Probability in the Engineering and Informational Sciences, **34**, 2020, 131–156.
doi:10.1017/S0269964818000426

ON STATE-INDEPENDENT IMPORTANCE SAMPLING FOR THE *GI*|*GI*|1 TANDEM QUEUE[1]

ANNE BUIJSROGGE, PIETER-TJERK DE BOER AND WERNER R.W. SCHEINHARDT
Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede, The Netherlands
E-mail: a.buijsrogge@utwente.nl; p.t.deboer@utwente.nl; w.r.w.scheinhardt@utwente.nl

In this paper, we consider a d-node $GI|GI|1$ tandem queue with i.i.d. inter-arrival process and service processes that are independent of each other. Our main interest is to estimate the probability to reach a high level N in a busy cycle of the system using simulation. As crude simulation does not give a sufficient precision in reasonable time, we use importance sampling. We introduce a method to find a state-independent change of measure and we show that this is equivalent to a change of measure that was earlier, but implicitly, described by Parekh and Walrand [8]. We also show that this change of measure is the only exponential state-independent change of measure that may result in an asymptotically efficient estimator. Lastly, we provide necessary conditions for this state-independent change of measure to give an asymptotically efficient estimator.

Keywords: $GI|GI|1$ queue, tandem queue, rare event simulation, importance sampling

1. INTRODUCTION

Rare events can play an important role in many practical situations, including in logistics or telecommunications systems. For instance, the event in which some storage buffer becomes too full may lead to expensive loss of material, while an overflowing data buffer may lead to loss of important information. Even though such events may have a very low probability of occurring, their impact on the performance of the system as a whole can be profound, which explains why it may be important to obtain accurate estimates of such probabilities. Many other examples exist, but the ones we mentioned here can be modeled as overflow in a queueing system, which is the topic of this paper.

Importance sampling is one of the methods used to estimate the (small) probability of a so-called *rare event* using stochastic simulation. In importance sampling, the event of interest is made less rare by changing the underlying probability distributions. This change of the probability distributions is also called the *change of measure* or *tilting*. During the

[1] This work is partly based on earlier unpublished work, see [1].

 0269-9648/18 $25.00

simulation, one keeps track of the likelihood ratio, which is the ratio between the probabilities in the original system and the probabilities in the changed system. The results of the simulation are weighted by this ratio and hence one obtains an unbiased estimator.

In this paper, we consider importance sampling for $GI|GI|1$ tandem queues. More specifically, we are interested in estimating the probability that in a busy cycle of the queueing system the total number of customers reaches some high level N. The goal is to obtain a so-called *asymptotically efficient* estimator, so that the relative error of the estimator grows less than exponentially with N.

One of the first papers to consider importance sampling in queueing networks is by Parekh and Walrand [8]. Their interest is in the same probability as in the current paper. To estimate this probability for the single queue, they propose a simple, explicit, change of measure, and for networks of queues, they implicitly describe how to find a change of measure. Their proposed change of measure is state-independent, that is, the change of measure does not depend on the current state of the system. In the remainder of this paper, the change of measure proposed by Parekh and Walrand will be referred to as the P&W change of measure. To determine this change of measure, an equation needs to be solved. Frater and Anderson [6] partially solved the equation proposed by P&W, resulting in a simpler but still implicit description of the P&W state-independent change of measure for a class of $GI|GI|1$ tandem queues.

Sadowsky [9] shows that for the single $GI|GI|m$ queue the P&W change of measure gives an asymptotically efficient estimator under some mild conditions. However, Glasserman and Kou [7] show that for the $M|M|1$ *tandem* queue the P&W change of measure may or may not give an asymptotically efficient estimator. They provide necessary conditions and (other) sufficient conditions for asymptotic efficiency. De Boer [3] extends these results, but also shows that the P&W change of measure is the only state-independent change of measure that can possibly yield an asymptotically efficient estimator for the $M|M|1$ tandem queue.

To the best of our knowledge, no results on asymptotic efficiency for the $GI|GI|1$ tandem queue had been obtained so far. The generalization from $M|M|1$ tandem queues to $GI|GI|1$ tandem queues is important because in practice, queueing networks usually do not have a Markovian arrival and/or service process. The contribution of this paper is threefold. First, we introduce another, simpler, method to obtain a change of measure for $GI|GI|1$ tandem queues, based on knowledge of the decay rate of the probability of interest which has been determined in Buijsrogge et al. [2]. This method is not implicit, and we show that it is equivalent to the earlier method, in the sense that it results in the same change of measure as P&W. Secondly, we show that the change of measure proposed by P&W is the only exponential state-independent change of measure that may give an asymptotically efficient estimator. Lastly, we provide necessary conditions for this exponential state-independent change of measure to give an asymptotically efficient estimator.

