A note on a state-independent change of measure for the $G|G|1$ tandem queue.

Anne Buijsrogge  
University of Twente  
a.buijsrogge@utwente.nl

Pieter-Tjerk de Boer  
University of Twente  
p.t.deboer@utwente.nl

Werner R.W. Scheinhardt  
University of Twente  
w.r.w.scheinhardt@utwente.nl

October 23, 2015

Abstract

In this report we compare two methods to find a state-independent change of measure in the $G|G|1$ tandem queue and we show that they are equivalent.

1 Introduction

Importance sampling is one of the methods used in rare-event simulation. In that method, the event of interest is made less rare by changing the underlying probability distribution. This change of the probability distribution is also called the change of measure or tilting. During the simulation, one keeps track of the likelihood ratio, which is the ratio between the probability distribution in the original system and the probability distribution in the changed system. The results of the simulation are weighted by this ratio and therefore we obtain unbiased estimators.

One of the first papers to consider importance sampling in queueing networks is [5] by Parekh and Walrand. They consider the probability that the total number of customers in the system reaches some level $N$ in a busy cycle. To estimate this probability efficiently using simulation, they propose a simple change of measure for both single server queues and networks of queues. Their proposed change of measure is state-independent, i.e. the change of measure does not depend on the state of the system. It entails that for the $M|M|1$ tandem queue, the change of measure is to exchange the arrival rate with the lowest service rate. In the remainder of this report, the change of measure proposed by Parekh and Walrand will be referred to as the P&W change of measure.

In [7], Sadowsky shows that the P&W change of measure is asymptotically efficient for the single $G|G|\infty$ queue. The notion of asymptotic efficiency means that the number of runs required in order to achieve a certain relative error, which is the standard deviation divided by the mean, grows less than exponentially with the overflow level.
Although it has been shown that the P&W change of measure works well for the single $G|G|1$ queue, Glasserman and Kou show in [4] that for the $M|M|1$ tandem queue the P&W change of measure may or may not be asymptotically efficient. They provide both necessary conditions and sufficient conditions for asymptotic efficiency. In [2], De Boer extends these results, but also shows that the P&W change of measure is the only state-independent change of measure that can possibly be asymptotically efficient for the $M|M|1$ tandem queue.

Much research has been done on efficient simulation of the $M|M|1$ tandem queue, but to the best of our knowledge there are no proven asymptotically efficient results for the $G|G|1$ tandem queue. In [3], Frater and Anderson use the results of P&W to describe a state-independent change of measure for a class of tandem $G|G|1$ queues more explicitly. However, in their paper no proofs for efficiency of the change of measure are given. In this report, we will consider the state-independent change of measure proposed by Parekh and Walrand for the $G|G|1$ tandem queue. In Section 3 we will show that the results of Frater and Anderson are equivalent to another idea proposed by Rosen, De Boer and Scheinhardt in [6]. The conclusions are presented in Section 5.

2 Model and preliminaries

2.1 The model

In this report, we consider $d$ $G|G|1$ queues in tandem. Let $A_k$ be the inter-arrival time between customers $k$ and $k+1$ and let $B_k^{(j)}$ be the service time of customer $k$ at queue $j$. All processes are i.i.d. and are independent of each other. After service completion at queue $j$, the customer enters queue $j+1$, so there is no probabilistic routing. When the customer finishes service at queue $d$, the customer leaves the system. Starting with customer 1 in queue 1 and all other queues empty, we are interested in the event that there are $N$ customers in the system before the system is empty again. We define $K_N$ as the index of the first customer who reaches the overflow level $N$. Likewise, $K_0$ is the index of the first customer after customer 1 who sees an empty system upon arrival.

The indicator $1_{K_N<K_0}$ defines if we have reached our rare event or not. The probability of this rare event, denoted by $p_N$, is equal to $E[1_{K_N<K_0}]$. The state of the system is the number of customers in each queue. We assume that the system is stable, that is, $E[B^{(j)}] < E[A]$ $\forall j = 1, ..., d$. However, if one of the queues is unstable, our event of interest is not rare and therefore no importance sampling is needed in order to obtain a good estimation of the probability of the event. In contrast to the results obtained in [2] and [4] for the $M|M|1$ tandem queue, where a discrete time Markov chain is considered, the model we are concerned with here is a continuous time queuing system.

