Proceedings of the 2012 Winter Simulation Conference
C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, eds.

SAMPLING POINT PROCESSES ON STABLE UNBOUNDED REGIONS AND EXACT SIMULATION OF QUEUES

Jose Blanchet
Jing Dong

Industrial Engineering and Operations Research Department
Columbia University
New York, NY 10027, USA

ABSTRACT

Given a marked renewal point process (assuming that the marks are i.i.d.) we say that an unbounded region is stable if it contains finitely many points of the point process with probability one. In this paper we provide algorithms that allow to sample these finitely many points efficiently. We explain how exact simulation of the steady-state measure valued state descriptor of the infinite server queue follows as a simple corollary of our algorithms. We provide numerical evidence supporting that our algorithms are not only theoretically sound but also practical. Finally, having simulation optimization in mind, we also apply our results to gradient estimation of steady-state performance measures.

1 INTRODUCTION

Let $N = \{N(t) : t \in (-\infty, \infty)\}$ be a two sided time stationary renewal point process. We write $\{A_n : n \in \mathbb{Z}_0\}$ for the times at which the process N jumps, where $\mathbb{Z}_0 = \mathbb{Z}\backslash\{0\}$ denotes the set of integers removing zero, and with $A_1 > 0 > A_{-1}$. For simplicity we assume that $A_n < A_{n+1}$ for every n. Further, we define $X_n = A_{n+1} - A_n$.

Now let $\{V_n : n \in \mathbb{Z}_0\}$ be a sequence of independent and identically distributed (i.i.d.) random variables (r.v.'s) which are independent of the process N. Define $Z_n = (A_n, V_n)$ and consider the marked point process $\mathscr{M} = \{Z_n : n \in \mathbb{Z}_0\}$ which forms a subset of $\mathbb{R}^2$. We say that a (Borel measurable) set $\mathscr{B}$ is stable if $|\mathscr{M} \cap \mathscr{B}| < \infty$ almost surely (where $|\mathscr{C}|$ is used to denote the cardinality of the set $\mathscr{C}$).

Under natural assumptions on the inter-arrival times underlying N and on the distribution of the V_n's (stated in Section 2) we propose and study a class of algorithms that allow to sample exactly (i.e. without any bias) a realization of the set $\mathscr{M} \cap \mathscr{B}$ for a large class of unbounded, stable sets $\mathscr{B}$.

Our approach builds on algorithms that are fully developed and studied in Blanchet and Dong (2012). As an application of the class of algorithms that we study here, we provide a procedure that allows to sample from the steady-state measure valued descriptor of an infinite server queue without any bias (i.e.exact simulation). Such a procedure, for instance, is obtained by considering the particular case in which $\mathscr{B}$ takes the form $\mathscr{B} = \{(t,v) : v > |t|, t \leq 0\}$. Given that point processes constitute a natural way of constructing queueing models in great generality, we believe that the class of algorithms that we propose here have the potential to be applicable to the design of exact sampling algorithms of more general queueing models. This is a research avenue that we plan to investigate in the future.

We argue empirically that it is cheaper to run our exact sampling procedure to fully delete the initial bias than it is to do a burn-in period that reduces the bias to a reasonable size, say 5%, when talking about, for instance, the steady-state queue length.

Finally, we apply our exact sampling algorithms for infinite server queues to perform steady-state sensitivity analysis. For instance, we consider quantities such as the derivative of the steady-state average

978-1-4673-4781-5/12/$31.00 ©2012 IEEE

remaining service time with respect to the arrival rate or service rate. These quantities are of great interests in stochastic optimization via simulation.

So, in summary, our contributions are as follows:

i) We provide the first exact sampling algorithm for stationary marked renewal processes on unbounded and stable sets, see Section 2.

ii) As a corollary of i) we explain how to obtain an exact sampling algorithm for the steady-state measure valued descriptor of the infinite server queue. We also show empirically that this algorithm is *practical* in the sense of being both easy to code and fast to run, see Section 3.

iii) Finally, we provide new procedures for the sensitivity analysis of steady-state performance measures of the infinite server queue, see Section 4.

Relevant literature

Following the seminal work by Propp and Wilson (1996), several exact sampling algorithms have been developed, particularly for spatial point processes. Kendall (1998) and Kendall and Møller (2000) developed algorithms and analytical tools based on so-called Dominated Coupling From the Past (DCFP). DCFP is based on the idea of introducing a stationary dominating process that is simulatable. Compared to our method, firstly they use spatial birth and death processes (generally of poisson type) as the coupled dominating processes. This would limit the target distribution to be absolutely continuous with respect to the Poisson measure. Secondly the number of steps simulated in the naive DCFP grows exponentially with the system scale (i.e. arrival rate in the infinite server queue setting); see Proposition 1 in Berthelsen and Møller (2002) for a detailed proof. Although several modifications have been proposed, still the number of steps involved in these backward construction appears to be significantly large, especially when sampling in infinite volume regions (Fernandez, Ferrari, and Garcia 2002); see Section 7 in Berthelsen and Møller (2002) for empirical comparisons.

Our method is based on a construction that is being used in Blanchet and Sigman (2011) and Blanchet and Dong (2012); see also Ensor and Glynn (2000) for related ideas. The method involves the technique of simulating the maximum of a negative drift random walk and the last passage time of independent and identically distributed random variables to an increasing boundary. As shown in Blanchet and Dong (2012) the complexity of our algorithm scales graciously as the system scale grows.

2 SAMPLING FROM STABLE UNBOUNDED REGIONS

We start by discussing the assumptions behind our development.

Assumptions:

A1) Assume that $E|V_n|^{1/\alpha} < \infty$ for some $\alpha > 0$, we also write $F(\cdot) = P(V_n \leq \cdot)$ for the cumulative distribution function (CDF) of V_n and put $\overline{F}(\cdot) = 1 - F(\cdot)$ for the tail CDF.

A2) We assume that $F(\cdot)$ is known and easily accessible either in closed form or via efficient numerical procedures. Moreover, we can simulate V_n conditional on $V_n \in [a,b]$ with $P(V_n \in [a,b]) > 0$. Finally we can find $u(k)$ such that $u(k) \geq \int_k^\infty P(|V_1|^{1/\alpha} > v)dv$ and $u(k) \to 0$ as $k \to \infty$.