Based on results for the $M|M|1$ tandem queue, it is clear that for the $GI|GI|1$ tandem queue the P&W change of measure does not always give an asymptotically efficient estimator. In [4,5], Dupuis et al. prove that a certain state-*dependent* change of measure is asymptotically efficient for Markovian networks. This change of measure roughly coincides with the P&W change of measure in most of the state space, but deviates from it near the edges. We expect the same for the $GI|GI|1$ case, which motivates our interest in the (state-independent) P&W change of measure: even though it fails to be asymptotically efficient in some models, it seems plausible that it will be an important ingredient for any asymptotically efficient state-dependent change of measure.

This paper is structured as follows. In Section 2, we introduce the model and the change of measure as derived from the decay rate obtained in Buijsrogge et al. [2]. In Section 3, we show that this is equivalent to the change of measure of Frater and Anderson [6] and thus

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

P&W. In Section 4, we show that this P&W change of measure is the only exponential state-independent change of measure that can give an asymptotically efficient estimator. Other necessary conditions for the state-independent change of measure to give an asymptotically efficient estimator are presented in Section 5. In Section 6, we give some numerical results, and the conclusions are presented in Section 7.

2. MODEL AND PRELIMINARIES

2.1. The model

In this paper, we consider d $GI|GI|1$ queues in tandem; in Section 5 and 6, we consider the special case $d = 2$. Let A_k be the inter-arrival time at queue 1 between customers k and $k+1$ and let $B_k^{(j)}$ be the service time of customer k at queue j. The arrival process and all service processes are assumed to be i.i.d. and are independent of each other. After service completion at queue $j < d$, the customer enters queue $j+1$, so there is no probabilistic routing. Note that the arrival process at queue $j > 1$ is obviously not independent and identically distributed. 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. Let $\mathcal{K} = \min(K_0, K_N)$. Then the indicator $\mathbb{1}\{\mathcal{K} = K_N\}$ defines if we have reached our rare event in the busy cycle or not, and the probability of this rare event, denoted by p_N, is equal to $\mathbb{E}[\mathbb{1}\{\mathcal{K} = K_N\}]$.

We denote the distribution functions of A_k and $B_k^{(j)}$ by F_A and $F_{B^{(j)}}$, respectively, and their moment generating functions by $M_A(t)$ and $M_{B^{(j)}}(t)$; for notational convenience, we let $\Lambda_A(t) = \log M_A(t)$ and $\Lambda_{B^{(j)}}(t) = \log M_{B^{(j)}}(t)$. Throughout this paper, we assume that for all $j = 1, \ldots, d$, $M_{B^{(j)}}(t)$ exists for some $t > 0$. We also assume that the system is stable, that is, $\mathbb{E}[B^{(j)}] < \mathbb{E}[A]$ $\forall j = 1, \ldots, d$. If at least one of the queues would be unstable, then our event of interest would not be rare and, therefore, no importance sampling would be needed in order to obtain a good estimation of the probability of the event. Also, we make the non-triviality assumptions that $\mathbb{P}(B^{(j)} > A) > 0$ for at least one queue j, so that the number of customers can reach any high level N in a busy cycle of the system, and that $\mathbb{P}\left(A > \sum_{j=1}^{d} B^{(j)}\right) > 0$, so that the system can become empty.

Under these assumptions, it is shown in [2] that the decay of p_N is given by

$$\lim_{N\to\infty} \frac{1}{N} \log p_N = \Lambda_A(-\theta^*), \tag{1}$$

where θ^* is the minimum of $\theta^{(1)}, \ldots, \theta^{(d)}$, with $\theta^{(j)}$ given as

$$\theta^{(j)} = \sup\{\theta : M_A(-\theta) M_{B^{(j)}}(\theta) \le 1\}, \tag{2}$$

or equivalently

$$\theta^{(j)} = \sup\{\theta : \Lambda_A(-\theta) + \Lambda_{B^{(j)}}(\theta) \le 0\},$$

for all queues j. We say $\theta^{(j)} = \infty$ when $M_A(-\theta) M_{B^{(j)}}(\theta) < 1$ for all $\theta > 0$, which only happens when $\mathbb{P}(B^{(j)} > A) = 0$, see Lemma A.1. As a consequence of the stability and non-triviality assumption, $0 < \theta^* < \infty$.

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

2.2. Importance sampling simulation

In importance sampling, the rare event is made less rare by changing the underlying probability distribution. For a single $GI|GI|1$ queue, so $d = 1$, it is suggested by Parekh and Walrand [8] to apply an exponential tilt $\theta = \theta^{(1)}$, with $\theta^{(1)}$ as in (2), for both the inter-arrival times and the service times in the following way,

$$dF_A^{\theta}(a) = \frac{e^{-\theta a}}{M_A(-\theta)} dF_A(a), \tag{3}$$

$$dF_{B^{(1)}}^{\theta}(b) = \frac{e^{\theta b}}{M_{B^{(1)}}(\theta)} dF_{B^{(1)}}(b), \tag{4}$$

where $F_A^{\theta}(a)$ and $F_{B^{(1)}}^{\theta}(b)$ denote the distribution functions under the change of measure. It is shown by Sadowsky in [9] that this change of measure results in an asymptotically efficient estimator, assuming that $\mathbb{E}[B^{(1)}] < \mathbb{E}[A]$ (stability), that $\mathbb{P}(B^{(1)} > A) > 0$ (non-triviality), and that $\mathbb{P}(B^{(1)} < M) = 1$ for some finite constant M (bounded service times). The last assumption is the only real restriction, but it was claimed that this is a mere technicality, and not essential for the result to hold.

