Analysis of state-independent IS measures for the two-node tandem queue
— extended abstract for RESIM2004 —

Pieter-Tjerk de Boer*

Abstract

The performance of state-independent changes-of-measure for estimating the overflow probability of the total population of a two-node Markovian tandem queue using importance sampling is studied. It is shown that only the change of measure proposed by Parekh and Walrand (1989) can be asymptotically efficient for large overflow levels. The behaviour of this change of measure all over the parameter space is categorized, using both analysis and numerical calculation.

1 Introduction

Since the end of the 1980s, there has been a steady interest in the problem of estimating rare event probabilities in queueing models, particularly networks of queues. This interest derives mainly from applications in the field of telecommunications, where buffers in routers and switches should be dimensioned such that overflow is a rare event.

The present work considers the use of *importance sampling* (IS) simulation to estimate the probability of overflow of the *total* population of two (M)/M/1 queues in tandem. Importance sampling simulation artificially makes the overflow less rare by changing the probability distributions (also called "change of measure" or "tilting") in the model, and compensating for this by keeping track of the likelihood ratio. In the present work, only *state-independent* changes of measure are considered; this means that only a global change of the arrival and service rates, without regard to the model's present state, is made.

In rare-event simulation of queueing models, one typically strives for *asymptotic efficiency*: the relative error (defined as the estimator's standard deviation divided by its mean) increases at most polynomially in the overflow level (at a fixed number of simulation replications), while the overflow probability itself decreases exponentially. An even better estimator may have *bounded relative error*: the relative error stays essentially constant as the overflow level increases. However, a worse change-of-measure may lead to an exponentially growing relative error (as does standard simulation), or even to the very undesirable situation where the estimator variance (and thus the relative error) becomes *infinite*.

In [PW89], a state-independent change-of-measure was proposed for Markovian queueing models, based on large-deviations theory; according to this change of measure, in the tandem queue model the arrival rate and the bottleneck service rate should be exchanged. This change of measure will henceforth be referred to as P&W. Experimental results in [PW89] showed that this method works well in some cases, and not so well in other cases. In [GK95] the P&W method was studied in detail for tandem networks; it was proven for some parts of the parameter space (i.e., ranges of arrival and service rates) that the resulting simulation is asymptotically efficient, and for some other parts that it is not. Experimental results in [dB00] and [dBN02] indicate that, at the cost of a significant increase in complexity, asymptotically efficient simulation is possible everywhere using a state-dependent change of measure.

So on the one hand we have P&W's state-independent tilting, which in several cases does not work well, and on the other hand we have the state-dependent tilting which works well even in those cases, but is much more complex. A natural question then is whether *other* state-independent tiltings than P&W could be asymptotically efficient. The fact that P&W is derived on the basis of large-deviations theory, and the fact that an adaptive approach to finding a state-independent tilting fails to result in asymptotic efficiency in some cases (see Section 7.4.4 in [dB00]), suggest that this might not be the case, but of course do not constitute a formal proof.

In the present work, the above question is answered for the two-node tandem queue, and extensions of [GK95]'s results to the entire parameter space are given.

2 Analytical results

This section lists some theorems with brief descriptions of the proofs; the full proofs will appear in a future full version of this paper.

* Dept. of Electrical Eng., Mathematics and Computer Science, University of Twente, Enschede, The Netherlands; `ptdeboer@cs.utwente.nl`

All of these theorems use the following notation: λ = the (untilted) arrival rate at the first queue, μ = the (untilted) service rate at the first server, and ν = the (untilted) service rate at the second server. Furthermore, λ', μ' and ν' represent the tilted rates.

We limit the study to cases where both queues are stable, i.e. $\lambda < \mu$ and $\lambda < \nu$; otherwise, the overflow event would not be rare. Furthermore, we limit ourselves to the case where the second server is the bottleneck, i.e., $\mu > \nu$.

