuals $n_1, n_2, \dots, n_N$.

An example of a process which is *not* quasireversible is obtained on retaining the general form for the functions $\psi_r, r = 1, 2, \dots, N$. However, in order to guarantee the existence of a simple product-form equilibrium distribution, one needs to assume that the quantities $c_1, c_2, \dots, c_N$ satisfy

$$c_r \mu_r = \nu_r, \qquad r = 1, 2, \dots, N$$

and

$$c_r \lambda_{rs} = c_s \lambda_{sr}, \qquad r, s = 1, 2, \ldots, N.$$

In this case, $m(\mathbf{n})$ is proportional to

$$\prod_{r=1}^{N} c_r^{n_r} \prod_{l=1}^{n_r} \frac{\psi_r(l-1)}{\phi_r(l)}, \tag{6}$$

for all $\mathbf{n}$, and an equilibrium distribution exists if and only if the invariant measure $\mathbf{m}$ can be normalized. If there is no restriction on the numbers of individuals in the various colonies, then, in equilibrium, $n_1, n_2, \ldots, n_N$ are independent with n_r having the distribution $\mathbf{m}_r = (m_r(n_r),\ n_r = 0, 1, \ldots)$, given by

$$m_r(n) = b_r c_r^n \prod_{l=1}^{n} \frac{\psi_r(l-1)}{\phi_r(l)}, \qquad n = 0, 1, \ldots\ ,$$

where b_r is a normalizing constant. In fact, we find that the process is reversible, for $\mathbf{m}$ is detail balanced. This reversible form of the migration process was first studied by Kingman [18]. It allows for no net circulation of individuals among the colonies. For this reason it is suitable for use as a model for social grouping behaviour (see, for example, Kelly [15], Section 6.2).

An interesting modification of the reversible migration process is obtained on setting

$$\psi_r(n_r) = \begin{cases} 1, & \text{if } n_r = 0, 1, \ldots, k_r - 1, \\ 0, & \text{if } n_r = k_r, k_r + 1, \ldots, \end{cases}$$

for suitable constants $\mathbf{k} = (k_1, k_2, \ldots k_N)$. Thus colony r can hold at most k_r individuals. Whilst colony r is "full", transitions that would have brought an individual to that colony are forbidden; thus migration from other colonies is "blocked". Perhaps surprisingly, the equilibrium distribution is obtained by simply normalizing (6) over $A_{\mathbf{k}}$, given by

$$A_{\mathbf{k}} = \{\mathbf{n} = (n_1, n_2, \ldots, n_N) : n_r \le k_r, r = 1, 2, \ldots, N\}.$$

Therefore, the sampling algorithm needs no modification, only that the "acceptance" condition should be replaced by $\mathbf{n} \in A_{\mathbf{k}}$.

A clustering process. Clustering processes have been used as stochastic models in a variety of diverse contexts including the study of social grouping behaviour (see, for example, Coleman and James [3]), of aggregation of slime mould (see, for example, Keller and Segel [13]) and of chemical processes (see, for example, Whittle [32]). The version we shall consider is described, in detail, in Pollett [27]. It is a generalization of Whittle's [32] reversible clustering process. As we shall only be dealing with the equilibrium distribution, we shall provide only a brief description of the model here.

Let $\mathcal{R}$ be a countable collection of cluster types and denote by n_r the number of clusters of type r. Thus, a typical state, $\mathbf{n} = (n_r,\ r \in \mathcal{R})$, of the process, indicating the numbers of clusters of each type, will be an element of S, the subset of $\{0, 1, \dots\}^{\mathcal{R}}$ whose elements have only finitely many non-zero entries. The parameters of the process are non-negative constants $\lambda_{rsu}(= \lambda_{sru}), \mu_{rsu}(= \mu_{sru})$ and γ_{ru}, where $r, s, u \in \mathcal{R}$ and non-negative, real-valued functions $\phi_r,\ r \in \mathcal{R}$. The quantity μ_{rsu} is the rate at which a given type u cluster divides into one of type r and one of type s, and λ_{rsu} is the rate at which a given type r cluster and a given type s cluster join to form a type u cluster. The parameter γ_{ru} is the rate at which a type r cluster transmutes into a type u cluster and the function $\phi_r(= \phi_r(n_r))$ measures the extent to which the overall dissociation, association and transmutation rates are affected by the numbers of clusters of the types involved.