Let us now consider d $GI|GI|1$ queues in tandem and let $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_d)$ be a vector of exponential tilts. Then $\mathbb{E}^{\boldsymbol{\theta}}[\cdot]$ and $\mathbb{P}^{\boldsymbol{\theta}}(\cdot)$ denote expected values and probabilities under this change of measure $\boldsymbol{\theta}$, and we denote the distribution function and the moment generating function of a random variable X under this change of measure as $F_X^{\boldsymbol{\theta}}(x) = \mathbb{P}^{\boldsymbol{\theta}}(X \leq x)$ and $M_X^{\boldsymbol{\theta}}(t) = \mathbb{E}^{\boldsymbol{\theta}}[e^{tX}]$, respectively. For the distribution functions of A and $B^{(j)}$ we have

$$dF_A^{\boldsymbol{\theta}}(a) = \frac{e^{-\theta_0 a}}{M_A(-\theta_0)} dF_A(a), \tag{5}$$

$$dF_{B^{(j)}}^{\boldsymbol{\theta}}(b) = \frac{e^{\theta_j b}}{M_{B^{(j)}}(\theta_j)} dF_{B^{(j)}}(b), \qquad j = 1, \ldots, d. \tag{6}$$

Note that the difference compared with Eqs. (3) and (4) is that now the inter-arrival time distribution and service time distributions of all queues j may be tilted differently. As a result, the moment generating functions of A and $B^{(j)}$ under the change of measure $\boldsymbol{\theta}$ are

$$M_A^{\boldsymbol{\theta}}(t) = \frac{M_A(t - \theta_0)}{M_A(-\theta_0)}, \tag{7}$$

$$M_{B^{(j)}}^{\boldsymbol{\theta}}(t) = \frac{M_{B^{(j)}}(t + \theta_j)}{M_{B^{(j)}}(\theta_j)}, \quad j = 1, \ldots, d,$$

and we find the expected values under the change of measure $\boldsymbol{\theta}$ to be

$$\mathbb{E}^{\boldsymbol{\theta}}[A] = \frac{M_A'(-\theta_0)}{M_A(-\theta_0)} = \frac{-d\Lambda_A}{d\theta}(-\theta_0), \tag{8}$$

$$\mathbb{E}^{\boldsymbol{\theta}}\left[B^{(j)}\right] = \frac{M_{B^{(j)}}'(\theta_j)}{M_{B^{(j)}}(\theta_j)} = \frac{d\Lambda_{B^{(j)}}}{d\theta}(\theta_j), \quad j = 1, \ldots, d. \tag{9}$$

In the sequel it will become clear that the "best" tilt $\boldsymbol{\theta^*} = (\theta_0^*, \ldots, \theta_d^*)$ is such that $\theta_0^* = \theta_{j^*}^*$ and $\theta_j^* = 0$, $j \neq 0, j^*$, where j^* is the "bottleneck queue" in some sense. Next, we discuss how to find queue j^* and the tilt-parameter $\theta_{j^*}^*$. In Section 3 it turns out that the change of measure described above is, in fact, the P&W change of measure, although this is not immediately clear from [8].

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

2.3. Specific change of measure θ^*

Based on our knowledge of the decay rate from [2], see (1), we propose a specific change of measure. We start by solving the equations in (2), then we define the notion of bottleneck queue in the following way.

DEFINITION 2.1: *Queue j^* is a θ-bottleneck queue, when $\theta^{(j^*)} = \theta^*$, where $\theta^{(j)}$ is defined in (2) for all j, and $\theta^* = \min_j \theta^{(j)}$.*

ASSUMPTION 2.1: *In addition to*

- *stability of the system, that is, $\mathbb{E}[B^{(j)}] < \mathbb{E}[A]\ \forall j = 1, \ldots, d$;*
- *non-triviality, that is, $\mathbb{P}(B^{(j)} > A) > 0$ for at least one queue j and $\mathbb{P}\big(A > \sum_{j=1}^{d} B^{(j)}\big) > 0$; and*
- *existence of $M_{B^{(j)}}(t)$ for some $t > 0$,*

(all mentioned earlier), we assume that

- *the bottleneck queue is unique, that is, $\theta^* < \theta^{(j)}$ for all $j \neq j^*$; and*
- *the inequality in definition (2) for $j = j^*$ holds with equality:*

$$M_A(-\theta^*)M_{B^{(j^*)}}(\theta^*) = 1. \tag{10}$$

Due to the uniqueness assumption, we are now ready to introduce the change of measure based on [2]: it is simply a $\boldsymbol{\theta}$-tilt as given in (5) and (6), where we choose the exponential tilt to be $\boldsymbol{\theta} = \boldsymbol{\theta^*}$ with $\boldsymbol{\theta^*} = (\theta^*, 0, \ldots, 0, \theta^*, 0, \ldots, 0)$. This means that we only tilt the inter-arrival times and the service times of the bottleneck queue j^*, with the same tilting parameter θ^*.

