Importance sampling in systems simulation
Hopmans, A.C.M.; Kleijnen, J.P.C.

Published in:
Mathematics and Computers in Simulation

Publication date:
1979

Link to publication in Tilburg University Research Portal

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
IMPORTANCE SAMPLING IN SYSTEMS SIMULATION: A PRACTICAL FAILURE? *

A.C.M. HOPMANS
Dr. Neher Laboratory, P.T.T., Leidschendam, The Netherlands

J.P.C. KLEIJNEN
Information Systems Group, Department of Business and Economics, Katholieke Hogeschool, Tilburg, The Netherlands

Abstract

A network of servers, known as a grading in telecommunication engineering, is simulated to estimate the probability of a customer being "blocked": all servers busy. Since blocking is a very rare event (1% to 5% chance), importance sampling was considered for reduction of the simulation variance. The basic idea of importance sampling is first explained by means of a non-dynamic system. For dynamic systems a method was proposed by Bayes in 1970, which is related to the "virtual measures" published by Carter and Ignall in 1975. For simple queuing systems, we derive the resulting variance, using the renewal or regenerative property of such systems. For our practical "grading" system several alternative importance regions are investigated. For practical reasons we choose to start an importance region immediately after a call gets blocked (not a renewal state). The analysis and simulation experiments for the resulting estimator yielded the estimated optimal length of the importance region and the optimal number of replications of the region. Unfortunately, a net increase in variance resulted.

1. Introduction

We feel that this paper is unusual in so far as it reports on an unsuccessful research effort. This effort tried to reduce the variability of simulation results by the application of a variance reduction technique (VRT). Not that we think that such unsuccessful investigations have been rare, but reporting such attempts seems to be rare indeed; see also [13]. Nevertheless the documentation of such abortive attempts can be useful: Practitioners may be warned against too optimistic expectations. Practitioners may be warned against too optimistic expectations. Theoreticians may be stimulated to revise our approach and devise a better VRT. Moreover, our efforts resulted in an improved insight into the behavior of the simulated system. Before we explain our particular VRT, we briefly characterize the practical system and its model to which we applied the VRT. Note that a glossary of the major symbols is provided in Appendix 1.

* This research was done by the authors as members of the Working Group on the Statistical Design and Analysis of Simulation Experiments, chaired by J.P.C. Kleijnen, under the auspices of the Section for Operations Research (SOR) of the Netherlands Society for Statistics (VVS). Many critical questions and helpful comments were received from the members of the Working Group, especially B. Sanders and R. van der Ven (P.T.T., Leidschendam), G. Horstmeier and R. Sierenberg (Delft University), T. Boulogne and R. van der Ham (ECT, Rotterdam).
The actual grading we investigated is more complicated and is shown in Fig. 3.

![Diagram of a realistic grading]

This practical grading shows \( g = 8 \) customer sources. Each source has \( k = 15 \) points of entry or "contacts", but since these "servers" are connected to form a common server for several customer sources, less than \( 8 \times 15 \) servers result. Actually Fig. 3 shows \( N = 45 \) servers. When a customer is generated, one of the 15 contacts is selected randomly. If this line is busy, then another line in its "column" is selected cyclically. If all the available 15 (not 45) servers are busy, then the call gets blocked. Observe that, unlike traditional queuing systems, our grading type does not permit customers to wait for service, i.e., if the system is busy new arrivals are lost. (Of course a customer may try again at a later point of time.) We further assume that the 8 customer sources generate demands independently of each other.

How can we estimate the blocking probability? Realistic gradings are too complicated for analytical solutions. Therefore simulation is used. The quantity to be estimated is very small, since a realistic grading is so designed that the blocking probability is between 1% and 5%. Hence reliable estimates require very long simulation runs, since during long periods of the simulated history nothing of interest happens, i.e., no blocking occurs!

Elsewhere we discussed a successful variance reduction technique applied to our grading problem. This VRT combined regression analysis with simulation, and is known as "control variates"; see [5] for details. Another VRT we applied, is so-called roulette simulation: Because of the Poisson (memoryless) character of our grading, the event-timing administration can be eliminated; see [4] and [11]. Using this roulette simulation we reduced the required computer time by a factor 2.

Note that the usual, crude estimator of the steady-state blocking probability \( B \), is

\[
\hat{B} = \frac{\text{number of calls lost}}{\text{total number of calls}}
\]

The variance of this estimator can be estimated by dividing the total simulation run into a number of subruns - in our case 15 subruns. These subruns can be assumed to give independent blocking probabilities; see [10, pp. 458-460].