A3) Recall that $X_n = A_{n+1} - A_n > 0$. Define $\psi(\theta) = \log E \exp(\theta X_n)$ and assume that there exists $\delta > 0$ such that $\psi(\delta) < \infty$. Finally, let us write $\mu = EX_n$.

A4) Define $G(\cdot) = P(X_n \leq \cdot)$ and $\overline{G}(\cdot) = 1 - G(\cdot)$. Suppose that $G(\cdot)$ is known and that it is possible to simulate from $G_{eq}(\cdot) := \mu^{-1} \int_\cdot^\infty \overline{G}(t)\,dt$. Moreover, let $G_\theta(\cdot) = E \exp(\theta X_n - \psi(\theta)) I(X_n \leq \cdot)$ be the associated exponentially tilted distribution with parameter θ for $\psi(\theta) < \infty$. We assume that we can simulate from $G_\theta(\cdot)$.

Consider the class of sets $\mathscr{B} \subset \mathbb{R}^2$ that are Borel measurable and such that

$$\mathscr{B} \subset \mathscr{C}_\alpha = \{(t,v) : |v| \geq |t|^\alpha\}.$$

Our goal in this section is to develop an algorithm that allows to sample without bias the random set $\mathscr{M} \cap \mathscr{C}_\alpha$, and therefore $\mathscr{M} \cap \mathscr{B}$. We will discuss extensions that follow immediately from our formulation at the end of this section. Figure 1 illustrates the different shapes that the set $\mathscr{C}_\alpha$ can take depending on the values of $\alpha > 0$.

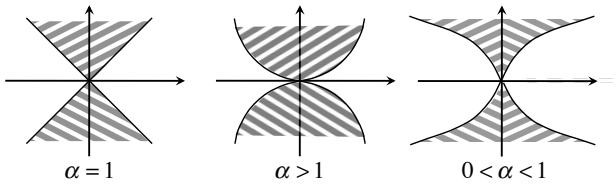

Figure 1: The area of $\mathscr{C}_\alpha$. The horizontal axis corresponds to the t coordinate while the vertical axis represents the v coordinate

We now proceed to explain our construction. As the stationary renewal point process is time reversible, starting at 0 the distribution of the forward process $\{Z_n : n > 0\}$ and the backward process $\{Z_n : n < 0\}$ are the same. In what follows we limit our discussion to the construction of the forward process and the simulation of the backward process is completely analogous.

We follow the idea in Blanchet and Dong (2012). Let $\varepsilon \in (0, \mu)$. Consider any random time κ, finite with probability one but large enough such that

$$A_{n+1} \geq n(\mu - \varepsilon) \text{ and } |V_{n+1}| \leq (n(\mu - \varepsilon))^\alpha$$

for all $n \geq \kappa$.

If such random time κ is well defined, we only need to simulate the stationary process up to κ to get a sample from the unbounded region.

Proposition 1 The random time κ defined above exists and it is finite with probability one.

Proof. By Chebyshev's inequality,

$$P(A_{n+1} < n(\mu - \varepsilon)) \leq E[\exp(\theta(n(\mu - \varepsilon) - A_{n+1}))] \leq \exp(-n(-\theta(\mu - \varepsilon) - \psi(-\theta)))$$

for any $\theta \geq 0$.
Let

$$I(-\varepsilon) = \max_{\theta \geq 0} \{-\theta(\mu - \varepsilon) - \psi(-\theta)\}$$

As $\psi(0) = 0$, $\psi'(0) = \mu$ and $\psi''(0) = Var(X) > 0$, $I(-\varepsilon) > 0$. Then

$$P(A_{n+1} < n(\mu - \varepsilon)) \leq \exp(-nI(-\varepsilon))$$

and

$$\sum_{n=1}^{\infty} P(A_{n+1} < n(\mu - \varepsilon)) \leq \frac{\exp(-I(-\varepsilon))}{1 - \exp(-I(-\varepsilon))} < \infty$$

By Borel-Cantelli lemma, $\{A_{n+1} \geq n(\mu - \varepsilon)\}$ eventually almost surely.
Similarly and independently we have

$$\sum_{n=1}^{\infty} P(|V_{n+1}| > (n(\mu - \varepsilon))^\alpha) = \sum_{n=1}^{\infty} P(|V_1|^{1/\alpha} > n(\mu - \varepsilon)) \leq \frac{1}{\mu - \varepsilon} \int_0^\infty P(|V_1|^{1/\alpha} > v) dv < \infty$$

Thus, again by Borel-Cantelli lemma, $\{|V_{n+1}| \leq (n(\mu - \varepsilon))^\alpha\}$ eventually almost surely. Therefore, $P(\kappa < \infty) = 1$ □

As $\{A_n : n \geq 1\}$ and $\{V_n : n \geq 1\}$ are independent of each other, we consider the following construction. Let $\kappa(A)$ be a random time satisfying that $A_{n+1} \geq n(\mu - \varepsilon)$ for $n \geq \kappa(A)$, and $\kappa(V)$ be a random time satisfying that $V_{n+1} \leq n(\mu - \varepsilon)$ for $n \geq \kappa(V)$. Clearly $\kappa(A)$ and $\kappa(V)$ are *not* stopping times and this makes the simulation of these times challenging. However, we will explain how to sample these times and then we can set $\kappa = \max\{\kappa(A), \kappa(V)\}$. Our construction will allow us to simulate $\{A_n : n \geq 1\}$ and $\{V_n : n \geq 1\}$ separately.

2.1 Simulation of : $1 \leq K \leq \max\{n, \kappa(A)\} + 1\}$

In this subsection we will introduce a method to simulate $\kappa(A)$ together with $\{A_k : k \geq 1\}$.

First, define A_1 according to the distribution $G_{eq}(\cdot)$. Sampling A_1 can be done according to A4).

Now, observe that $A_{n+1} = A_1 + X_1 + ... + X_n$ and define

$$\tilde{S}_n = n(\mu - \varepsilon) - (A_{n+1} - A_1) = \sum_{i=1}^{n} Y_i,$$

where $Y_i = (\mu - \varepsilon) - X_i$. Note that the Y_i's are i.i.d. with $EY_i = -\varepsilon$. If we set $\tilde{S}_0 = 0$, then $\{\tilde{S}_n : n \geq 0\}$ is a random walk with negative drift. We are interested in sampling up to the *last time* n at which $\tilde{S}_n > 0$.