We will refer to this change of measure as *the $\boldsymbol{\theta^*}$-tilt*. As mentioned earlier, in Section 3 we will show that this $\boldsymbol{\theta^*}$-tilt coincides with the P&W change of measure for cases in which this is properly defined (that is, when (10) holds), and in Section 4 we will show that it is the only reasonable exponential change of measure, since taking $\boldsymbol{\theta} \neq \boldsymbol{\theta^*}$ will result in an estimator that is not asymptotically efficient.

Remark 2.1: The notion of bottleneck queue mentioned in Definition 2.1 does not necessarily coincide with that of the ρ-bottleneck queue, that is, queue j^* is not necessarily the queue with the largest server utilization ρ. However, both notions may yield the same bottleneck; for example, in case of an $M|M|1$ tandem queue, this is always the case.

3. COMPARISON WITH FRATER AND ANDERSON

In this section, we compare our method to obtain j^* and θ^* for the change of measure for the $GI|GI|1$ tandem queue to the earlier developed method by Frater and Anderson [6]. They presented one way to obtain a change of measure for the $GI|GI|1$ tandem queue in the early 90s. Their method is based on Parekh and Walrand [8] and is written in an implicit form. In Section 3.1, we present the method of Frater and Anderson; then in Section 3.2, we show that the two are equivalent in all cases where they are properly defined.

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

3.1. Method by Frater and Anderson [6]

In [6], the change of measure proposed by Parekh and Walrand is further explored. Based on large deviations theory, Parekh and Walrand defined a cost function H that needs to be minimized in order to find the change of measure. Frater and Anderson simplify this function (see (37) in [6]) to

$$H(\lambda_1', \mu_1', \ldots, \mu_d', R) = \frac{1}{\lambda_1' - \mu_R'} \left[\lambda_1' h_A \left(\frac{1}{\lambda_1'} \right) + \sum_{j=1}^{d} \mu_j' h_{B^{(j)}} \left(\frac{1}{\mu_j'} \right) \right], \tag{11}$$

where λ_1' is the arrival rate at queue 1 and μ_j' the service rate of queue j, where each rate is just the inverse of the corresponding expectation, and where the primes denote that the values should be optimized to find the change of measure. Furthermore, $h_A(\cdot)$ and $h_{B^{(j)}}(\cdot)$ denote the Cramér transforms of the inter-arrival time distribution and the service time distribution at queue j, respectively (where the Cramér transform of a random variable X is defined as $h_X(y) = \sup_s[sy - \log M_X(s)]$). Finally, R is the index of the *rightmost unstable queue* under the change of measure, that is, R is the largest index j for which $\mu_j' < \lambda_1'$ under the change of measure. (Note that Frater and Anderson write M instead of R.)

Then they explain how to find the minimum of (11). They show that for all queues $j \neq R$ the optimal value of μ_j' is μ_j and since $h_{B^{(j)}}(1/\mu_j) = 0$ (see [8]) this implies that H reduces to a function of λ_1', μ_R' and R in the following way,

$$H(\lambda_1', \mu_R', R) = \frac{1}{\lambda_1' - \mu_R'} \left[\lambda_1' h_A \left(\frac{1}{\lambda_1'} \right) + \mu_R' h_{B^{(R)}} \left(\frac{1}{\mu_R'} \right) \right], \tag{12}$$

see (43) in [6]. 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, that is, the value of R that minimizes $H(\lambda_1', \mu_R', R)$,
2. given R, to find the values of λ_1' and μ_R' that minimize $H(\lambda_1', \mu_R', R)$.

Assuming the first problem is solved, that is, given R, the solution of the second problem is not hard, using a similar method as for the single $GI|GI|1$ queue, and Frater and Anderson show how to obtain the optimal values of λ_1' and μ_R', referring to [8]. From these values, again using [8], the change of measure now follows, which prescribes exponential tilting of the distributions such that their rates become equal to the optimal rates. This change of measure turns out to be precisely as in (5) and (6) above, with the tilting vector given by $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}} \equiv (\widetilde{\theta}^{(R)}, 0, \ldots, 0, \widetilde{\theta}^{(R)}, 0, \ldots, 0)$, with $\widetilde{\theta}_j = 0$ for all $j \neq 0, R$, and $\widetilde{\theta}_0 = \widetilde{\theta}_R = \widetilde{\theta}^{(R)} > 0$, where the latter is such that it satisfies

$$M_A(-\widetilde{\theta}^{(R)}) M_{B^{(R)}}(\widetilde{\theta}^{(R)}) = 1. \tag{13}$$

As a result, the expectations of A and $B^{(R)}$ under the change of measure become $1/\lambda_1'$ and $1/\mu_R'$, as they should, so given the optimal value of R the problem is solved.