2. Importance Sampling: Non-Dynamic Situations

The basic idea of importance sampling or IS, was introduced by Kahn and Marshall [8] in 1953 as follows: Suppose we wish to estimate \( \xi \), the value of the following integral

\[
\xi = \int g(x) f(x) dx = E[g(x)]
\]

where \( f(x) \) is a density function so that eq. (2.1) defines the expected value, denoted by \( E[g(x)] \). The crude estimator would sample \( x \) from \( f(x) \) and compute

\[
\hat{\xi} = \frac{1}{n} \sum_{i=1}^{n} g(x_i)
\]

However, we can also write eq. (2.1) as

\[
\xi = \int g(x) f(x) \frac{h(x)}{h(x)} dx
\]

where we can substitute \( h(x) \) another density function than \( f(x) \). Then we may sample \( x \) from \( h(x) \) and compute

\[
\hat{\xi} = \int g(x) f(x) \frac{h(x)}{h(x)} dx \quad \text{or} \quad \hat{\xi} = \frac{1}{n} \sum_{i=1}^{n} \frac{g(x_i)}{f(x_i)}
\]
where \( f(x)/h(x) \) may be interpreted as a weighing factor. The quantity \( \xi \) is estimated by the average of \( g^2(x_i) \) \( (i=1, \ldots, n) \), analogous to eq. (2.2). It can be derived that the minimum variance estimator \( g^2_0(x) \) results for

\[
h_0(x) = \frac{g(x) f(x)}{\xi}
\]

provided \( g(x) \geq 0 \) for all \( x \). In other words, we sample heavily from the important region of \( x \), i.e., from the region where \( x \) yields high values for the response \( g(x) \) unless the probability of such values is small. Unfortunately, we cannot calculate \( h_0(x) \) since it contains the unknown \( \xi \) itself! Nevertheless eq. (2.5) can suggest an adequate approximation to \( h_0(x) \). For instance, in [10, p. 166] the following integral is studied:

\[
\xi = \int_{\lambda v}^{\infty} \left( \frac{1}{\lambda} e^{-\lambda x} \right) dx \quad (\lambda, v > 0)
\]

Using \( f(x) = \lambda e^{-\lambda x} \) \( (x > 0) \) we obtain

\[
h_0(x) = \frac{\lambda}{\xi} x e^{-\lambda x} \quad \text{if } x \geq \nu
\]

\[
= 0 \quad \text{if } x < \nu
\]

One possible exponential approximation is to shift the original exponential distribution with parameter \( \lambda \) a distance \( v \) to the right. This reduced the variance drastically: for 4 combinations of \( \lambda \) and \( v \), the variance ranged between 0.7% and 6.5% of the original variance! Publication [10] gives many more references to importance sampling in non-dynamic situations.

3. Importance Sampling in Simple Dynamic Systems

In the simulation of dynamic systems IS is much harder to apply. Various IS approaches are summarized in [10, pp. 173-186]. Two other studies, however, form the basis for the present study. One approach is that of "virtual measures" introduced by Carter and Ignall [3]. The other approach, closer related to our study, is the method presented by Bayes [1]. The latter approach will be explained in the present section.

Consider a simple queuing system with one server and one customer source. We wish to estimate the probability of a queue \( q \) longer than some constant \( c \). This constant \( c \) is so high that, hopefully, the above probability is very small (rare event). This \( c \) may represent the size of a queuing area in a computer system, a doctor's office, etc. The crude estimator is

\[
P(q > c) = \frac{\text{total time during which } q > c}{\text{total simulated time}} \quad (3.1)
\]

Obviously the "rare event" \( q > c \) tends to happen more frequently when the system enters a "heavy loaded" period. In other words the rare event is expected to occur more frequently when a customer enters a system where already many customers are waiting, say \( q = 9 \). Bayes [1] proposes to repeat that part of the simulation run which started from such a situation; see the dotted lines in Fig. 4. He further proposes to stop such a "replication" (dotted line) as soon as the queue drops to the initial value, here \( q = 5 \).

Obviously we have to correct for the fact that the important regions are sampled more frequently. Therefore we take the averages of the times during which the rare event occurred \( \tau \), and the lengths of the important regions \( t \); the time it takes before the critical region is reached is denoted by \( \theta \); see Fig. 5 (next page). Summarizing, we wish to estimate

\[
P_{\text{IS}}(q > 15 | \text{crude}) = \frac{\sum \tau_i}{\sum (\theta_i + \tau_i)} \quad (3.2)
\]

The IS estimator with \( m \) replications \( (m > 1) \) - see Fig. 6 - is the analogue of eq. (3.1):

\[
P_{\text{IS}}(q > 15 | \text{IS}) = \frac{\sum \tau_i}{\sum (\theta_i + \tau_i)} \quad (3.3)
\]

with the averages per replicated important region \( i \):

\[
\bar{\tau}_i = \frac{\sum \tau_{ij}}{m}, \quad \text{and} \quad \bar{\tau}_i = \frac{\sum \tau_{ij}}{m}
\]

(See next page for Fig. 6.)