We define the following sequence of random times:

$$\Delta_1 = 0,\ \Gamma_1 = \inf\{n \geq \Delta_1 : \tilde{S}_n - \tilde{S}_{\Delta_1} > 0\},$$

and for $j \geq 2$

$$\Delta_j = \inf\{n \geq \Gamma_{j-1}\mathbf{1}\{\Gamma_{j-1} < \infty\} \vee \Delta_{j-1} : \tilde{S}_n \leq 0\},$$
$$\Gamma_j = \inf\{n \geq \Delta_j : \tilde{S}_n - \tilde{S}_{\Delta_j} > 0\}.$$

Now, let $\gamma = \inf\{j \geq 1 : \Gamma_j = \infty\}$ and note that $\Delta_{\gamma+1} = \Delta_\gamma$ and that $\tilde{S}_n \leq 0$ for $n \geq \Delta_\gamma$, which in particular implies that $A_{n+1} \geq n(\mu - \varepsilon)$ for $n \geq \Delta_\gamma$. Therefore, we have that $\Delta_\gamma = \kappa(A)$.

In what follows we will explain how to simulate the Δ_j's and Γ_j's sequentially and jointly with the underlying random walk until time Δ_γ. One important observation is that for every $j \geq 1$, $\Delta_j < \infty$ almost surely by the strong law of large numbers.

Let us write $\mathscr{F}_n = \sigma\{Y_1, Y_2, ..., Y_n\}$ for the σ-field generated by the Y_j's up to time n. Let $\xi \geq 0$ and define

$$T_\xi := \inf\{n \geq 0 : \tilde{S}_n > \xi\},$$

then by the strong Markov property we have that for $j \leq \gamma$,

$$P(\Gamma_j = \infty | \mathscr{F}_{\Delta_j}) = P(\Gamma_j = \infty | \tilde{S}_{\Delta_j}) = P(T_0 = \infty) > 0,$$

where we use $P(\cdot)$ to denote the nominal probability measure under which $\tilde{S}_0 = 0$.

It is important then to note that

$$P(\gamma = k) = P(T_0 < \infty)^{k-1} P(T_0 = \infty)$$

for $k \geq 1$. In other words, γ is geometrically distributed. The procedure that we have in mind is to simulate Δ_γ in time intervals, and the number of time intervals is precisely γ.

Let $\psi_Y(\theta) = \log E \exp(\theta Y_i)$. As the moment generating function of X_i is finite in a neighborhood of the zero, $\psi_Y(\cdot)$ is also finite in a neighborhood of zero and $EY_i = \psi_Y'(0) = -\varepsilon$, $\mathrm{Var}(Y_i) = \psi_Y''(0) > 0$. Then by the convexity of $\psi_Y(\cdot)$, one can always select $\varepsilon > 0$ sufficiently small so that there exists $\eta > 0$ with

$\psi_Y(\eta) = 0$ and $\psi_Y'(\eta) > 0$. The root η allows us to define a new measure P_η based on exponential tilting so that

$$\frac{dP_\eta}{dP}(Y_i) = \exp(\eta Y_i).$$

Moreover, under P_η, $\tilde{S}_n$ is random walk with positive drift equal to $\psi_Y'(\eta)$ (Asmussen 2003 P. 365). Therefore $P_\eta(T_0 < \infty) = 1$ and $P(T_0 < \infty) = E_\eta(\exp(-\eta \tilde{S}_{T_0}))$. More generally, $P_\eta(T_\xi < \infty) = 1$ and

$$q(\xi) := P(T_\xi < \infty) = E_\eta(\exp(-\eta \tilde{S}_{T_\xi}))$$

for each $\xi \geq 0$. Based on the above analysis we now introduce a convenient representation to simulate a Bernoulli random variable $J(\xi)$ with parameter $q(\xi)$ namely,

$$J(\xi) = I(U \leq \exp(-\eta \tilde{S}_{T_\xi})). \tag{1}$$

where U is a uniform random variable independent of everything else under P_η.

Identity (1) provides the basis for an implementable algorithm to simulate a Bernoulli with success probability $q(\xi)$. Sampling $\{\tilde{S}_1, ..., \tilde{S}_{T_0}\}$ conditional on $T_0 < \infty$, as we shall explain now, corresponds to basically the same procedure. First, let us write $P^*(\cdot) = P(\cdot | T_0 < \infty)$. The following result provides an expression for the likelihood ratio between P^* and P_η.

Lemma 2 We have that

$$\frac{dP^*}{dP_\eta}(\tilde{S}_1, ..., \tilde{S}_{T_0}) = \frac{\exp(-\eta \tilde{S}_{T_0})}{P(T_0 < \infty)} \leq \frac{1}{P(T_0 < \infty)}.$$

Proof.

$$\begin{aligned} P(\tilde{S}_1 \in H_1, ..., \tilde{S}_{T_0} \in H_{T_0} | T_0 < \infty) &= \frac{P(\tilde{S}_1 \in H_1, ..., \tilde{S}_{T_0} \in H_{T_0}, T_0 < \infty)}{P(T_0 < \infty)} \\ &= \frac{E_\eta[\exp(-\eta \tilde{S}_{T_0}) I(\tilde{S}_0 \in H_0, ..., \tilde{S}_{T_0} \in H_{T_0})]}{P(T_0 < \infty)}. \end{aligned}$$

□

The previous lemma provides the basis for a simple acceptance / rejection procedure to simulate $\{\tilde{S}_1, ..., \tilde{S}_{T_0}\}$ conditional on $T_0 < \infty$. More precisely, we propose $(\tilde{S}_1, ..., \tilde{S}_{T_0})$ from $P_\eta(\cdot)$. Then one generates a uniform random variable U independent of everything else and accept the proposal if

$$U \leq \frac{1}{1/P(T_0 < \infty)} \times \frac{dP^*}{dP_\eta}(\tilde{S}_1, ..., \tilde{S}_{T_0}) = \exp(-\eta \tilde{S}_{T_0}).$$

This criterion coincides with $J(0)$ according to (1). So, the procedure above simultaneously obtains both a Bernoulli r.v. $J(0)$ with parameter $q(0)$, and the corresponding path $\{\tilde{S}_1, ..., \tilde{S}_{T_0}\}$ conditional on $T_0 < \infty$.