However, finding the optimal R is difficult since (the index of) the rightmost unstable queue under the change of measure depends on this change of measure itself. Only for a certain class of problems, Frater and Anderson show that R can be chosen simply as the "rho-bottleneck" (see Remark 2.1 above). For the general case, they need to calculate for each possible value of R the optimal λ_1' and μ_R', and then substitute these in $H(\lambda_1', \mu_R', R)$ to obtain a function $H(R)$ that only depends on R, after which the optimal value $\widetilde{R}$ needs

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

to be picked such that it minimizes $H(R)$. When there are multiple candidates for R they seem to suggest that R should be chosen as large as possible, but this is not entirely clear to us.

Finally, it has to be checked whether under the resulting change of measure $\widetilde{\boldsymbol{\theta}}$, the corresponding $\widetilde{R}$ is indeed the rightmost unstable queue. If this is not the case it is not clear how to proceed, but we will show in Section 3.2 that $\widetilde{R}$ is indeed the rightmost unstable queue under the change of measure.

3.2. Comparison of the two methods

In this section, we show that the method based on the decay rate (as described in Section 2.3) and the method by Frater and Anderson [6] (as described in Section 3.1) are equivalent. First of all it is clear that both methods consider the same type of exponential tilting based on (5) and (6), and that the tilting vectors $\boldsymbol{\theta}^*$ and $\widetilde{\boldsymbol{\theta}}$ have the same structure, so that only the inter-arrival times and the service times of one of the queues are tilted. Frater and Anderson find optimal values for λ_1' and μ_R', where R is the particular queue to be tilted, but as described above this optimization is equivalent to finding the corresponding $\widetilde{\theta}^{(R)}$. (In fact, their use of λ_1' and μ_j', $j = 1, \ldots, d$ in minimizing (11) and (12) can be seen as an alternative (one-to-one) parametrization to optimize the tilting parameters θ_0 and θ_j, $j = 1, \ldots, d$.) Given the optimal value of R, the value of the tilting parameter $\widetilde{\theta}^{(R)}$ is given in the same way as the $\theta^{(j)}$ in our method, compare (13) with (2), and note that (13) also shows that [6] and [8] only consider cases in which (2) holds with equality (as we assume for our j^*, see Assumption 2.1).

As a consequence, the change of measure is exactly the same for both methods *if* the same queue is tilted. Therefore, we only need to show that the bottleneck queue j^* as described in Section 2.3 minimizes the function $H(R)$, and then do the "Frater and Anderson check" to see if queue j^* is indeed the rightmost unstable queue under the change of measure, as described in Section 3.1. We show these statements in the following two lemmas.

The first lemma relates the $\boldsymbol{\theta}^*$-bottleneck queue to minimizing $H(R)$. We start by briefly motivating how to rewrite $H(R)$. As mentioned, for fixed R, Frater and Anderson choose the values for λ' and μ_R' which minimize the function $H(\lambda_1', \mu_R', R)$ in (12). The optimization is done in exactly the same manner as was done by Parekh and Walrand in [8] for the single $GI|GI|1$ queue. We will not copy the details but only mention they set the partial derivatives of $H(\lambda_1', \mu_R', R)$ with respect to λ' and μ_R' equal to zero, and combine this with properties of the Cramér transform and with the implicit assumption that (13) holds; for more details see Equations (37)–(44) in [8]. The result is simply that $H(R)$ can be written as

$$H(R) = -\Lambda_A(-\widetilde{\theta}^{(R)}).$$

LEMMA 3.1: *$H(j^*) < H(j)$ for all $j \neq j^*$.*

PROOF: As mentioned before, $\widetilde{\theta}^{(R)}$ coincides with our θ^* when $j^* = R$. Indeed, $H(j) = -\Lambda_A(-\theta^{(j)})$ is minimal for the choice $j = j^*$ since $\theta^* < \theta^{(j)}$ for all $j \neq j^*$, and $-\Lambda_A(-\theta)$ is a strictly increasing function of θ. ■

In the second lemma, we check that queue j^* is the rightmost unstable queue in the $\boldsymbol{\theta}^*$-tilted system.

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

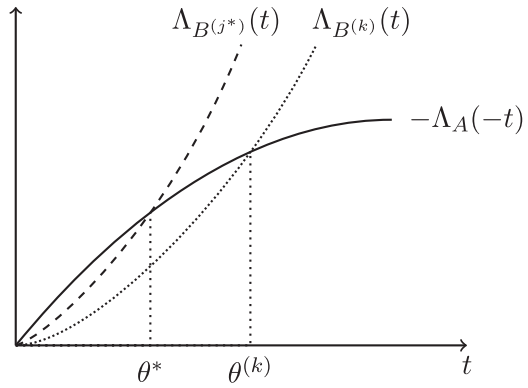

FIGURE **1.** A graphical interpretation of the inequalities presented in (14).

LEMMA 3.2: *Under Assumption 2.1, queue j^* is the rightmost unstable queue in the $\boldsymbol{\theta^*}$-tilted system and, in particular,* $\mathbb{E}^{\boldsymbol{\theta^*}}[B^{(j^*)}] > \mathbb{E}^{\boldsymbol{\theta^*}}[A]$.