Theorem 1 *For the two-node tandem queue simulation problem, no state-independent change of measure is asymptotically efficient if*

$$\mu^2\nu^4 > (1-2\nu)((1-\mu)^2 + \mu^2\nu^2).$$

The proof starts by defining two sample paths. Their contributions to the first and second moments of the estimator are calculated; from this, two necessary conditions for asymptotic efficiency follow. The inequality in the theorem is true in the region where these two conditions are contradictory.

Theorem 2 *For the two-node tandem queue simulation problem, every state-independent change of measure for which either $\lambda' \neq \nu$, or $\mu' \neq \mu$, or $\nu' \neq \lambda$, is not asymptotically efficient. I.e., the P&W tilt is the only one that may be asymptotically efficient.*

The proof considers starting the simulation from any initial state and proposing a "guess" for the second moment that would be obtained (as a function of the initial state); next, it is shown that with any non-P&W tilting this is actually a lower bound on the true second moment; finally it is established that this lower bound already increases so quickly with the buffer size that asymptotic efficiency is impossible.

Theorem 3 *For the two-node tandem queue simulation problem, a state-independent change of measure according to P&W (i.e., with $\lambda' = \nu$, $\mu' = \mu$ and $\nu' = \lambda$) is not asymptotically efficient if*

$$\nu > \frac{3-\mu}{2} - \frac{1}{2}\sqrt{-3\mu^2 - 2\mu + 5}.$$

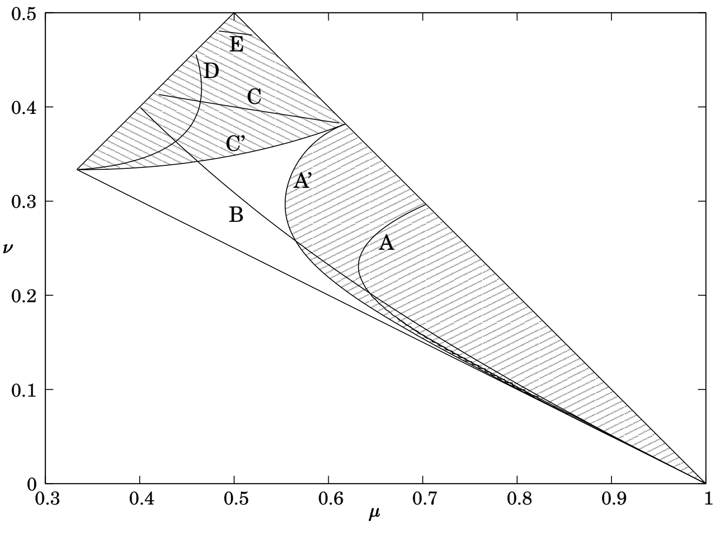
The proof uses a similar approach to the one used for the previous theorem.

Theorem 4 *For the two-node tandem queue simulation problem, a state-independent change of measure according to P&W (i.e., with $\lambda' = \nu$, $\mu' = \mu$ and $\nu' = \lambda$) leads to infinite variance if*

$$M^2 < \frac{4\nu\mu\lambda}{1-M},$$

with M defined as $M = \frac{(\lambda+\mu)^2}{\mu+\nu}$.

The proof uses a similar approach to the one used for the previous two theorems. The difference however is that the "guess" is such that multiplying it by any positive number does not invalidate the proof that it is a lower bound on the true second moment; consequently, the second moment must be infinite.



Right of A: P&W is a.e. according to [GK95]

Right of A': P&W is a.e. according to Theorem 5

Above B: a.e. proof turns into bounded relative error proof ([GK95])

Above C: P&W is not a.e. according to [GK95]

Above C': P&W is not a.e. according to Theorem 3

Left of D: P&W gives infinite variance according to Theorem 4

Above E: no state-independent tilting is a.e., according to Theorem 1 (but according to Theorem 2, this is actually true everywhere above C')

Figure 1: Analytically found boundaries in μ, ν space

Theorem 5 *For the two-node tandem queue simulation problem, the P&W change of measure is asymptotically efficient if*

$$\frac{\mu+\nu}{(\lambda+\mu)^2}\left(\nu+\lambda\mu\left(\frac{1}{\lambda+\mu}+\frac{\lambda\mu^2}{(\mu-\nu)\nu^2}\right)\right)<1.$$

This proof uses the same approach as the one for Theorem 5.7 in [GK95], but at two places uses tighter bounds.