It can be shown (Kelly [15], Exercise 8.5.2) that if there exist positive quantities, $(c_r,\ r \in \mathcal{R})$, satisfying

$$c_r c_s \sum_{u \in \mathcal{R}} \lambda_{rsu} = \sum_{u \in \mathcal{R}} c_u \mu_{rsu}, \qquad r, s \in \mathcal{R}$$

and

$$\sum_{r \in \mathcal{R}} \sum_{s \in \mathcal{R}} c_r c_s \lambda_{rsu} + \sum_{r \in \mathcal{R}} c_r \gamma_{ru} = c_u \left(\sum_{r \in \mathcal{R}} \sum_{s \in \mathcal{R}} \mu_{rsu} + \sum_{s \in \mathcal{R}} \gamma_{us} \right), \qquad u \in \mathcal{R},$$

then, when an equilibrium distribution exists, the equilibrium probability, $p(\mathbf{n})$, that the state of the process is $\mathbf{n}$, is proportional to

$$\prod_{r \in \mathcal{R}} \frac{c_r^{n_r}}{\prod_{l=1}^{n_r} \phi_r(l)}.$$

Conditions for the existence of an equilibrium distribution, and the precise form of that distribution, will depend on the functional relationship between the numbers of clusters of the various types. For example, if, as is the case in models for social grouping behaviour and for polymerization processes, clusters are comprised of a number of immutable units and the type of a cluster determines the number of units in that cluster, that is, $\mathcal{R} = \{1, 2, \dots, N\}$, then the state space of the process will be given by

$$A_N = \left\{ \mathbf{n} = (n_1, n_2, \dots, n_N) : \sum_{r=1}^{N} r n_r = N \right\}$$

and so an equilibrium distribution, $\mathbf{p} = (p(\mathbf{n}),\ \mathbf{n} \in A_N)$, always exists and it is given by

$$p(\mathbf{n}) = B_N \prod_{r=1}^{N} \frac{c_r^{n_r}}{\prod_{l=1}^{n_r} \phi_r(l)}, \qquad \mathbf{n} \in A_N,$$

where B_N is the normalizing constant. Further, if, as is commonly the case, the rates of dissociation, association and transmutation depend *linearly* on the numbers of clusters of the types involved, that is, $\phi_r(n_r) = n_r$, then

$$p(\mathbf{n}) = B_N \prod_{r=1}^{N} \frac{c_r^{n_r}}{n_r!}, \qquad \mathbf{n} \in A_N.$$

It is easy to show that the process is a truncation of an *open* clustering process, that is, one altered by allowing the immigration and the emigration of clusters; to obtain the transition rates of this process, one simply sets all the emigration parameters to 1 and the immigration parameter for type r clusters equal to c_r, for each $r \in \mathcal{R}$ (for details see Pollett [23]). The open process has the property that the equilibrium numbers of clusters, n_r, $r \in \mathcal{R}$, are independent random variables, in the "linear" case, independent *Poisson* random variables with n_r having mean c_r. Thus, in order to generate a random state of the process, one can simply generate values of these random variables, thus producing $\mathbf{n}$, a state of the open process, until one finds a value of $\mathbf{n}$ which lies in A_N. We can therefore use the algorithm specified above in order to obtain a random sample of the state of the linear clustering process.

As a final remark, we note that the method remains unchanged if the state space A_N is replaced by

$$A_N = \left\{ \mathbf{n} = (n_1, n_2, \ldots, n_N) : \sum_{r=1}^{N} r n_r \le N \right\}.$$

This might be appropriate if, in a model for social grouping behaviour which allows the immigration and the emigration of individuals, an upper limit, N, is placed on the population size.

5. Variance reduction

The sampling algorithm which we have described can be improved, sometimes dramatically, by employing any one of a number of standard variance reduction techniques (for a review of these methods see Kleijnen [19] or Devroy [7]). With each of these, one replaces the existing estimator of a given quantity of interest with another estimator, based on a sample of no greater size, which has a smaller variance. Thus one can obtain a tighter confidence interval without having to increase the number of sample points, or, equivalently, one can achieve the desired level of precision using a smaller sample.