PROOF: We show that: (i) queue j^* is unstable under the $\boldsymbol{\theta^*}$-tilt and $\mathbb{E}^{\boldsymbol{\theta^*}}[B^{(j^*)}] > \mathbb{E}^{\boldsymbol{\theta^*}}[A]$; and (ii) all queues $k > j^*$ are stable under the $\boldsymbol{\theta^*}$-tilt, which proves the lemma.

(i) We say that a queue is unstable when the service rate of that queue is smaller than the local arrival rate to that queue. Under the $\boldsymbol{\theta^*}$-tilt the service rate at queue j^* is $1/\mathbb{E}^{\boldsymbol{\theta^*}}[B^{(j^*)}]$, while the arrival rate at queue j^* is $\min\{1/\mathbb{E}^{\boldsymbol{\theta^*}}[A], 1/\mathbb{E}^{\boldsymbol{\theta^*}}[B^{(1)}], \ldots, 1/\mathbb{E}^{\boldsymbol{\theta^*}}[B^{(j^*-1)}]\}$. We will show that both $\mathbb{E}^{\boldsymbol{\theta^*}}[B^{(j^*)}] > \mathbb{E}^{\boldsymbol{\theta^*}}[A]$ and $\mathbb{E}^{\boldsymbol{\theta^*}}[B^{(j^*)}] > \mathbb{E}^{\boldsymbol{\theta^*}}[B^{(k)}]$ for all $k = 1, \ldots, j^* - 1$, implying instability of queue j^*. To show that $\mathbb{E}^{\boldsymbol{\theta^*}}[B^{(j^*)}] > \mathbb{E}^{\boldsymbol{\theta^*}}[A]$, we let $f(\theta) = \Lambda_A(-\theta) + \Lambda_{B^{(j^*)}}(\theta)$. We know that $f(0) = f(\theta^*) = 0$ and $f'(0) < 0$. By convexity of the log moment generating functions it must hold that $f'(\theta^*) > 0$ and so it follows, using (8) and (9), that $\mathbb{E}^{\boldsymbol{\theta^*}}[B^{(j^*)}] > \mathbb{E}^{\boldsymbol{\theta^*}}[A]$. We conclude this part of the proof by showing that $\mathbb{E}^{\boldsymbol{\theta^*}}[B^{(j^*)}] > \mathbb{E}^{\boldsymbol{\theta^*}}[B^{(k)}] = \mathbb{E}[B^{(k)}]$ for all $k = 1, \ldots, j^* - 1$. For a graphical interpretation, see Figure 1.

Again using (8) and (9), we have for any $k \neq j^*$

$$\mathbb{E}[B^{(k)}] = \frac{d\Lambda_{B^{(k)}}(0)}{d\theta} \leq \frac{\Lambda_{B^{(k)}}(\theta^*)}{\theta^*} < \frac{\Lambda_{B^{(j^*)}}(\theta^*)}{\theta^*} \leq \frac{d\Lambda_{B^{(j^*)}}(\theta^*)}{d\theta} = \mathbb{E}^{\boldsymbol{\theta^*}}[B^{(j^*)}], \tag{14}$$

where the first and the final inequality follow from the convexity of $\Lambda_{B^{(k)}}(\theta)$ and $\Lambda_{B^{(j^*)}}(\theta)$, and the second inequality follows by definition and uniqueness of θ^* (that is, if $\Lambda_{B^{(j^*)}}(\theta^*) > \Lambda_{B^{(k)}}(\theta^*)$ queue k would be the bottleneck queue instead of queue j^*, and if $\Lambda_{B^{(j^*)}}(\theta^*) = \Lambda_{B^{(k)}}(\theta^*)$ the bottleneck queue would not be unique). Hence we have that queue j^* is unstable under the $\boldsymbol{\theta^*}$-tilt and, in particular, $\mathbb{E}^{\boldsymbol{\theta^*}}[B^{(j^*)}] > \mathbb{E}^{\boldsymbol{\theta^*}}[A]$.

(ii) Finally, we show that queue j^* is the *rightmost* unstable queue under the $\boldsymbol{\theta^*}$-tilt. If $j^* = d$ this statement is trivial, so suppose for the remainder of the proof that $j^* < d$. By (i), the arrival rate for queue $j^* + 1$ is equal to the service rate of the unstable queue j^*. Queue $j^* + 1$ is stable, as we have $\mathbb{E}^{\boldsymbol{\theta^*}}[B^{(j^*+1)}] = \mathbb{E}[B^{(j^*+1)}] < \mathbb{E}^{\boldsymbol{\theta^*}}[B^{(j^*)}]$ by Eq. (14). Since queue $j^* + 1$ is stable, the arrival rate for queue $j^* + 2$ also

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

equals the service rate of queue j^*. Now look at any queue $k \in [j^*+2, d]$ (if any). If all queues between j^* and k are stable, queue k is also stable because $\mathbb{E}^{\boldsymbol{\theta}^*}[B^{(k)}] = \mathbb{E}[B^{(k)}] < \mathbb{E}^{\boldsymbol{\theta}^*}[B^{(j^*)}]$, which follows immediately from Eq. (14), and hence the arrival rate at queue k equals the service rate of queue j^*. Thus, by induction, the result follows. ■