Algorithm 1 (Outputs $(\tilde{S}_0, ..., \tilde{S}_{\Delta_\gamma})$)

Step 0. Set $K = 0$, and $S_0 = 0$

Step 1. Simulate $(\tilde{S}_1, ..., \tilde{S}_{T_0})$ from P_η and compute $J := J(0)$ according to (1).

Step 2. If $J = 1$, then let $S_{K+j} = \tilde{S}_j$ for $j = 1, ..., T_0$ and update $K \longleftarrow K + T_0$. Then, go back to Step 1. Otherwise, $J = 0$ (i.e. $\Delta_\gamma = K$), stop and output $(S_0, ..., S_K)$

Remark: It has been proved in Blanchet and Dong (2012) that the expected number of times we need to repeat Step 1 does not change with the system scale (i.e. the arrival rate).

We noted earlier that $\Delta_\gamma = \kappa(A)$ and Algorithm 1 together with the initial procedure to sample A_1 allows us to simulate $(A_{j+1} : 0 \leq j \leq \kappa(A))$, and we know that $A_{n+1} \geq n(\mu - \varepsilon)$ for $n \geq \kappa(A)$. We need to simulate A_{n+1} for $n \leq \kappa = \max\{\kappa(A), \kappa(V)\}$, and $\kappa(V)$ is independent of $\kappa(A)$. So, there might be cases for which we will have to sample A_{n+1} for $n > \kappa(A)$. Since $A_{n+1} = A_1 - \tilde{S}_n + n(\mu - \varepsilon)$ it suffices to explain how to simulate $\tilde{S}_n$ for $n > \Delta_\gamma$. In turn, it suffices to explain how to simulate $(\tilde{S}_n : n \geq 0)$ with $\tilde{S}_0 = 0$ conditional on $T_0 = \infty$. We will once again apply an acceptance / rejection procedure but this time we will use the original (nominal) distribution as the proposal distribution. Define

$$P'(\cdot) = P(\cdot | T_0 = \infty).$$

The following result provides an expression for the likelihood ratio between P' and P.

Lemma 3 We have that

$$\frac{dP'}{dP}(\tilde{S}_1, ..., \tilde{S}_l) = \frac{I(T_0 > l)(1 - q(-\tilde{S}_l))}{P(T_0 = \infty)} \leq \frac{1}{P(T_0 = \infty)}.$$

Proof.

$$\begin{aligned} P(\tilde{S}_1 \in H_1, ..., \tilde{S}_l \in H_l | T_0 = \infty) &= \frac{P(\tilde{S}_1 \in H_1, ...\tilde{S}_l \in H_l, T_0 = \infty)}{P(T_0 = \infty)} \\ &= \frac{E[I(\tilde{S}_1 \in H_1, ..., \tilde{S}_l \in H_l) I(T_0 > l) P(T_0 = \infty | \tilde{S}_0, ..., \tilde{S}_l)]}{P(T_0 = \infty)}. \end{aligned}$$

The result then follows from the strong Markov property and homogeneity of the random walk. □

We are in good shape now to apply acceptance / rejection to sample from P'. The previous lemma indicates that to sample $\{\tilde{S}_0, ..., \tilde{S}_l\}$ given $T_0 = \infty$ we can propose from the original (nominal) distribution and accept with probability $q(-\tilde{S}_l)$ as long as $\tilde{S}_j \leq 0$ for all $0 \leq j \leq l$. So, in order to perform the acceptance test we need to sample a Bernoulli with parameter $q(-\tilde{S}_l)$, but this is easily done using identity (1). Thus we obtain the following procedure.

Algorithm 2 (Given $n \geq 0$ outputs $\{A_1, A_2, ..., A_{\max\{n, \kappa(A)\}+1}\}$)

Step 1. Run Algorithm 1 and obtain $\{S_0, S_1, ..., S_K\}$.
Step 2. If $K = \kappa(A) \geq n$, jump to Step 6. Otherwise, $K < n$, let $l = n - K \geq 1$.
Step 3. Simulate $\{\tilde{S}_0, \tilde{S}_1, ..., \tilde{S}_l\}$ from the original (nominal) distribution with $\tilde{S}_0 = 0$.
Step 4. If $\tilde{S}_j \leq 0$ for all $0 \leq j \leq l$ then sample a Bernoulli $J(-\tilde{S}_l)$ with parameter $q(-\tilde{S}_l)$ using (1) and continue to Step 5. Otherwise (i.e. $\tilde{S}_j > 0$ for some $1 \leq j \leq l$) go back to Step 3.
Step 5. If $J(-\tilde{S}_l) = 1$, go back to Step 3. Otherwise, $J(-\tilde{S}_l) = 0$, let $S_{K+i} = S_K + \tilde{S}_i$ for $i = 1, 2, ..., l$
Step 6. Let $m = \max\{n, \kappa(A)\}$. Simulate A_1 with CDF $G_{eq}(\cdot) = \mu^{-1} \int_\cdot^\infty \bar{G}(t) dt$. Set $A_{n+1} = A_1 - S_n + n(\mu - \varepsilon)$ for $n = 1, ..., m$. Output $\{A_1, ..., A_{m+1}\}$.

2.2 Simulation of : $1 \leq N$

In this section we will introduce a method to simulate $\kappa(V)$ together with the $\{V_n : n \geq 1\}$.

Let $p(n) = P(|V_1| > (n(\mu - \varepsilon))^\alpha)$. We define $\Upsilon_0 = 0$ and $\Upsilon_i = \inf\{n > \Upsilon_{i-1} : |V_{n+1}| > (n(\mu - \varepsilon))^\alpha\}$ for $i = 1, 2, ...$ We also define two independent sequences of random variables, $\{\hat{V}_{n+1} : n \geq 1\}$, and $\{\bar{V}_{n+1} : n \geq 1\}$ as follows. The elements in each sequence are i.i.d., $\hat{V}_{n+1}$ is distributed as V_{n+1} conditional

on $|V_{n+1}| > (n(\mu-\varepsilon))^\alpha$, and $\bar{V}_{n+1}$ follows the distribution of V_{n+1} conditional on $|V_{n+1}| \le (n(\mu-\varepsilon))^\alpha$. We simulate V_1 following its nominal distribution independent of everything else.