We will denote the moment generating function of distribution $X$ by $M_X(t)$. Throughout this report we assume that the moment generating functions of the service time distributions of queue $j$, $M_{B^{(j)}(t)}$, exist for some $t > 0$.

2.2 Importance sampling simulation

In importance sampling, the rare-event is made less rare by changing the underlying probability distribution. For a single $G|G|1$ queue, it is shown in [7] that making an
exponential tilt $\theta$ for both the inter-arrival times and service times in the following way

$$
\begin{align*}
\text{d}F^\theta_A(a) &= \frac{e^{-\theta a}}{M_A(-\theta)} \text{d}F_A(a), \\
\text{d}F^\theta_B(b) &= \frac{e^{\theta b}}{M_B(\theta)} \text{d}F_B(b),
\end{align*}
$$

(1)

(2)
gives an asymptotically efficient estimator when $\theta$ is the solution of $M_A(-\theta)M_B(\theta) = 1$. In this expression, $F^\theta_A$ and $F^\theta_B$ denote the distribution functions of the inter-arrival times and service times respectively. The superscript $\theta$ denotes the exponential tilt $\theta$.

Let us now consider $d\ G|G|1$ queues in tandem. As we will see, in both of the methods explained in Section 3 only one of the service time distributions will be tilted; the other service time distributions remain the same. Which queue to tilt and what the exponential tilt parameter $\theta$ should be is explained later in Section 3. When the rare event is reached, let the likelihood ratio $L$ be,

$$
L = \prod_{k=1}^{K_N-1} \frac{\text{d}F_A}{\text{d}F^\theta_A}(A_k) \prod_{k=1}^{K_N-\sum_{k=1}^{n_k+1} n_j > 0} \frac{\text{d}F_B(j)}{\text{d}F^\theta_B(B_k)},
$$

(3)

if queue $j$ is exponentially tilted and where $n_i$ is the number of customers in queue $i$ upon reaching level $N$. Let $\mathbb{E}^\theta[X]$ be the expected value of $X$ under the change of measure with exponential tilt $\theta$. Then under the tilt $\theta$, $L1_{K_N<K_0}$ is an unbiased estimator for $p_N$, i.e., $p_N = \mathbb{E}^\theta[L1_{K_N<K_0}]$.

### 3 Description of the two methods

Frater and Anderson presented one way to obtain a change of measure in the $G|G|1$ tandem queue in the early 90s in [3]. No proofs of asymptotic efficiency of this change of measure are provided. Independently of Frater and Anderson, Rosen et al. have thought of a new way to obtain a change of measure for the $G|G|1$ tandem queue in [6]. In this section we will present both methods; then in Section 4 we will show that the two are equivalent in all cases where they are properly defined.

#### 3.1 Method 1 - Frater and Anderson

In [3], by Frater and Anderson, the change of measure proposed by Parekh and Walrand is further explored. Based on large deviations theory, Parekh and Walrand define a cost function $H$ that needs to be minimized in order to find the change of measure. For the tandem network with $d\ G|G|1$ queues, this cost function equals

$$
H = \frac{1}{\sum_i (\gamma'_i - \mu'_i) 1_{\gamma'_i > \mu'_i}} \left[ \lambda'_i h_{\lambda_1} \left( \frac{1}{\lambda'_1} \right) + \sum_{i=1}^{d} \mu'_i h_{\mu_i} \left( \frac{1}{\mu'_i} \right) \right],
$$

(3)

where $\gamma_i$ is the total arrival rate at queue $i$, $\mu_i$ is the service rate of queue $i$, $\lambda_1$ is the arrival rate at queue 1, $h_{\lambda_1}(\cdot)$ is the Cramér transform of the inter-arrival time distribution and $h_{\mu_i}(\cdot)$ is the Cramér transform of the distribution of the services times at queue $i$. The primes denote that these values should be optimized. The Cramér
transform is defined as \( h_D(y) = \sup_s [sy - \log M_D(s)] \), where \( D \) is the distribution. The minimization of this cost function will give us the change of measure.