THEOREM 3.3: *Under Assumption 2.1 and when $\widetilde{R}$ is unique, we have $j^* = \widetilde{R}$, and hence the $\boldsymbol{\theta}^*$-tilt described in Section 2.3 and the P&W method as described in Frater and Anderson give the same change of measure.*

PROOF: Consider the $\boldsymbol{\theta}^*$-tilt with bottleneck queue j^*, then j^* is the rightmost unstable queue in the $\boldsymbol{\theta}^*$-tilted system by Lemma 3.2. We also know, in view of the uniqueness of j^*, that $\theta^* < \min_{j \neq j^*} \theta^{(j)}$ and by Lemma 3.1 that j^* minimizes $H(R)$. Hence $j^* = \widetilde{R}$. The equivalence of the two corresponding changes of measure is then immediate from the fact that both are based on (5) and (6) with $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$. ■

We note that $-H(j^*) = \Lambda_A(-\theta^*)$ is the rate of decay, see (1). This implies that the large deviations approximation made in Parekh and Walrand is actually good.

4. THE θ-TILT IS NOT ASYMPTOTICALLY EFFICIENT WHEN $\theta \neq \theta^*$

Having determined that the $\boldsymbol{\theta}$-tilt is the same as the P&W change of measure by Parekh and Walrand [8], in this section we show that it is the only exponential state-independent change of measure that may give an asymptotically efficient estimator. In Section 4.1, we introduce the likelihood ratio $L^{\boldsymbol{\theta}}$ of a path that reaches level N in a busy cycle of the system and give the mathematical definition of asymptotic efficiency in terms of the second moment of this random variable. Then in Section 4.2, we show the main result of this section, Theorem 4.3.

4.1. Definitions

Suppose we use the exponential change of measure $\boldsymbol{\theta}$. Remembering that by definition we have $\mathcal{K} = \min(K_0, K_N)$, we let the likelihood ratio $L^{\boldsymbol{\theta}}$ of a path that consists of $\mathcal{K}$ arrivals be,

$$L^{\boldsymbol{\theta}} = \prod_{k=1}^{\mathcal{K}-1} \frac{dF_A}{dF_A^{\boldsymbol{\theta}}}(A_k) \prod_{j=1}^{d} \prod_{k=1}^{k_j} \frac{dF_{B^{(j)}}}{dF_{B^{(j)}}^{\boldsymbol{\theta}}}(B_k^{(j)}). \tag{15}$$

Here, k_j is the number of initiated services in queue j just before the $\mathcal{K}$-th arrival, formally defined as $k_j = \mathcal{K} - 1 - \sum_{k=1}^{j} n_k + \mathbb{1}\{n_j > 0\}$ for $j = 1, \ldots, d$, where n_j is the number of customers in queue j just before the $\mathcal{K}$-th arrival. When $\mathcal{K} = K_N$ it holds that $\sum_{k=1}^{d} n_k = N - 1$, so in that case we can also write $k_j = \mathcal{K} - N + \sum_{k=j+1}^{d} n_k + \mathbb{1}\{n_j > 0\}$.

Remark 4.1: In principle, one could reduce the estimator variance a bit further by dividing the likelihood ratio in (15) by the likelihood ratio of the remaining service times upon reaching level N (but for a clearer presentation we decided not to do this).

Under the tilt $\boldsymbol{\theta}$, $L^{\boldsymbol{\theta}}\mathbb{1}\{\mathcal{K} = K_N\}$ is an unbiased estimator for p_N, that is, $p_N = \mathbb{E}^{\boldsymbol{\theta}}[L^{\boldsymbol{\theta}}\mathbb{1}\{\mathcal{K} = K_N\}]$. The goal of importance sampling simulation is to get an asymptotically efficient estimator, which can be defined as follows (see for example [7]).

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

DEFINITION 4.1: *An unbiased estimator is asymptotically efficient if*

$$\liminf_{N\to\infty} \frac{\log \mathbb{E}^{\boldsymbol{\theta}}\left[\left(L^{\boldsymbol{\theta}}\right)^2 \mathbb{1}\{\mathcal{K}=K_N\}\right]}{\log p_N} \geq 2.$$

Note that we always have $\limsup_{N\to\infty} \log \mathbb{E}^{\boldsymbol{\theta}}[(L^{\boldsymbol{\theta}})^2 \mathbb{1}\{\mathcal{K}=K_N\}]/\log p_N \leq 2$ by Jensen's inequality. Hence, alternatively, we could replace the inequality in Definition 4.1 by an equality sign (and the liminf by a limit).

The meaning of the definition is that for an asymptotically efficient estimator, the second moment vanishes at twice the rate of the estimator itself. As a consequence, the relative error increases sub-exponentially.