The results from all the above theorems (and from [GK95] for comparison) are summarized in Figure 1.

3 Numerical study of P&W

For a relatively simple model like the two-node tandem queue, it is feasible to *numerically* calculate the expectation of the second moment (and of the first moment, i.e., the overflow probability itself) for a given tilting and a given overflow level. By repeating this for several overflow levels and studying the dependence of the resulting variance on the overflow level, one can classify the simulation as asymptotically efficient or not. Furthermore, cases of infinite variance can be identified. Doing this on a grid of points in the μ, ν parameterspace for the P&W tilting allows us to produce a complete picture of the behaviour of P&W, also in the areas where neither of the Theorems 3, 4 and 5 apply (i.e., the white area in Figure 1).

The result of a scan of the entire μ, ν parameter space for the two-node tandem queue with P&W tilting is shown in Figure 2. Note that, in contrast to other results in this paper, this scan also covers the area in which the first queue is the bottleneck (i.e, $\mu < \nu$).

Theorems 3 and 4 provide sufficient conditions for exponential growth and infinite variance; these results (i.e., curves C' and D from Figure 1) have also been drawn in the figure, for comparison with the numerical results. Clearly, the numerically found boundaries follow the sufficient conditions rather closely.

Finally, note that although the proofs only imply bounded relative error above line B in Figure 1, the numerical results suggest that wherever the simulation is asymptotically efficient, it also has a bounded relative error; asymptotic efficiency without bounded relative error was never observed in the numerical study.