Frater and Anderson simplify (3) to

\[
H = \frac{1}{\lambda_1 - \mu_R'} \left[ \lambda_1' h_{\lambda_1} \left( \frac{1}{\lambda_1} \right) + \sum_{i=1}^d \mu_i' h_{\mu_i} \left( \frac{1}{\mu_i} \right) \right],
\]

(4)

where \( R \) is the index of the rightmost unstable queue under the change of measure, i.e., \( R \) is the largest index \( j \) for which \( \mathbb{E}^\theta [B^{(j)}] > \mathbb{E}^\theta [A] \) under the change of measure \( \theta \).

Frater and Anderson explain how to find the minimum of (4). They note that for all queues \( j \neq R \) the optimal value of \( \mu_j' \) is \( \mu_j \). This result implies that \( H \) reduces in the following way

\[
H = \frac{1}{\lambda_1 - \mu_R'} \left[ \lambda_1' h_{\lambda_1} \left( \frac{1}{\lambda_1} \right) + \mu_R h_{\mu_R} \left( \frac{1}{\mu_R} \right) \right],
\]

(5)

since \( h_{\mu_j} \left( \frac{1}{\mu_j} \right) = 0 \); see [1]. Next, they note the two problems that remain in order to find the change of measure:

1. to find the value of \( R \) that is optimal, i.e., the value of \( R \) that minimizes \( H \),
2. given \( R \), to find the values of \( \lambda_R^* \) and \( \mu_R^* \),

where \( \lambda_R^* \) and \( \mu_R^* \) are the values for \( \lambda_1' \) and \( \mu_R' \) that minimize \( H \). The solution of the second problem is also given by Frater and Anderson, although they do not express this explicitly, but they refer to [5]. Given \( R \), the P&W change of measure for the single \( G|G|1 \) queue, see (1)-(2), can be generalized in the following way:

\[
dF_A^\theta^* (a) = \frac{e^{-\theta^* a}}{M_A(-\theta^*)} dF_A(a),
\]

(6)

\[
dF_B(R)^*(b) = \frac{e^{\theta^* b}}{M_B(R)(\theta^*)} dF_B(b).
\]

(7)

Note that the only difference compared with equation (1)-(2) is that now the service time distribution of queue \( R \) is tilted. All other queues remain untitled. So, if we have found the rightmost unstable queue in the tilted system, then we know that the change of measure is as explained above and \( \theta^* \) should be the solution of

\[
M_A(-\theta^*) M_B(R)(\theta^*) = 1.
\]

(8)

However, the difficulty is to find the rightmost unstable queue, because this depends on \( \theta^* \) itself. Frater and Anderson note that in the worst case this requires a search over all possible values of \( R \). However, they do not show how to do this except for some special cases.

Let us quickly summarize the method of Frater and Anderson. The function \( H \), that needs to be minimized, is a function of \( \lambda_1', \mu_R' \) and \( R \). If we take a fixed value for \( R \), \( \lambda_R^* \) and \( \mu_R^* \) can be calculated and a similar method as is used in [5] to find the change of measure. The question that remains is which \( R \) we need in order to minimize \( H \), hence \( H \) is a function of \( R \) and we will denote it by \( H(R) \).
If we have found the rightmost unstable queue $R$, we know that the function $H(R)$ in (5) should obtain the minimum value with respect to $R$. In order to find the rightmost unstable queue $R$, we should calculate all possible $\theta_j$ by solving

$$M_A(-\theta_j)M_{B(j)}(\theta_j) = 1, \quad \forall j,$$

and calculate $H(j)$ for all $j$. Then we define $R^*$ as the rightmost unstable queue, in the $\theta_R$-tilted system, that minimizes $H(R)$, i.e., $H(R^*) < H(j) \forall j \neq R^*$. Finally, we have to check whether under the resulting change of measure $\theta_{R^*}$, $R^*$ is indeed the rightmost unstable queue. If this is not the case it is not clear how to proceed, but we will show that this is indeed the rightmost unstable queue under the change of measure. 

### 3.2 Method 2 - Rosen et al.

In [6], Rosen et al. propose the following change of measure. In contrast to Frater and Anderson, Rosen et al. start by solving the equations in (9). Then they define the bottleneck queue in the following way.

**Definition 3.1.** Queue $\hat{R}$ is a $\theta$-bottleneck queue, or simply bottleneck queue, when $\theta_{\hat{R}} = \min_j(\theta_j)$, where $\theta_j$ can be found by solving (9).