Let $\sigma = \inf\{i \ge 0 : \Upsilon_i = \infty\}$. Then $V_{n+1} \le (n(\mu-\varepsilon))^\alpha$ for $n \ge \Upsilon_{\sigma-1}+1$. We next introduce a method to sample $\Upsilon_1, \Upsilon_2, ...$ sequentially and jointly with the V_n's up until $\Upsilon_{\sigma-1}$.

The following lemma provides the basis to guarantee the termination of our procedure.

Lemma 4 If $E|V_1|^{1/\alpha} < \infty$, then

$$P(\Upsilon_1 = \infty) = \prod_{i=1}^{\infty}(1-p(i)) \ge \exp(-2E|V_1|^{1/\alpha}/(\mu-\varepsilon)) > 0,$$

consequently $E\sigma \le \exp(2E|V|^{1/\alpha}/(\mu-\varepsilon)) < \infty$.

Remark: The bound on $E\sigma$ can be improved. This improvement is important for the theoretical asymptotic analysis of GI/GI/∞ application, see Blanchet and Dong (2012).

Proof.

$$\begin{aligned} P(\Upsilon_1 = \infty) = \prod_{n=1}^{\infty}(1-p(n)) &\ge \prod_{n=1}^{\infty}\exp(-2p(n)) \\ &\ge \exp(-\frac{2}{\mu-\varepsilon}\int_0^{\infty} P(|V_1|^{1/\alpha} > v)dv) = \exp(-\frac{2E|V_1|^{1/\alpha}}{\mu-\varepsilon}) \end{aligned}$$

For $i = 2,3,...$ conditional on $\Upsilon(i-1) = k$:

$$P(\Upsilon_i = \infty | \Upsilon_{i-1} = k) = \prod_{n=k+1}^{\infty}(1-p(n)) \ge \exp(-\frac{2\int_k^{\infty} P(|V_1|^{1/\alpha} > v)dv}{\mu-\varepsilon} \ge \exp(-\frac{2E|V_1|^{1/\alpha}}{\mu-\varepsilon})$$

Thus σ is stochastically dominated by a geometric random variable with parameter $p = \exp(-2E|V_1|^{1/\alpha}/(\mu-\varepsilon))$, the result then follows.

□

Notice that

$$\prod_{i=k+1}^{l}(1-p(i)) \ge P(\Upsilon_i = \infty | \Upsilon_{i-1} = k) \ge \prod_{i=k+1}^{l}(1-p(i)) \times \exp(-\frac{2\int_l^{\infty} P(|V_1|^{1/\alpha} > v)dv}{\mu-\varepsilon}) \qquad (2)$$

for $l \ge k+1$.

Thus if we are simulating $I \sim$ Bernoulli(r_i) with $r_i := P(\Upsilon_i = \infty | \Upsilon_{i-1})$, then with probability one we can check whether $U \le P(\Upsilon_i = \infty | \Upsilon_{i-1})$ for $U \sim$ Unif$[0,1]$ by making l sufficiently large without calculating the infinite product in the definition of $P(\Upsilon_i = \infty | \Upsilon_{i-1})$.

On the other hand, if we define $\prod_{j=1}^{0}(1-p(j)) := 1$, then

$$P(\Upsilon_1 = n | \Upsilon_1 < \infty) = p(n)\frac{\prod_{j=1}^{n-1}(1-p(j))}{P(\Upsilon_1 < \infty)} \le p(n)\frac{1}{P(\Upsilon_1 < \infty)}.$$

Consider a random variable N with the following probability density function

$$P(N = n) = cp(n)$$

for $n = 1,2,...$, where $c = (\sum_{n=1}^{\infty} p(n))^{-1}$. Then $P(\Upsilon_1 = n | \Upsilon_1 < \infty)/P(N = n) \le 1/(cP(\Upsilon_1 < \infty))$.

So we can simulate Υ_1 given $\Upsilon_1 < \infty$ using acceptance / rejection with N as the proposal random variable. Generalizing the idea to Υ_i, we can obtain the following algorithm

Algorithm 3 (Given $\Upsilon_{i-1} = k$, outputs Υ_i conditional on $\Upsilon_i < \infty$)

Step 1. Let $c = (\sum_{n=k+1}^{\infty} p(n))^{-1}$. Simulate N with probability density function $P(N=n) = cp(n)$ for $n = k+1, k+2, ...$

Step 2. Simulate $U \sim \text{Unif}[0,1]$ independently. If $U \leq \prod_{j=k+1}^{N-1}(1-p(j))$, set $\Upsilon_i = N$ and stop. Otherwise go back to Step 1

We conclude this section with our procedure to simulate $\{V_1, V_2, ...V_{\kappa(V)+1}\}$.

Algorithm 4 (Outputs $\{V_1, V_2, ...V_{\kappa(V)+1}\}$)

Step 0. Set $\Upsilon_0 = 0$, $i = 1$. Simulate V_1 from its nominal distribution.

Step 1. Simulate $I \sim \text{Bernoulli}(r_i)$ with $r_i := P(\Upsilon_i = \infty | \Upsilon_{i-1})$ (see (2)).

Step 2. If $I = 1$, set $\kappa(V) = \Upsilon_{i-1} + 1$. Simulate $V_{\kappa(V)+1}$ by sampling from $\bar{V}_{\kappa(V)+1}$ and stop. Otherwise $I = 0$, sample Υ_i conditional on $\Upsilon_i < \infty$ and the value of Υ_{i-1} using Algorithm 3. Simulate the process between $\Upsilon_{i-1}+2$ and $\Upsilon_i + 1$ by sampling from $\bar{V}_n$ for $\Upsilon_{i-1}+2 \leq n \leq \Upsilon_i$ and $\hat{V}_n$ for $n = \Upsilon_i + 1$. Set $i = i+1$ and then go back to Step 1.