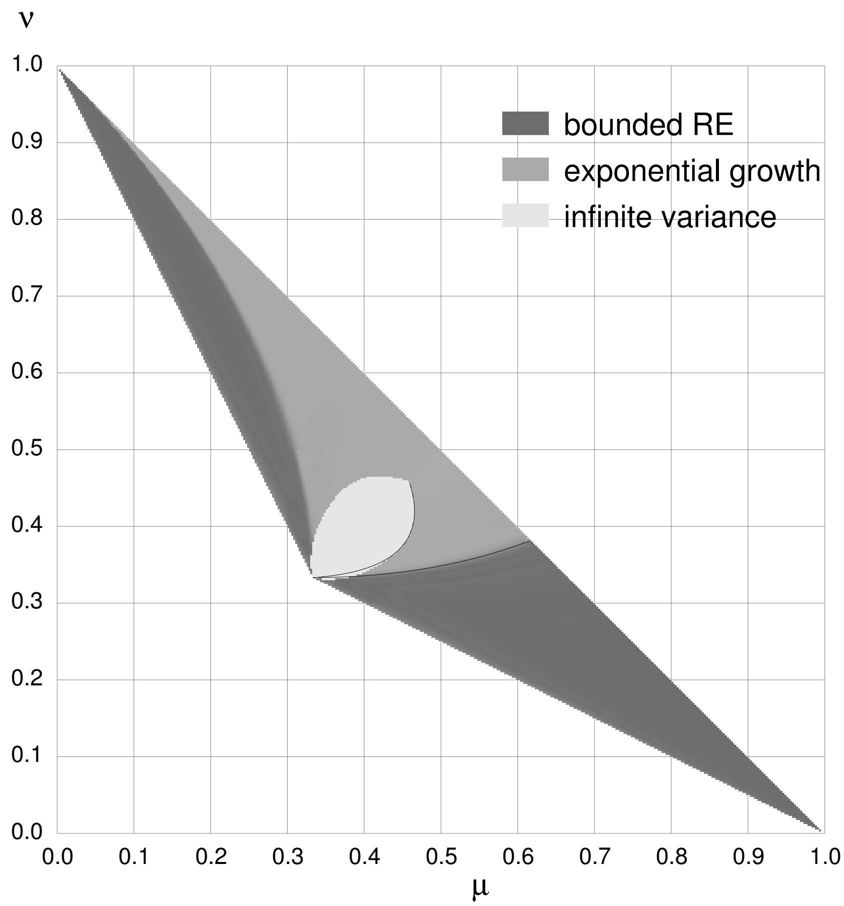


Figure 2: Behaviour of the two-node tandem queue simulation under P&W tilting

4 Conclusions

In this paper, a detailed analysis of simulating a Markovian two-node tandem queue using importance sampling with a state-*in*dependent change of measure has been presented. The main conclusion is that the only such change of measure that may be asymptotically efficient, is the one heuristically derived on the basis of large deviations theory in [PW89]. Furthermore, the behaviour of the simulation under this change of measure has been characterized further, thus extending the results from [GK95].

Obviously, simulating this particular two-node queueing network is not very interesting by itself; typically, the overflow probability can be calculated numerically. However, the results from this paper do have implications for future research in the general area of efficient simulation of queueing models.

First of all, the fact that P&W is on the one hand not asymptotically efficient in parts of the parameter space, and on the other hand is the only state-independent tilting that may be asymptotically efficient at all, means that more complicated changes of measure are needed. Formally, this has only been established for this particular model, but given this model's simplicity, it is to be expected that this is also true for a large class of other models. This gives a strong motivation for research into more complicated methods, like the adaptive state-dependent method from [dBN02]. Also, it means that methods (like [dBKR02]) that adaptively try to find the best state-independent tilting (rather than just assuming P&W) at least for this model will never be asymptotically efficient when P&W is not. (Still, such methods should not be discarded; they are simple and can still be used in the many situations where an asymptotically efficient state-independent tilting does exist; and when the overflow level of interest is not too high, even a non-asymptotically efficient simulation may be acceptable.)

Secondly, the results regarding the limitations of P&W tilting are of use in testing new simulation methods. It is particularly interesting to test new simulation methods in parameter regions where P&W is not asymptotically efficient or yields infinite variance. Previously, P&W was only known to be not asymptotically efficient in a rather limited region of the parameter space (above curve C in Figure 1). The analysis and numerical study in this paper have extended this region. Furthermore, we now know that there is also a region in which the variance under P&W is actually infinite. (Infinite variance under P&W previously was known for a different system, namely a two-node tandem queue with feedback, see [RJ02].)

Future extensions of this research could focus on finding analytical results for the parameter regions that are not covered by the present theorems; on generalizing the statement that only P&W can be asymptotically efficient (Theorem 2) to a larger class of models; and on non-Markovian queueing networks.

References

[dB00] Pieter-Tjerk de Boer. *Analysis and efficient simulation of queueing models of telecommunication systems*. PhD thesis, University of Twente, 2000.

[dBKR02] P. T. de Boer, D. P. Kroese, and R. Y. Rubinstein. A fast cross-entropy method for estimating buffer overflows in queueing networks. In *Fourth Workshop on Rare Event Simulation and Related Combinatorial Optimization Problems, RESIM/COP'02*, 2002.

[dBN02] Pieter-Tjerk de Boer and Victor F. Nicola. Adaptive state-dependent importance sampling simulation of markovian queueing networks. *European Transactions on Telecommunications*, 13(4):303–315, 2002.

[GK95] Paul Glasserman and Shing-Gang Kou. Analysis of an importance sampling estimator for tandem queues. *ACM Transactions on Modeling and Computer Simulation*, 5(1):22–42, January 1995.

[PW89] S. Parekh and J. Walrand. A quick simulation method for excessive backlogs in networks of queues. *IEEE Transactions on Automatic Control*, 34:54–66, 1989.

[RJ02] R. S. Randhawa and S. Juneja. Simulating rare events by combining importance sampling and temporal difference methods. Draft, 2002.

[Sad91] J. S. Sadowsky. Large deviations theory and efficient simulation of excessive backlogs in a $GI/GI/m$ queue. *IEEE Transactions on Automatic Control*, 36:1383–1394, 1991.