For a mathematical derivation of the "obvious" estimator (3.4) we refer to Appendix 2. Note that the original (background) simulation run is formed by continuing one arbitrary replication. It is convenient (and unbiased) to continue the last replication.

The variance of both the crude and the IS estimators in this simple simulation model, can be derived analytically, using the renewal or regenerative property. Systems with the renewal property may enter a specific state from which
FIG. 5. Importance sampling symbols $\theta$, $t$ and $\tau$

FIG. 6. Replicated important regions ($m=2$)
the simulation can be continued, "forgetting" the past (simulated) history. The simplest example is provided by a queuing system with Poisson (memoryless) arrivals: Once the system becomes empty (all servers idle) the system starts from scratch. In our particular case, it is enough to know that \( q \) jumped from 9 to 10 (event \( e_1 \)) in order to be able to continue the simulation run; remember that the assumed Poisson arrival and service processes imply a memoryless system.

Hence all cycles (epochs, tours) starting in \( e_1 \) are identically and independently distributed; an alternative renewal point is \( e_1 \). For further discussion of the renewal property in a simulation context we refer to \[17\]. Applying the regenerative property we prove in Appendix 3 that

\[
\text{var}(\hat{B}) = \frac{\sigma^2_{t} + 2p \text{cov}(t, t) + p^2 \sigma^2_{t+2}}{n(E(t+0))^2}
\]

for \( n \to \infty \) (3.7)

where \( n \) denotes the number of cycles. Using a similar derivation for \( \text{var}(\hat{B}_{IS}) \) we find that the gross variance reduction is

\[
\text{VR}_{\text{gross}} = \frac{m}{1+(m-1)p} \sigma^2_{t}/\sigma^2_{z} \quad (n \to \infty)
\]

where \( \sigma^2_{z} \) is a shorthand notation for the numerator of eq. (3.7). The net variance reduction corrects for the \( (m-1) \) extra subruns of length \( t \) with

\[
E(t) = P(q \geq 10) E(T)
\]

where \( T = \theta + t \). Hence the extra simulation length with which to correct the gross variance reduction, yields the factor

\[
\frac{(m-1) P(q \geq 10)}{E(T)} = \frac{(m-1) P(q \geq 10) + 1}{E(T)}
\]

so that the net variance reduction follows from eq. (3.8) and eq. (3.10):

\[
\text{VR}_{\text{net}} = \frac{m}{1+(m-1) P(q \geq 10)} \left( 1 + \frac{\sigma^2_{t}}{\sigma^2_{z}} \right) \left[ \frac{1}{1+(m-1) P(q \geq 15)} \right] \cdot \frac{1}{\sigma^2_{z}}
\]

(3.11)

It is not easy to see in which direction \( \text{VR}_{\text{net}} \) reacts to changes in the start of the importance region (e.g. starting from 12 instead of 10), and changes in the probability of the rare event (e.g. defining the rare event as \( q \geq 20 \) instead of \( q \geq 15 \)). We have not investigated this problem. However, in the next sections we do investigate a similar selection problem for our more complicated practical system.

The "importance boundary" denotes the start of the importance region in which \( m \geq 1 \) replications are simulated. If \( m = 1 \) then IS "degenerates" to crude sampling. In crude simulation the estimator of the steady-state blocking probability \( B \) is the average of \( M \) subrun probabilities \( B_k \):

\[
\hat{B} = \frac{\sum_{k=1}^{M} B_k}{M}
\]

with subrun probability estimator

\[
\hat{B}_k = \frac{L_k}{SS_k}
\]

(4.2)

where

- \( L_k \): number of calls blocked or "lost" in subrun \( k \),
- \( SS_k \): total number of calls in subrun \( k \) (sample size).

Since \( SS_k \) is kept constant in all subruns, we may drop the index \( k \), or \( SS_k = SS = 10,000 \).

The application of IS to the grading of Fig. 3 becomes troublesome because of the complexity of this system. In the preceding section a renewal state (completely specifying the system's state) was the value of the queue length \( q \), assuming Poisson, memoryless arrival and service processes. In theory, assuming Poisson processes for the grading, a possible renewal state could be defined by specifying for each individual line whether this "server" is busy or idle. However, there are as many as \( N = 45 \) lines so that a return to this specific system state will take very long, as the total number of possible states is \( 2^{45} \approx (3.5)(10^{13}) \). A renewal state does not necessarily yield a good starting point for an importance region. Fig. 3 shows that 15 busy lines can already block some customer source. The other extreme, all \( N = 45 \) lines busy, would imply that all customer sources are blocked. Normally calls get blocked before this extreme is reached.