3 APPLICATION TO THE INFINITE SERVER QUEUE

As a direct application of the ideas discussed in the previous section we study steady-state simulation for the infinite server queue. The following diagram indicates how to construct the steady-state measure valued descriptor assuming that we can sample all the points inside the set

$$\mathscr{C} = \{(t,v) : v \geq |t|, t \leq 0\}.$$

Let $Q(t,y)$ denote the number of people in the system at time t with residual service time strictly greater than y and $E(t)$ denote the time elapsed since the previous arrival at time t (i.e. $E(\cdot)$ is the age process associated with $N(\cdot)$). Figure 2 below depicts the region $\mathscr{C}$. Every point in $|\mathscr{M} \cap \mathscr{C}|$ is projected to the vertical line at time zero by drawing a -45^0 line. The final position in the vertical line if positive, represents the corresponding remaining service time. Since the underlying point process is time stationary, the whole configuration of points obtained by this procedure at time zero is a snap shot of the steady-state distribution of the infinite server queue.

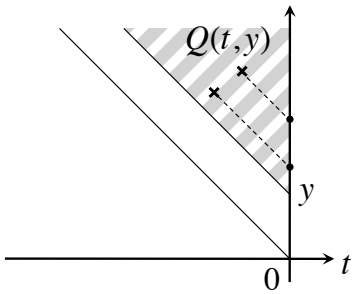

Figure 2: The points lies in the shaded area correspond to people who are still in the system at time 0 with remaining service time greater than y.

3.1 Algorithm for the Infinite Server Queue

As depicted in Figure 2 after projecting into the vertical line at $t = 0$, we obtain the stationary remaining service requirements of the customers at time zero. We shall use $R_1, R_2, ..., R_{Q(0,0)}$ to denote the remaining service times. The labeling is arbitrary although we will assign smaller indexes to customers that have

spent less time in the system. Our algorithm proceeds as follows.

Algorithm 5 (Outputs $\{R_1, R_2, ..., R_{Q(0,0)}\}$ and $E(0)$)

Step 1. Use Algorithm 4 to simulate the $\{V_n, 1 \leq n \leq \kappa(V)+1\}$.
Step 2. Use Algorithm 2 to simulate the $\{A_1, A_2, ..., A_{\max\{\kappa(V),\kappa(A)\}+1}\}$.
Step 3. Set $\kappa = \max(\kappa(V), \kappa(A))$. If $\kappa > \kappa(V)$, simulate V_n by sampling from $\bar{V}_n$ for $n = \kappa(V)+2, ..., \kappa+1$.
Step 4. Set $q = 0$, $i = 0$ and repeat the following procedure until $i = \kappa$:
set $i = i+1$*; if* $V_i > A_i$*, set* $q = q+1$ *and* $R_q = V_i - A_i$.
Output $\{R_1, R_2, ...R_q\}$ and A_1.

3.2 Empirical Performance

Let $Y = \{Y(t) : t \geq 0\}$ be a continuous time Markov process on the state space Ω and f is a real-valued function defined on Ω. The ergodic theorem guarantees in great generality (assuming a unique stationary distribution $\pi(\cdot)$) that

$$\frac{1}{t}\int_0^t f(Y(s))ds \to \int_\Omega f(y)\pi(dy)$$

as $t \to \infty$ almost surely for every positive, measurable function $f(\cdot)$. In the setting of the infinite server queue such a stationary distribution exists if $EV_n < \infty$ and $EX_n < \infty$. The most natural estimator for $E_\pi f(Y) := \int_\Omega f(y)\pi(dy)$ is therefore

$$\Phi(t, Y(0)) := \frac{1}{t}\int_0^t f(Y(s))ds,$$

where $Y(0)$ is the initial state. The estimator $\Phi(t, Y(0))$ is generally biased unless $Y(0)$ is sampled from the stationary distribution $\pi(\cdot)$ (Asmussen and Glynn 2007 P. 97). Our algorithm has the obvious advantage of removing the initial transient.

In what follows we conduct some simulation experiment to evaluate the practical performance of our algorithm. The idea is to fix a reasonable tolerance error, say 10%, for a given performance measure. Then we want to empirically find how large a burn-in period one would need in practice to reduce the initial transient bias to about 10%. In order to effectively quantify the error we select a class of systems for which $\pi(\cdot)$ can be explicitly evaluated.

We consider an infinite server queue with Poisson arrivals and Lognormal service times. As we are interested in the efficiency of our algorithm for relatively large systems, we set the arrival rate $\lambda = 100$ and the service time $V_n \sim \text{Lognormal}(-0.25, 0.5)$ (i.e. V_n has the same distribution as $\exp(-.25 + .5 \times N(0,1))$, where $N(0,1)$ denotes a standard Gaussian random variable).

Let $Y(t) = (Q(t,\cdot), E(t)) \in \mathscr{D}[0,\infty) \times \mathbb{R}_+$, then $Y(t)$ is a Markovian measure valued descriptor of the infinite server queue (of course in the Poisson arrival case one does not need to keep track of $A(\cdot)$).

We first compare the performance of our algorithm to the burn-in period defined as the period needed to reduce the initial transient as indicated earlier. Let $f(Y(t)) = Q(t,0)$, i.e. the number of people in the system at time t. We measure the computation effort of the algorithm in terms of the number of arrivals (we call this the number of steps) simulated. Given $\varepsilon > 0$ we let $n(\varepsilon)$ denote the minimum number of steps required so that $|E\Phi(A_{n(\varepsilon)}, (\phi, 0)) - E_\pi Q(0,0)|/E_\pi Q(0,0) \leq \varepsilon$, where $(\phi, 0)$ denotes a system that starts empty with $E(0) = 0$ (recall that $E(\cdot)$ is the age process associated with $N(\cdot)$, i.e. when $E(0) = x$, A_1 is distributed as X_n conditional on $X_n > x$). Table 1 shows the relation between ε and $n(\varepsilon)$, obtained empirically based on the average of 10^4 independent replications

Compared to the results in Table 1, our algorithm is unbiased. The average number of steps involved is $n = 592.6369$ based on the average of 10^4 independent replications and the average computer time needed for a single replication is 0.0249 s.

Table 1: Bias of $\Phi(S_{n(\varepsilon)})$.