In the context of estimating blocking probabilities in a circuit-switched network, the method of *antithetic variates* and the method of *stratified sampling* have been used with some success. The method of antithetic variates involves performing two simulation runs, the second of which uses a complementary sequence of random numbers. In particular, if $u_1, u_2, \ldots$ is the sequence of uniform(0,1) random numbers used in the first run, then $1-u_1, 1-u_2, \ldots$ are used in the second. If we denote by $\hat{g}_r^1$ and $\hat{g}_r^2$ the sample-mean estimators for the blocking probability g_r, each based on a sample of size n, in, respectively, the first and the second run, then the average of these, $\frac{1}{2}(\hat{g}_r^1 + \hat{g}_r^2)$, has a variance which is less than that of the original sample-mean estimator based on a sample of size $2n$, and so is used in preference. The method works because $\hat{g}_r^1$ and $\hat{g}_r^2$ are negatively correlated; this is difficult to establish in any other way but empirically. Indeed, the more negatively correlated, the greater the reduction in variance. The author has demonstrated between a 30 and a 40 percent improvement in estimating blocking probabilities in moderately large circuit-switched networks (see Pollett [28]).

The method of stratified sampling is a well-known statistical technique. It was used by Harvey and Hills [9] to estimate blocking probabilities in circuit-switched networks, and they have reported that a specified precision can be achieved 30 times faster than with the individual state selection method. They divide the collection of routes into two classes, "short" routes and "long" routes. The state space is then partitioned into classes which consist of states that indicate the same number of calls on the "long" routes. Members of the partition are selected at random using an individual state selection and then individual states in that class are selected in order to provide an estimate for g_r. A weighted average of these is then used, for this provides an estimate with a reduced variance (for details see [9]).

Another variance reduction technique which has been widely used, particularly in the context of queueing networks, is the method of *control variates.* This involves estimating a quantity whose precise numerical value is known in advance. For example, if one is simulating a single server queue, a suitable control is the probability that the server is busy, for this is known to be equal to the traffic intensity. The method has been used with varying degrees of success in simulating a wide variety of systems (see Rubinstein [31] and the references contained therein) and we expect it to be of some use in the situations considered above. However, it is difficult to see how the method could be applied to circuit-switched networks, for simple explicit formulae for *any* of the quantities of interest are unavailable.

Acknowledgements. The financial assistance of the Australian Research Council and the University of Queensland is gratefully acknowledged.

References

1. Burman, D.Y., Lehoczky, J.P. and Lim, Y., *Insensitivity of blocking probabilities in a circuit-switching network*, J. Appl. Probab. **21** (1984), 850–859.
2. Buzen, J.P., *Computational algorithms for closed queueing networks with exponential servers*, Comm. A.C.M. **16** (1973), 525–531.
3. Coleman, J.S. and James, J., *The equilibrium size distribution of freely forming groups*, Sociometry **24** (1961), 36–45.
4. Cox, D.R. and Smith, W.L., *Queues*, Methuen, London, 1961.
5. Crane, M.A. and Lemoine, A.J., *An Introduction to the Regenerative Method for Simulation Analysis, Lecture Notes in Control and Information Sciences*, vol. 4, Springer–Verlag, Berlin, Heidelberg, New York, 1977.
6. Darvey, I.G., Ninham, B.W. and Staff, P.J., *Stochastic models for second order chemical reaction kinetics. The equilibrium state*, J. Chem. Phys. **45** (1966), 2145–2155.