An alternative starting point for the importance region is provided by the total number of busy lines, or TBL. Compared to the above renewal states, we ignore the identity of the lines. The arrows in Fig. 7 show that the importance region starts as soon as we cross the boundary line from below. This starting point is not a renewal state! To continue the simulation we would have to know not only the TBL value but also the identity of the busy lines. Even with Poisson processes, a different assignment

4. Importance Sampling in a Practical "Grading" System

Let us introduce the following terminology:
of identities would result in a different subsequent history. Therefore it does not make much sense to decide to end the importance replication when the line in Fig. 7 returns to the boundary line from above, a procedure described in the preceding section and originally proposed by Bayes [1].

Selecting TBL as a boundary condition does not provide a renewal state, but it does yield a more frequently occurring system state from which to start replications. Since it is no renewal state the length of the replication SS is made constant, instead of being dependent upon the return to the same TBL value from "above". The length of a replication is defined by the total number of generated calls. Replications starting from the same boundary point are made independent by the use of different random number streams.

Other boundary conditions may be considered. We restricted our study to the following options:

1. The total number of busy lines TBL; see above.
2. If all 15 lines serving one specific customer source are busy, and this particular source generates a call, then this call gets blocked. Therefore we start an importance region as soon as any customer source shows 15 busy servers.
3. Immediately after a call gets blocked, an importance region is started.

In pilot studies we found that the first two options do not lead to important regions in which many more calls get blocked than in the other regions (called 0 in Fig. 6). Therefore we shall concentrate in this paper on the results with the more promising option 3.

5. Results for a Specific Importance Boundary

As we mentioned in the preceding section we conjectured that an important region starts as soon as a call gets blocked. In other words we expect that lost calls are clustered. This conjecture is checked by performing a pilot simulation run, and measuring the number of calls between two consecutive blocked calls: "interarrival time" of blocked calls, or IA. The resulting frequency diagram is shown in Fig. 8 with double logarithmic scaling. This figure suggests that a good approximation is

\[
P(IA = k) = 0.175 k^{-1.06} \quad (k = 2, 3, \ldots, 512)
\]  

The mean and median are 67.8 and 12 respectively, i.e., the distribution is very asymmetric and suggests that the "rare events" (lost calls) occur in clusters. This result seems an encouraging indication of a useful importance boundary definition! The length of replication \( j \) (\( j = 1, \ldots, m \)) in the important region is denoted by a constant SSR. Within subrun \( k \) (\( k = 1, \ldots, 15 \)) the important region may be entered again later on; see index \( i \) below.

Consider a subrun \( k \) (without importance region replications, i.e., \( m = 1 \)). Let \( n_k \) denote the number of times an important region is entered within subrun \( k \). Hence \( n_k \) calls are blocked in subrun \( k \) outside the importance regions. Each importance region is replicated \((m-1)\) extra times, and has length SSR. The number of calls lost within an importance replication is denoted by IR. Since a subrun has total
length SS, the estimated blocking probability is
\[ \hat{r}_k = \left( \eta_k + \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{LR_{i,j}}{m} \right) / SS \] (5.2)

Obviously this estimator is an unbiased estimator of the steady-state blocking probability, since per important region each of the replications has the same expected value for the number of blocked calls, as the replication that forms part of the original (background) run.

To derive the variance of the importance sampling estimator of eq. (5.2), we denote the numerator by \( L_k \); the denominator is a constant. Note that \( L_k \) comprises a summation over a stochastic number of terms, \( n_i \) in the first \( \Sigma \) sign. Hence we use a well-known formula - see [9, p. 390] - namely:
\[ \text{var}(y) = \text{E}(\text{var}(y|x)) + \text{var}(\text{E}(y|x)) \] (5.3)

In Appendix 4 we derive that
\[ \text{var}(\hat{r}_k) = \left\{ \left( 1 + E(LR) \right)^2 \text{var}(\eta) + \frac{E(\eta) \text{var}(LR)}{m} / SS^2 \right\} \] (5.4)

In that appendix we assume that certain variables are independent, an assumption that seems realistic. If the assumption, however, would be violated, then positive correlation might be expected, so that eq. (5.4) would need the addition of some positive terms. Hence eq. (5.4) may be a lower bound, so that the derived variance reduction may be an upper bound.