**Remark 3.1.** This notion of bottleneck queue does not necessarily coincide with that of the $\rho$-bottleneck queue, i.e., queue $\hat{R}$ is not necessarily the queue with the smallest server utilization $\rho$. However, it may coincide; e.g. in case of an $M|M|1$ tandem queue.

When the bottleneck queue is unique (i.e., when $\theta_j < \theta_{\hat{R}}$ for all $j \neq \hat{R}$), the proposed change of measure by Rosen et al. is an exponential tilt $\hat{\theta} = \theta_{\hat{R}}$ for both the inter-arrival times and the service times of queue $\hat{R}$ in the following way

$$dF_{\hat{R}}^\theta(a) = \frac{e^{-\theta a}}{M_A(-\theta)}dF_A(a),$$

$$dF_{B(\hat{R})}^\theta(b) = \frac{e^{\hat{\theta} b}}{M_{B(\hat{R})}(\hat{\theta})}dF_B(b).$$

Note that this change of measure is similar to the change of measure in (6)-(7). The difference is that there is always a, possibly, different queue that is tilted and that the tilt is, possibly, different.

### 4 Comparison of the two methods

In this section we show that both methods of Section 3 are equivalent. Since for both methods the $\theta_j$'s are calculated in the same way for all $j$ (since (8) and (9) are equivalent) and so the change of measure is the same if the same queue is tilted, we only need to show that the rightmost unstable queue $R^*$, as described in Section 3.1, is equal to the $\theta$-bottleneck queue as described in Definition 3.1, i.e. $R^* = \hat{R}$.

We note that the moment generating functions of the inter-arrival times and service
times of the rightmost unstable queue under a change of measure \( \theta \) are

\[
M^\theta_A(t) = \frac{M_A(t - \theta)}{M_A(-\theta)},
\]

\[
M^\theta_B(i)(t) = \frac{M_B(i)(t + \theta)}{M_B(i)(\theta)},
\]

and so we find the expected values under this change of measure \( \theta \) to be

\[
\mathbb{E}^\theta[A] = \frac{M'_A(-\theta)}{M_A(-\theta)},
\]

\[
\mathbb{E}^\theta[B^{(i)}] = \frac{M'_B(i)(\theta)}{M_B(i)(\theta)}.
\]

For all other queues, the moment generating function is not affected by the change of measure. Now, we are ready to prove the following lemmas.

**Lemma 4.1.** Suppose there is one bottleneck queue, i.e., \( \theta_{\widehat{R}} < \theta_k \forall k \neq \widehat{R} \), and let \( \widehat{\theta} = \theta_{\widehat{R}} \). Then queue \( \widehat{R} \) is the rightmost unstable queue in the \( \widehat{\theta} \)-tilted system.

**Proof.** For notational convenience\(^1\) we let \( \Lambda_A(\theta) = \log M^\theta_A(\theta) \) and \( \Lambda_B(j)(\theta) = \log M^\theta_B(j)(\theta) \) \( \forall j = 1, \ldots, d \). Recall that we assumed \( \mathbb{E}[B^{(j)}] < \mathbb{E}[A] \forall j \), otherwise \( \widehat{\theta} = 0 \). In the remainder of this proof, we show that: (i) queue \( \widehat{R} \) is unstable under the change of measure if all queues \( k < \widehat{R} \) remain stable in the \( \widehat{\theta} \)-tilted system; (ii) queue \( \widehat{R} \) is unstable under the change of measure if there is at least one queue \( k < \widehat{R} \) that is unstable in the \( \widehat{\theta} \)-tilted system; and (iii) all queues \( k > \widehat{R} \) are stable under the change of measure, i.e., queue \( \widehat{R} \) is the rightmost unstable queue in the \( \widehat{\theta} \)-tilted system.