ε	$n(\varepsilon)$	computer time (s)
10.26%	6×10^2	0.0310
5.71%	1×10^3	0.0382
1.17%	5×10^3	0.1367

In addition, in Table 2 we compare the performance of the estimators $\Phi(A_n,(\phi,0))$ and $\Phi(A_{n'},(Q(0,\cdot),A_1))$, where $Q(0,\cdot)$ and A_1 are sampled according to Algorithm 5. n and n' are calibrated so that the computation budget is basically the same in both estimators. Under our procedure, $E\kappa$, the average number of arrivals required to terminate is approximately equal to 600. So for instance, the first row in Table 2 corresponds to $n=10^4$. This means that $n'\approx 9.4\times10^3 = 10^4-600$. The true value of $E_\pi Q(0,0)$ is 88.2497. The sample mean and sample standard deviation are calculated using the method of Batch means. The result in Table 2 shows that our mixed method performs better than the batch means with relatively small computation budget, while with large budget, the two methods are about the same.

Table 2: Simulation result with different initial states.

	$(\phi,0)$		$(Q(0,\cdot),A_1)$	
n	Sample Mean	Sample Std	Sample Mean	Sample Std
1×10^4	86.1274	1.0104	88.1713	0.6018
5×10^4	89.0893	0.4587	88.2956	0.3770
1×10^5	88.5151	0.3531	88.1270	0.2976
5×10^5	88.3022	0.1481	88.3581	0.1402

4 APPLICATION TO SENSITIVITY ANALYSIS OF INFINITE SERVER QUEUE

In this section, we apply our algorithm to sensitivity analysis of the infinite server queue. We consider a sequence of systems indexed by (λ,ν), $\lambda>0$, $\nu>0$. Given (λ,ν), the interarrival times are multiplied by $1/\lambda$, obtaining X_n/λ for all n, and the service times are multiplied by $1/\nu$, thus we have V_n/ν for all n. We assume that $EV_n<\infty$ and $EX_n<\infty$. We will use the notation $Q_{\lambda,\nu}(\cdot)$ to denote the infinite server queue descriptor for the (λ,ν)-system. Our strategy rests on the application of Infinitesimal Perturbation Analysis (IPA), see for instance Glasserman (2003) P. 386. We assume here that the interarrival times have a continuous distribution.

We illustrate the methodology by computing the sensitivity of the steady-state average remaining service time, which we denote by $E_\pi\bar{R}(\lambda,\nu)$; namely,

$$E_\pi\bar{R}(\lambda,\nu)=E_\pi\frac{1}{Q_{\lambda,\nu}(0,0)}\int_0^\infty yQ_{\lambda,\nu}(0,dy).$$

We also consider

$$E_\pi R^\infty(\lambda,\nu)=E_\pi(\inf\{y\geq 0: Q_{\lambda,\nu}(0,y)=0\}),$$

in words, the steady-state maximum remaining service time. In order to apply IPA we need to define a few quantities.

First, let us define $\bar{\Xi}(\lambda,\nu)$ to be the average elapsed service time of the customers that are present at time zero (given the construction of the stationary process $\{Q_{\lambda,\nu}(t,\cdot):t\in(-\infty,\infty)\}$, see Figure 2). That is,

$$\bar{\Xi}(\lambda,\nu)=\frac{1}{Q_{\lambda,\nu}(0,0)}\sum_{n=-1}^{-\infty}\frac{|A_n|}{\lambda}I\left(\frac{|A_n|}{\lambda}<\frac{V_n}{\nu}\right)$$

Likewise, define $\bar{V}(\lambda,\nu)$ as the average of the total service requirement of the customers that are present at time zero, namely

$$\bar{V}(\lambda,\nu)=\frac{1}{Q_{\lambda,\nu}(0,0)}\sum_{n=-1}^{-\infty}\frac{V_n}{\nu}I\left(\frac{|A_n|}{\lambda}<\frac{V_n}{\nu}\right).$$

Next, we define $\Xi^{(\infty)}(\lambda,\nu)$ as the elapsed service time of the customer with the maximum remaining service time at time zero and $V^{(\infty)}(\lambda,\nu)$ as his total service time requirement. Specifically, if we let $m=\arg\max\{n:V_n/\nu-|A_n|/\lambda\}$ then

$$\Xi^{(\infty)}(\lambda,\nu)=\frac{|A_m|}{\lambda}\text{ and }V^{(\infty)}(\lambda,\nu)=\frac{V_m}{\nu}$$

We then obtain the following representation for the derivatives of $E_\pi\bar{R}(\lambda,\nu)$ and $E_\pi R^\infty(\lambda,\nu)$ with respect to λ and ν.

Lemma 5 We have that

i)

$$\frac{\partial}{\partial\lambda}E_\pi\bar{R}(\lambda,\nu)=\frac{1}{\lambda}E_\pi\bar{\Xi}(\lambda,\nu)\text{ and }\frac{\partial}{\partial\nu}E_\pi\bar{R}(\lambda,\nu)=-\frac{1}{\nu}E_\pi\bar{V}(\lambda,\nu).$$

ii)

$$\frac{\partial}{\partial\lambda}E_\pi R^\infty(\lambda,\nu)=\frac{1}{\lambda}E_\pi\Xi^{(\infty)}(\lambda,\nu)\text{ and }\frac{\partial}{\partial\nu}E_\pi R^\infty(\lambda,\nu)=-\frac{1}{\nu}E_\pi V^{(\infty)}(\lambda,\nu)]$$

Proof. We only give a proof of part i) here as the proof of part ii) is entirely analogous.
Let R_n denote the remaining service time of the nth customer at time zero and V_n as his total service time requirement, then $R_n\le V_n$. Thus if $EV_n<\infty$, we have

$$E_\pi\bar{R}(\lambda,\nu)<\infty$$

for any $\lambda>0$,$\nu>0$.
For a fixed sample path ω constructed backward in time, let $R_n(\lambda,\nu,\omega)$, $n<0$, denote the remaining service time of customer n (counting backward in time) at time 0 in system (λ,ν). Then $R_n(\lambda,\nu,\omega)=(V_n(\omega)/\nu-|A_n(\omega)|/\lambda)^+$ and

$$\lim_{h\to0}\frac{R_n(\lambda+h,\nu,\omega)-R_n(\lambda,\nu,\omega)}{h}=\frac{|A_n(\omega)|}{\lambda^2}1\{\frac{V_n(\omega)}{\nu}\ge\frac{|A_n(\omega)|}{\lambda}\}$$

$$\lim_{h\to0}\frac{R_n(\lambda,\nu+h,\omega)-R_n(\lambda,\nu,\omega)}{h}=-\frac{V_n(\omega)}{\nu^2}1\{\frac{V_n(\omega)}{\nu}\ge\frac{|A_n(\omega)|}{\lambda}\}$$