The variance without importance sampling follows from eq. (5.4) by substituting \( m = 1 \). Hence the gross variance reduction (neglecting repeated sampling effort) is:
\[ \text{VR gross} = \left\{ \left( 1 + E(LR) \right)^2 \text{var}(\eta) + \frac{E(\eta) \text{var}(LR)}{m} / SS^2 \right\} \] (5.5)

where the symbols \( f_1 \) and \( f_2 \) are introduced to simplify the following presentation. The effect of repetitions in the importance region is shown by the factor \( 1/m \) in eq. (5.5). Those replications have more effect as the magnitude of \( f_2 \) is large relative to \( f_1 \). Obviously the sum of \( f_1 \) and \( f_2 \), i.e., the numerator in eq. (5.5), is independent of the partitioning of the total simulation run into "important" and "unimportant" regions. The shares of \( f_1 \) and \( f_2 \) in the constant \( c \), depend (among other things) on the length of the importance replication SSR. It is interesting to consider two limiting cases:

**Case 1: SSR = 0**

Since there are no replications in the important region, \( LR = 0 \). Hence
\[ c = f_1 + f_2 = (1 + 0)^2 \text{var}(\eta) + E(\eta) 0 = \text{var}(\eta) \] (5.6)

**Case 2: SSR approaches SS**

A subrun starts in an unimportant region. As soon as a call gets blocked, the rest of the subrun is replicated as an important region. (Hence SSR is a weak upper bound for SSR.) Consequently \( \eta_k = 1 \) and
\[ c = f_1 + f_2 = (1 + E(LR))^2 0 + 1 \text{var}(LR) = \text{var}(LR) \] (5.7)

Comparing cases 1 and 2, we see that \( f_2 \) is maximal relative to \( f_1 \), if SSR approaches SS. Considering eq. (5.5) this means that in that case the effect of replications in the important region is maximal. So we might jump at the conclusion that the length of the importance replication should be as long as possible. However, as the replication moves on, the effect of its starting point diminishes! What is the net effect of these two conflicting reasonings? We shall present some numerical results below.

The gross variance reduction in eq. (5.5) needs correction for the extra sampling effort \( ESE \), which has expected value
\[ E(ESE) = E(\eta)(m-1) \text{SSR} \] (5.8)

Hence the net variance reduction is
\[ \text{VR net} = \left\{ \frac{f_1 + f_2}{f_1 + \frac{f_2}{m}} \frac{SS + E(\eta)(m-1)\text{SSR}}{SS + E(\eta)(m-1)\text{SSR}} \right\} \] (5.9)

where \( f_1 \) and \( f_2 \) both depend on SSR. To maximize eq. (5.9) we need to select optimal values for SSR, length of replication, and \( m \), number of replications. Note that the factor \( m \) is not involved in any of the other factors in eq. (5.9). VRnet is maximal if its denominator is minimal. Hence we determine the partial derivative \( \partial VR / \partial SSR \) and solve \( \partial VR / \partial SSR = 0 \). The optimal number of replications is found to be
\[ m_0 = \frac{[f_2 (SS + E(\eta) \text{SSR})]^{1/2}}{E(\eta) \text{SSR}} \] (5.10)

The functions \( f_1(\text{SSR}) \) and \( f_2(\text{SSR}) \) are not explicitly known, so that we cannot compute the optimal value \( m_0 \) from eq. (5.10). Neither can we compute the optimal SSR as \( \partial VR / \partial \text{SSR} \) cannot be made explicit. Therefore we estimate \( f_1 \) and \( f_2 \) besides \( E(\eta) \), for various SSR values, using a pilot simulation run. In this simulation run \( m = 1 \) so that no importance sampling is needed! This results in Table 1 (see next page) which we can explain as follows.

(1) As columns 2 and 3 show, when a greater part (SSR) of the total subrun is considered as forming an importance region, then the remaining number of blocked calls (new entries of an importance region) necessarily decreases, i.e., \( \xi \) decreases where \( \xi \) is the average of 15 subruns.
The gross variance reduction was 1.0085 but the net variance reduction (accounting for the one extra run per importance region) was only 0.859. Of course this estimate could again be inaccurate.

The above conclusions are based on estimates only, but fortunately the numbers in columns 6 and 7 do not show wild oscillations. We checked our conclusions by actually executing an importance sampling experiment with \( \text{SSR} = 25 \) and \( m = 2 \). The gross variance reduction was 1.0085 but the net variance reduction (accounting for the one extra run per importance region) was only 0.859. In other words, the variance even increased by 14%. Of course this estimate could again be inaccurate.

### 6. Conclusion

Importance sampling was originally developed for the evaluation of integrals such as eq. (2.1). In that area dramatic variance reductions have been realized, e.g., a factor 100. The extension of this technique to dynamic (auto-correlated) systems was tried by several authors. The variant that inspired our study was developed by Bayes [1] and shows some relationships with the "virtual measures" of Carter and Ignall [3]. However, we applied Bayes' procedure to a much more complicated system, namely a server network or "grading" occurring in telephone exchanges. In such a grading renewal state could be detected but they could not be utilized since the return to a renewal state takes too long for practical purposes, and does not necessarily start an important region.

The crucial issue is to define situations (states) which initialize an "important region", i.e., a part of the simulation run in which many important - but rare events are expected to occur. Three alternative "importance boundaries" were investigated. This report concentrated on the boundary that seemed most promising, namely, an important region starting immediately after a customer (call) gets blocked. The reason for this choice was that some lost calls tend to occur in clusters.