(i) Under the change of measure \( \widehat{\theta} \) we will first assume that all queues \( k < \widehat{R} \) remain stable, i.e., \( \mathbb{E}[B^{(k)}] < \mathbb{E}[A] \). Then we have for the server utilization of queue \( \widehat{R} \) under the \( \widehat{\theta} \)-tilt:

\[
\rho^\widehat{\theta}_{\widehat{R}} = \frac{-\mathbb{E}^\widehat{\theta}[B^{(\widehat{R})}]}{-\mathbb{E}^\widehat{\theta}[A]},
\]

\[
= \frac{M'_B(\widehat{R})(\widehat{\theta}) M_A(-\widehat{\theta})}{M_B(\widehat{R})(\widehat{\theta}) M'_A(-\widehat{\theta})},
\]

\[
= \frac{\mathbb{E}[B^{(\widehat{R})}e^{\widehat{\theta}B^{(\widehat{R})}}]}{\mathbb{E}[Ae^{\widehat{\theta}A}]},
\]

\[
= \frac{\mathbb{E}[B^{(\widehat{R})}e^{\widehat{\theta}(B^{(\widehat{R})}-A)}]}{\mathbb{E}[Ae^{\widehat{\theta}(B^{(\widehat{R})}-A)}]}.
\]

\(^1\)Note that this notation is slightly different to the notation used by Sadowsky who defined \( \Lambda_A(\theta) = \log M^\theta_A(-\theta) \) and \( \Lambda_B(j)(\theta) = \log M^\theta_B(j)(-\theta) \) \( \forall j = 1, \ldots, d \).
Furthermore we know that

\[
M_{B(\hat{R})-A}(0) = 1,
\]
\[
M'_{B(\hat{R})-A}(0) = \mathbb{E} \left[ B(\hat{R}) - A \right] < 0,
\]
\[
M_{B(\hat{R})-A}(\hat{\theta}) = 1,
\]

where the third equation follows from (9). Note that \( M_{B(\hat{R})-A}(\theta) \) is a convex function and decreasing at 0, where \( M_{B(\hat{R})-A}(0) = 1 \). At \( \theta = \hat{\theta} > 0 \) the function value of \( M_{B(\hat{R})-A}(\theta) \) is again 1, so by convexity, \( M'_{B(\hat{R})-A}(\hat{\theta}) > 0 \). Hence we know that \( \mathbb{E} \left[ (B(\hat{R}) - A)e^{\hat{\theta}(B(\hat{R}) - A)} \right] > 0 \) and it follows that \( \rho^\theta_R > 1 \).

(ii) Suppose there is at least one queue with index smaller than \( \hat{R} \) that is unstable under the \( \hat{\theta} \)-tilt. Among these queues, let \( k \) be the queue with the highest index. So in particular we have that \( k < \hat{R} \) and that queue \( k \) is unstable, i.e. \( \mathbb{E} \left[ B^{(k)} \right] > \mathbb{E}^\hat{\theta} \left[ A \right] \). Then the arrival rate at queue \( \hat{R} \) is equal to the service rate at queue \( k \). Note that the service rate of queue \( k \) is not tilted by \( \hat{\theta} \). So we have \( \rho^\theta_R = \frac{\mathbb{E}^\hat{\theta} \left[ B^{(\hat{R})} \right]}{\mathbb{E} \left[ B^{(k)} \right]} \).

We will show that \( \mathbb{E}^\hat{\theta} \left[ B^{(\hat{R})} \right] > \mathbb{E} \left[ B^{(k)} \right] \), and hence \( \rho^\theta_R > 1 \). For a graphical interpretation of this part of the proof, see Figure 1.

![Figure 1: A graphical interpretation of the inequalities presented in (10).](image)

Note that in general for a random variable \( D \), exponentially tilted by \( \theta \), it holds that \( \frac{d\Lambda_D(\theta)}{d\theta} = \frac{M_D(\theta)}{M_D(\theta)} = \mathbb{E}^\theta [D] \). We know that

\[
\mathbb{E} \left[ B^{(k)} \right] = \frac{d\Lambda_{B^{(k)}}(0)}{d\theta} \leq \frac{\Lambda_{B^{(k)}}(\theta_k)}{\theta_k} = \frac{-\Lambda_A(-\theta_k)}{\theta_k} \\
< \frac{\Lambda_A(-\hat{\theta})}{\hat{\theta}} = \frac{\Lambda_{B(\hat{R})}(\hat{\theta})}{\hat{\theta}} \leq \frac{d\Lambda_{B(\hat{R})}(\hat{\theta})}{d\theta} = \mathbb{E}^\hat{\theta} \left[ B(\hat{R}) \right],
\]
where the first and the final inequality follow from convexity of \( \Lambda_{B(k)}(\theta) \) and \( \Lambda_{B(\hat{R})}(\theta) \), and the second inequality follows from \( \hat{\theta} < \theta_k \) and concavity of \( -\Lambda_A(-\theta) \). Hence we have that queue \( \hat{R} \) is unstable under the \( \theta \)-tilt.