Thus the derivative $\frac{\partial}{\partial\lambda}\bar{R}(\lambda,\nu)$ and $\frac{\partial}{\partial\nu}\bar{R}(\lambda,\nu)$ exists.
Let Ξ_n denote the elapsed service time of the nth customer at time zeros and define $\Xi_n=V_n$ if he is no longer in the system at time zero, then $\Xi_n\le V_n$. Therefore $E_\pi\frac{\partial}{\partial\lambda}\bar{R}(\lambda,\nu)<\infty$ and $E_\pi\frac{\partial}{\partial\nu}\bar{R}(\lambda,\nu)<\infty$.
As $|(\bar{R}_n(\lambda+h,\nu)-\bar{R}_n(\lambda,\nu))/h|\le\max_{\kappa_{\lambda+h,\nu}<n<0}V_n/\lambda^2$ and $|(\bar{R}_n(\lambda,\nu+h)-\bar{R}_n(\lambda,\nu))/h|\le\max_{\kappa_{\lambda,\nu+h}<n<0}V_n/\nu^2$, by Lebesgue Dominated Convergence Theorem, we have

$$\frac{\partial}{\partial\lambda}E_\pi\bar{R}(\lambda,\nu)=E_\pi\frac{\partial}{\partial\lambda}\bar{R}(\lambda,\nu)\text{ and }\frac{\partial}{\partial\nu}E_\pi\bar{R}(\lambda,\nu)=E_\pi\frac{\partial}{\partial\nu}\bar{R}(\lambda,\nu)$$

As the interarrival times have a continuous distribution, $P(V_n/\nu=|A_n|/\lambda)=0$ for $n<0$.
Combining the change of limit and the sample path analysis we have

$$\frac{\partial}{\partial\lambda}E_\pi\bar{R}(\lambda,\nu)=\frac{1}{\lambda}E_\pi\bar{\Xi}(\lambda,\nu)\text{ and }\frac{\partial}{\partial\nu}E_\pi\bar{R}(\lambda,\nu)=-\frac{1}{\nu}E_\pi\bar{V}(\lambda,\nu)$$

□

Table 3 shows the simulated results of an infinite server queue with base (i.e. $\lambda = 1$) interarrival times distributed as Gamma$(2,2)$ and base (i.e. $\nu = 1$) service times distributed as Lognormal$(-0.25, 0.5)$.

Table 3: Simulation result from exact sampling.

(λ,ν)	$\frac{\partial}{\partial\lambda}E_\pi\bar{R}(\lambda,\nu)$	$\frac{\partial}{\partial\nu}E_\pi\bar{R}(\lambda,\nu)$	$\frac{\partial}{\partial\lambda}E_\pi R^\infty(\lambda,\nu)$	$\frac{\partial}{\partial\nu}E_\pi R^\infty(\lambda,\nu)$
$(80,1)$	7.0741×10^{-3}	-1.1320	6.1022×10^{-3}	-2.8389
$(100,1)$	5.6470×10^{-3}	-1.1316	4.9379×10^{-3}	-2.9495
$(120,1)$	4.7236×10^{-3}	-1.1337	4.2337×10^{-3}	-3.0684

ACKNOWLEDGMENTS

Support from the NSF foundation through the grants CMMI-0846816 and CMMI-1069064 is gratefully acknowledged.

REFERENCES

Asmussen, S. 2003. *Applied Probability and Queues*. 2 ed. New York: Spinger.

Asmussen, S., and P. Glynn. 2007. *Stochastic Simulation*. New York: Spinger.

Berthelsen, K., and J. Møller. 2002. "A primer on perfect simulation for spatial point process". *Bull Braz Math Soc* 33(3):351–367.

Blanchet, J., and J. Dong. 2012. "Exact sampling of loss systems". *in preparation*.

Blanchet, J., and K. Sigman. 2011. "On exact sampling of stochastic perpetuities". *J. Appl. Probab.* 48A:165–182.

Ensor, K., and P. Glynn. 2000. "Simulating the maximum of a random walk". *Journal of Statistical Planning and Inference* 85:127–135.

Fernandez, R., P. Ferrari, and N. Garcia. 2002. "Perfect simulation for interacting point processes, loss networks and Ising models". *Stoch. Process. Appl.* 102(1):63–88.

Glasserman, P. 2003. *Monte Carlo Methods in Financial Engineering*. New York: Spinger.

Kendall, W. 1998. "Perfect simulation for area-interaction point processes". In *Probability Towards 2000*, edited by L. Accardi and C. Heyde, 218–234. New York: Spinger.

Kendall, W., and J. Møller. 2000. "Perfect simulation using dominating processes on ordered spaces, with application to locally stable point pocesses". *Adv. Appl. Prob.* 32:844–865.

Propp, J., and D. Wilson. 1996. "Exact sampling with coupled Markov chains and applications to statistical mechanics". *Random Structures and Algorithms* 9:223–252.

AUTHOR BIOGRAPHIES

JOSE BLANCHET is a faculty member of the IEOR Department at Columbia University. Jose holds a Ph.D. in Management Science and Engineering from Stanford University. Prior to joining Columbia he was a faculty member in the Statistics Department at Harvard University. Jose is a recipient of the 2009 Best Publication Award given by the INFORMS Applied Probability Society and of the 2010 Erlang Prize. He also received a PECASE award given by NSF in 2010. He has research interests in applied probability and Monte Carlo methods. He serves in the editorial board of Advances in Applied Probability, Journal of Applied Probability, Mathematics of Operations Research and QUESTA. His email is jose.blanchet@columbia.edu.

JING DONG is a Ph.D. student in the IEOR Department at Columbia University, and a member of INFORMS. She received her B.S. in Actuarial Science from Hong Kong University. Her research interests is in rare event simulation and stochastic modeling. Her email is jd2736@columbia.edu.