Next we were confronted with two tactical questions: how long to continue sampling in the importance region, and how often to repeat this sampling? We derived a formula for the (net) variance reduction (correcting for the additional sampling effort). This formula could not be solved analytically for the optimal sampling length and replication number. Therefore estimates were substituted based on a pilot simulation run. The results indicate that the way we applied importance sampling in our particular system, resulted in a net variance increase!

The lesson for practitioners may be not to use importance sampling in complicated, dynamic systems simulation, since the resulting variance reduction may very well be poor. Moreover, its application is not so straightforward as that of some other variance reduction techniques. Nevertheless a side-benefit was that during our analysis we gained an improved understanding of the way our system behaves as a stochastic process. Our study may be of interest to theoreticians, in so far as it provides a challenge to improve our importance sampling technique which seems of particular value in systems characterized by "rare" events.

### Notes

1) For readers familiar with traffic engineering we notice that at this point we ignore the hunting order of the selectors (traffic sources).

2) We assume identical Poisson customer sources and exponential service times. Then a Markov process results. This system would require the solution of 28 equations (2m = 3.5 x 10^{13}). For Markov systems occurring in telecommunications, we refer to Olsson, e.g., [12].

3) Our terminology is such that \( m \) "replications" means that \( m \) "replication" is part of the background or base run, and \( m-1 \) "replications" are duplicates.

4) Observe that \( \tau_i \) may consist of non-consecutive epochs during which \( q \geq 15 \), within the \( j \)th replication. Further \( t \) is the time between the events \( \varepsilon_i \) and the next event \( \varepsilon_j \), and \( \theta \) is the time between \( \varepsilon_2 \) and a next event \( \varepsilon_1 \).
5) Note that $E(t)/E(t+\hat{t}) \neq E[r/(t+\hat{t})]$, so that the ratio estimator of eq. (3.3) is a biased estimator. Asymptotically this estimator becomes unbiased. Alternative ratio estimators are surveyed in [6]. However, in crude estimation it is possible to fix the total simulation runlength so that the denominator of eq. (3.3) becomes deterministic.

6) Replications starting at a "later" boundary point (say, the right-hand arrow in Fig. 7), are theoretically dependent on the previous history, and hence on the last replication of the preceding importance region. If importance regions are "far" apart, this dependence may be ignored for practical purposes.

7) Fig. 8 shows that all observations are close to a linear line, with the exception of the starting point, denoted by O.

8) Actually Fig. 9 is misleading in so far as replications may end at different points of time, since each replication length is determined by a fixed number of calls. For completeness sake we further mention that the simulation is started in the empty state (all lines free), and the total run is cut into 15 subruns, each comprising 10,000 calls. No subrun starts in an important region.

9) $$\frac{\delta(\text{denominator})}{\delta m} = f_2 m \frac{E(\eta)}{SSR} m^{-2} +$$

$$+ E(\eta) \frac{SSR}{f_1 - SS} f_2 m^{-2}$$

It is easy to check that eq. (5.10) defines a minimum indeed, and not a maximum or saddle-point.

10) We compute

$$LR = \frac{1}{15} \sum_{k=1}^{15} \frac{n_k}{\eta} \sum_{i=1}^{n_k} LR_{ki1}$$

and

$$SSR_k = \frac{1}{\eta} \sum_{i=1}^{n_k} LR_{ki1}^2 - \left( \frac{1}{\eta} \sum_{i=1}^{n_k} LR_{ki1} \right)^2 / n_k$$

so that

$$SSR = \frac{1}{15} \sum_{k=1}^{15} SSR_k$$

Then

$$\hat{t}_1 = (1 + \bar{R})^2 s_\eta$$

and

$$\hat{t}_2 = \bar{R} S^2_{SR}$$

11) Remember that below eq. (5.4) we noted that if actually some variables are dependent then our formula gives an upper bound for the variance reduction, so that this 6.3% is an estimated upper bound.
An estimator for \( P(q \geq IB) \) is

\[
\hat{P}(q \geq IB) = \sum_{i=1}^{n} \frac{\theta_i}{\theta_i + \bar{\epsilon}_i}
\]  \hspace{1cm} (A2.8)

where

\[
\bar{\epsilon}_i = \frac{\theta_i + \bar{\epsilon}_i}{T}
\]  \hspace{1cm} (A2.9)

and

\[
T = \sum_{i=1}^{n} (\theta_i + \bar{\epsilon}_i)
\]  \hspace{1cm} (A2.10)

so that

\[
\hat{P}(q \geq IB) = \frac{\sum_{i=1}^{n} \bar{\epsilon}_i}{\sum_{i=1}^{n} (\theta_i + \bar{\epsilon}_i)}
\]  \hspace{1cm} (A2.11)