(iii) Finally, we will show that queue \( \hat{R} \) is the rightmost unstable queue. The arrival rate for queue \( \hat{R} + 1 \) is equal to the service rate of the unstable queue \( \hat{R} \). If queue \( \hat{R} + 1 \) is stable, then the arrival rate for queue \( \hat{R} + 2 \) is also the service rate of queue \( \hat{R} \), etcetera. Now suppose we look at some queue \( k > \hat{R} \). Then

\[
\hat{\rho}_k = \frac{\mathbb{E}[B(k)]}{\mathbb{E}[B(\hat{R})]} < 1,
\]

where the inequality follows immediately from equation (10).

Lemma 4.2. If \( \theta_J < \theta_j \) then \( H(J) < H(j) \forall J,j \).

Proof. Frater and Anderson minimize the cost function \( H(R) \), where

\[
H(R) = \frac{1}{\lambda_1^* - \mu_R^*} \left[ \lambda_1^* h_{\lambda_1} \left( \frac{1}{\lambda_1^*} \right) + \mu_R^* h_{\mu_R} \left( \frac{1}{\mu_R^*} \right) \right].
\]

Now we use similar reasoning as in [5] for the single \( G|G|1 \) queue. Suppose that \( \lambda_1^* \) and \( \mu_R^* \) achieve the infimum of \( H(R) \). We can derive formula (40) in [5] in a similar way, which in our case becomes

\[
\theta_R \frac{1}{\lambda_1^*} + h_{\lambda_1} \left( \frac{1}{\lambda_1^*} \right) = \theta_R \frac{1}{\mu_R^*} - h_{\mu_R} \left( \frac{1}{\mu_R^*} \right).
\]

Hence we can rewrite \( H(R) \) by using formula (11),

\[
H(R) = \frac{1}{\lambda_1^* - \mu_R^*} \left[ \lambda_1^* h_{\lambda_1} \left( \frac{1}{\lambda_1^*} \right) + \mu_R^* h_{\mu_R} \left( \frac{1}{\mu_R^*} \right) \right],
\]

\[
= -h_{\mu_R} \left( \frac{1}{\mu_R^*} \right) + \theta_R \frac{1}{\mu_R},
\]

\[
= \log M_{B(R)}(\theta_R),
\]

\[
= -\log M_A(-\theta_R).
\]

The second step follows directly from formula (41) of [5] and the final step is by definition of \( \theta_R \). As \( H(R) \) is a concave increasing function, the statement follows.

Theorem 4.3. Both methods, when properly defined, give the same result, i.e., \( R^* = \hat{R} \).

Proof. Suppose we use the method of Rosen et al. and we find the \( \theta \)-bottleneck queue \( \hat{R} \). Then we know that \( \hat{R} \) is the rightmost unstable queue in the \( \theta \)-tilted system by Lemma 4.1. We also know that \( \hat{\theta} \leq \theta_j \forall j \neq \hat{R} \) and, in view of Remark 4.1 below, by Lemma 4.2 we know that \( \hat{R} \) also minimizes \( H(R) \). Hence \( R^* = \hat{R} \).

Remark 4.1. In this report we did not address the case when \( \theta^* \) is not unique, i.e., there is no unique solution to \( \min_j(\theta_j), j = 1, \ldots, d \). In neither [3] nor [6] was this case considered.
5 Conclusions

In [3], Frater and Anderson proposed a change of measure for the $G|G|1$ tandem queue, but only specified it implicitly, in the form of a minimization over all possible “guesses” for which queue would become the rightmost unstable queue. Independently, Rosen et al. proposed another change of measure for the $G|G|1$ tandem queue in [6]. In this report, we have shown that the two methods result in the same change of measure for the $G|G|1$ tandem queue for all cases where they are properly defined.

References