7. Devroy, L., *Non-uniform random variate generation*, Springer-Verlag, New York, 1986.
8. Dunstan, F.D.J. and Reynolds, J.F., *Normal approximations for distributions arising in the stochastic approach to chemical reaction kinetics*, J. Appl. Probab. **18** (1981), 263–267.
9. Harvey, C. and Hills, C.R., *Determining the grades of service in a network*, Proc. 9th Int. Teletraffic Cong. (1979).
10. Henderson, W., Pearce, C.E.M., Pollett, P.K. and Taylor, P.G., *Connecting internally balanced quasireversible Markov processes*, Adv. Appl. Probab. **24** (1992) (to appear).
11. Jackson, J.R., *Jobshop–like queueing systems*, Mgmt. Sci. **10** (1963), 131–142.
12. Jansen, U. and König, D., *Insensitivity and steady-state probabilities in product form for queueing networks*, Elektron. Informationsverarbeit. Kybernet. **16** (1980), 385–397.
13. Keller, E.F. and Segel, L.A., *Initiation of slime mould aggregation viewed as an instability*, J. Theoret. Biol. **26** (1970), 399–415.
14. Kelly, F.P., *Networks of queues*, Adv. Appl. Probab. **8** (1976), 416–432.
15. Kelly, F.P., *Reversibility and Stochastic Networks*, Wiley, Chichester, 1979.
16. Kelly, F.P., *Networks of quasi–reversible nodes*, Ed. R. Disney, Applied Probability–Computer Science, the interface: Proceedings of the ORSA–TIMS Boca Raton Symposium, Birkhauser, Boston, Cambridge, Ma, 1981.
17. Kelly, F.P., *Blocking probabilities in large circuit–switched networks*, Adv. Appl. Probab. **18** (1986), 473–505.
18. Kingman, J.F.C., *Markov population processes*, J. Appl. Probab. **6** (1969), 1–18.
19. Kleijnen, J.P.C., *Statistical Techniques in Simulation (in two parts)*, Marcel Dekker, New York, 1974/75.
20. Matis, J.H. and Hartley, H.O., *Stochastic compartmental analysis: model and least squares estimation from time series data*, Biometrics **27** (1971), 77–102.
21. McKenna, J. and Mitra, D., *Integral representations and asymptotic expansions for closed queueing networks: normal usage*, Bell System Tech. **61** (1982), 661–683.
22. McQuarrie, D.A., *Stochastic approach to chemical kinetics*, J. Appl. Probab. **4** (1976), 413–478.
23. Pollett, P.K., *Altering the q-matrix : the problem of varied arrival rates*, Eds E. A. Cousins and C. E. M. Pearce, Proceedings of the 7th National Conference of the Australian Society for Operations Research, 1985, pp. 206–234.
24. Pollett, P.K., *Connecting reversible Markov processes*, Adv. Appl. Probab. **18** (1986), 880-900.
25. Pollett, P.K., *Analysis of response times and optimal allocation of resources in message and packet switched networks*, Asia–Pacific J. Operat. Res. **3** (1986), 134–149.
26. Pollett, P.K., *A new method for estimating the performance of a communications network using simulation*, Bull. Aust. Soc. Operat. Res. **7** (1987), 6–9.
27. Pollett, P.K., *Preserving partial balance in continuous–time Markov chains*, Adv. Appl. Probab. **19** (1987), 431–453.
28. Pollett, P.K., *A method involving antithetic sampling for estimating the blocking probabilities in a circuit–switched network*, Aust. Telecomm. Res. **22** (1988), 39–44.
29. Pollett, P.K. and Vassallo, A., *Diffusion approximations for some simple chemical reaction schemes* (1991).
30. Reiser, M., *Numerical methods in separable queueing networks*, Neuts, M.F. (ed.), Algorithmic Methods in Probability, TIMS Studies in Management Sciences, vol. 7, North–Holland, Amsterdam, 1977.
31. Rubinstein, R.Y., *Simulation and the Monte Carlo Method*, Wiley, New York, 1981.
32. Whittle, P., *Statistical processes of aggregation and polymerisation*, Proc. Cam. Phil. Soc. **61** (1965), 475–495.
33. Whittle, P., *Nonlinear migration processes*, Bull. Inst. Int. Statist. **42** (1967), 642–647.
34. Whittle, P., *Equilibrium distributions for an open migration process*, J. Appl. Probab. **5** (1968), 567–571.

35. Whittle, P., *Reversibility and acyclicity*, Ed. J. Gani, Perspectives in Probability and Statistics: Papers in Honour of M.S. Bartlett, Applied Probability Trust, Sheffield. Distributed by Academic Press, London, 1975, pp. 217–224.
36. Whittle, P., *Partial balance and insensitivity*, J. Appl. Probab. **22** (1985), 168–176.
37. Whittle, P., *Partial balance, insensitivity and weak coupling*, Adv. Appl. Probab. **18** (1986), 706–723.

P.K. Pollett, Department of Mathematics, The University of Queensland, Queensland 4072, AUSTRALIA.

E-mail address: `pkp@markov.maths.uq.oz.au`