Substitution of eq. (A2.7) and eq. (A2.11) into eq. (A2.1) yields

\[
\hat{P}(q \geq 15) = \frac{\sum_{i=1}^{n} \bar{\epsilon}_i}{\sum_{i=1}^{n} (\theta_i + \bar{\epsilon}_i)}
\]  \hspace{1cm} (A2.12)

Appendix 3: Variances of Estimators in a Simple Queuing System

We derive \( \text{var}(\hat{\beta}) \) following [7]. Define

\[
z_i = \tau_i - p(t_i + \theta_i)
\]  \hspace{1cm} (A3.1)

so that

\[
E(z_i) = 0
\]  \hspace{1cm} (A3.2)

and

\[
\sigma_z^2 = \text{var}(z_i)
\]  \hspace{1cm} (A3.3)

Note that the \( z_i \) are independent because of the renewal property. We further have

\[
\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i
\]  \hspace{1cm} (A3.4)

so that \( \bar{z} \sim N(0, \sigma_z^2) \) for \( n \to \infty \). Hence

\[
\bar{z} = \frac{\sigma_z}{\bar{z}} = \frac{\sigma_z}{\sqrt{n} \sigma_z} = \sqrt{n}
\]

\[
\sigma_z = \sqrt{n} (\bar{z} - \bar{\zeta})
\]  \hspace{1cm} (A3.5)

has a \( N(0,1) \) distribution for \( n \to \infty \). Substituting eq. (A3.1) with \( \bar{z} = \bar{\zeta} + \bar{\zeta} \) yields

\[
\frac{\bar{z}}{\sigma_z} = \sqrt{n} \left( \frac{\bar{z}}{\sigma_z} - \bar{\zeta} \right) \approx \sqrt{n} \left( \frac{\bar{z}}{\sigma_z} - \bar{\zeta} \right)
\]  \hspace{1cm} (A3.6)

which has an asymptotic standard normal distribution. Consequently

\[
\text{var}(\hat{\beta}) = \frac{\sigma_z^2}{\bar{z}^2} = \frac{\sigma_z^2}{\bar{z}^2} \quad (n \to \infty)
\]  \hspace{1cm} (A3.7)

For \( \text{var}(z) \) we can write

\[
\sigma_z^2 = \sigma_t^2 - 2 \sigma_t \sigma_\theta \cos(\theta_t + \theta) + \sigma_\theta^2
\]  \hspace{1cm} (A3.8)

where \( \sigma_t, \sigma_\theta \) are independent, eq. (A3.7) and eq. (A3.8) yield

\[
\text{var}(\hat{\beta}) = \frac{\sigma_t^2 - 2 \sigma_t \sigma_\theta \cos(\theta_t + \theta) + \sigma_\theta^2}{n \bar{z}^2}
\]  \hspace{1cm} (A3.9)

When we apply Importance Sampling, \( \theta_i \) is followed by several timepaths of length \( t_{ij} \). We obtain several \( \tau_{ij} \)'s. As an estimator we use:

\[
\hat{P}_{IS}(q \geq 15) = \frac{1}{m} \sum_{i=1}^{m} \left( \theta_i + \bar{\epsilon}_i \right)
\]  \hspace{1cm} (A3.10)

where

\[
\bar{\tau}_{ij} = \frac{1}{m} \sum_{i=1}^{m} \tau_{ij}
\]  \hspace{1cm} (A3.11)

and

\[
\bar{\bar{\tau}}_{ij} = \frac{1}{m} \sum_{i=1}^{m} \bar{\tau}_{ij}
\]  \hspace{1cm} (A3.12)

where \( m \) denotes the number of replications. Note that eq. (A3.10) is biased. To find the variance we proceed analogous to (A3.1): Define

\[
\tau_i = \bar{\tau}_{ii} - p(t_i + \theta_i)
\]  \hspace{1cm} (A3.13)

so that

\[
\bar{\tau} = \frac{1}{m} \sum_{i=1}^{m} \tau_i
\]

We can relate \( \sigma_{\bar{\tau}}^2 \) to \( \sigma_z^2 \) as follows. Consider the three terms of eq. (A3.14):

(i) \( \sigma_{\bar{\tau}}^2 = \frac{\sigma_z^2}{m} \)  \hspace{1cm} (A3.15)

because of the independence of the replications.

(ii) \( \sigma_{\bar{\tau}+\bar{\theta}}^2 = \sigma_{\bar{\tau}} \sigma_{\bar{\theta}} = \frac{\sigma_z \sigma_\theta}{m} \)  \hspace{1cm} (A3.16)

where the first equality holds, since \( \bar{\tau} \) and \( \bar{\theta} \) are independent.

(iii) \( \sigma_{\bar{\tau}+\bar{\theta}}^2 = \sigma_{\bar{\tau}}^2 + \sigma_\theta^2 = \frac{\sigma_z^2 + \sigma_\theta^2}{m} \)  \hspace{1cm} (A3.17)

where the first equality holds since \( \bar{\tau} \) and \( \bar{\theta} \) are independent. Consequently eq. (A3.14) becomes:

\[
\sigma_{\bar{\tau}}^2 = \frac{1}{m} \sigma_z^2 + \frac{p^2}{m} \sigma_\theta^2
\]  \hspace{1cm} (A3.18)
For \( \text{var}(\hat{L}) \) we find:

\[
\text{var}(\hat{L}) = \frac{1}{n} \frac{\sigma^2 \left( \frac{2}{m} + \frac{m-1}{m} \frac{2}{\sigma^2} \right)^2}{\left( \frac{2}{m} \frac{2}{\sigma^2} \right)^{n+1}} \quad (n \to \infty) \quad (A3.19)
\]

Hence the gross variance reduction is:

\[
\frac{\sigma^2}{\sigma^2} = \frac{m}{1 + (m-1) \frac{2}{\sigma^2}} \quad (n \to \infty) \quad (A3.20)
\]

In the main text we derived the net variance reduction:

\[
\text{VR}_{\text{net}} = m \left( 1 + (m-1) P(q^{\geq 10}) \right) \left( 1 + (m-1) \frac{2}{\sigma^2} \right)^{-1} \quad (A3.21)
\]

with the constant \( \sigma^2 = \sigma^2 \). Since \( P < 1 \) we can write:

\[
\text{VR}_{\text{net}} = \frac{m}{\left( 1 + (m-1) P(q^{\geq 10}) \right) \left( 1 + (m-1) \frac{2}{\sigma^2} \right)} \quad (A3.22)
\]

As the importance boundary increases, the replications decrease in lengths. Hence \( \sigma^2 \) increases and \( \sigma^2 \) decreases (compare the geometric distribution). This yields less variance reduction. If then, however, \( \sigma^2 \) decreases, the effect increases!

Appendix 4: Variance of \( \hat{R}_k \)

Applying eq. (5.3) to the numerator, denoted by \( L_k \), of eq. (5.2) yields

\[
\text{var}(L_k) = E(\text{var}(\hat{n}_k + \sum_{i=1}^{\eta} \sum_{j=1}^{m} LR_{ij} | n_k = \eta)) + \eta \sum_{i=1}^{n} \sum_{j=1}^{m} LR_{ij} | n_k = \eta) + m \text{var}(E(\hat{n}_k + \sum_{i=1}^{\eta} \sum_{j=1}^{m} LR_{ij} | n_k = \eta)) = T_2 + T_1 \quad (A4.1)
\]

We assume that \( n_k \) and \( LR_{ij} \) are independent, which is a realistic assumption. We know that \( LR_{ij} \) is independent of the other replications (using different random numbers), say \( LR_{ij'} \). Finally, we ignore possible dependence between replications in subsequent encounters with an important event within the same subrun \( k \), i.e., \( LR_{ij} \) and \( LR_{ij'} \) are assumed to be independent. This assumption is realistic if important regions are "far" apart so that autocorrelations vanish. Hence

\[
T_2 = E(n \text{var}^\frac{1}{m} \sum_{j=1}^{m} LR_{ij} | n_k = \eta)) = \frac{1}{n} E \text{var}(LR) \quad (A4.2)
\]

and

\[
T_1 = var(n + \eta E(LR_k | n_k = \eta)) = \frac{1}{n} (1 + \text{var}(LR))^2 \text{var}(\eta) \quad (A4.3)
\]

Hence

\[
\text{var}(\hat{R}_k) = \frac{\text{var}(L_k)}{SS^2} = \frac{(1 + \text{var}(LR))^2 \text{var}(\eta) + \text{E}(\eta) \frac{1}{m} \text{var}(LR)}{SS^2} \quad (A4.4)
\]

References


8. KAHN, H. and A.W. MARSHALL, Methods of reducing sample size in Monte Carlo computations. JOURNAL OPERATIONS RESEARCH SOCIETY OF AMERICA, 1, no. 5, Nov. 1953, pp. 263-278.

9. KEEPing, E.S., INTRODUCTION TO STATISTICAL INFERENCE. D. van Nostrand Company, Inc.,


12. OLSSON, K.M., Accuracy of observed mean values of waiting times, queue lengths and proportion of delayed calls in Markovian many-server systems. ERICSSON TECHNICS, no. 2, 1977, pp. 106-123.

13. SCHRUBEN, L.W., Reply to Fox. MANAGEMENT SCIENCE, 24, no. 8, April 1978, pp. 862.