Using (1), we find that the estimator is asymptotically efficient if

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\boldsymbol{\theta}}\left[\left(L^{\boldsymbol{\theta}}\right)^2 \mathbb{1}\{\mathcal{K}=K_N\}\right] \leq 2\Lambda_A(-\theta^*). \tag{16}$$

4.2. Main result

In this section, we show that using an exponential tilt other than the P&W change of measure cannot give an asymptotically efficient estimator. By the above, we need to show that an estimator based on the tilt $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$ satisfies

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\boldsymbol{\theta}}\left[\left(L^{\boldsymbol{\theta}}\right)^2 \mathbb{1}\{\mathcal{K}=K_N\}\right] > 2\Lambda_A(-\theta^*). \tag{17}$$

Before we state the theorem, we need the following lemmas. Even though the statements seem obvious, they are not entirely trivial (especially the first one when $d > 1$); we present the proofs in the appendix.

LEMMA 4.1: *Suppose we have d $GI|GI|1$ queues in tandem. Under the change of measure $\boldsymbol{\theta}^*$ for which $\mathbb{E}^{\boldsymbol{\theta}^*}[B^{(j^*)}] > \mathbb{E}^{\boldsymbol{\theta}^*}[A]$, we have for all N that $\mathbb{P}^{\boldsymbol{\theta}^*}(\mathcal{K}=K_N) \geq \mathbb{P}^{\boldsymbol{\theta}^*}(E) > 0$, where E is the event that the system never empties. Moreover, we have as $N\to\infty$ that $K_N/N \to \mathbb{E}^{\boldsymbol{\theta}^*}[B^{(j^*)}]/\mathbb{E}^{\boldsymbol{\theta}^*}[B^{(j^*)}] - \mathbb{E}^{\boldsymbol{\theta}^*}[A]$ with probability 1.*

LEMMA 4.2: *Consider a sequence $\{X_N\}$ of random variables that converges to a constant c with probability 1 as $N\to\infty$ and let E be an event with $\mathbb{P}(E) > 0$. Then $\mathbb{E}[\lim_{N\to\infty} X_N \mid E] = \mathbb{E}[\lim_{N\to\infty} X_N] = c$.*

Now we are ready to prove our theorem.

THEOREM 4.3: *Consider d $GI|GI|1$ queues in tandem. Under Assumption 2.1 the $\boldsymbol{\theta}^*$-tilt is the only exponential state-independent change of measure that can possibly give an asymptotically efficient estimator.*

PROOF: Consider an exponential tilt $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$, then the goal is to show (17) when $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$. To rewrite the second moment of the likelihood ratio in terms of the expectation under $\boldsymbol{\theta}^*$,

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

rather than $\boldsymbol{\theta}$, notice that

$$\mathbb{E}^{\boldsymbol{\theta}}\left[\left(L^{\boldsymbol{\theta}}\right)^2 \mathbb{1}\{\mathcal{K}=K_N\}\right] = \mathbb{E}^{\boldsymbol{\theta}^*}\left[L^{\boldsymbol{\theta}} L^{\boldsymbol{\theta}^*} \mathbb{1}\{\mathcal{K}=K_N\}\right].$$

Let E denote the event that the system never empties, as in Lemma 4.1. We find

$$\begin{aligned}
&\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\boldsymbol{\theta}^*}\left[L^{\boldsymbol{\theta}} L^{\boldsymbol{\theta}^*} \mathbb{1}\{\mathcal{K}=K_N\}\right] \\
&\quad \geq \liminf_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\boldsymbol{\theta}^*}\left[L^{\boldsymbol{\theta}} L^{\boldsymbol{\theta}^*} \middle| E\right] + \liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}^{\boldsymbol{\theta}^*}(E) \\
&\quad \geq \liminf_{N\to\infty} \log \mathbb{E}^{\boldsymbol{\theta}^*}\left[\left(L^{\boldsymbol{\theta}} L^{\boldsymbol{\theta}^*}\right)^{1/N} \middle| E\right] \qquad \text{(by Jensen's inequality and Lemma 4.1)} \\
&\quad \geq \log \mathbb{E}^{\boldsymbol{\theta}^*}\left[\liminf_{N\to\infty} \left(L^{\boldsymbol{\theta}} L^{\boldsymbol{\theta}^*}\right)^{1/N} \middle| E\right] \qquad \text{(by Fatou's Lemma)} \\
&\quad \geq \mathbb{E}^{\boldsymbol{\theta}^*}\left[\liminf_{N\to\infty} \frac{1}{N} \log\left(L^{\boldsymbol{\theta}} L^{\boldsymbol{\theta}^*}\right) \middle| E\right] \qquad \text{(by Jensen's inequality)} \\
&\quad \geq \mathbb{E}^{\boldsymbol{\theta}^*}\left[\liminf_{N\to\infty} \frac{1}{N} \log L^{\boldsymbol{\theta}} \middle| E\right] + \mathbb{E}^{\boldsymbol{\theta}^*}\left[\liminf_{N\to\infty} \frac{1}{N} \log L^{\boldsymbol{\theta}^*} \middle| E\right]. \qquad \textbf{(18)}
\end{aligned}$$

From (15) and then (5) and (6) it follows that

$$\frac{1}{N}\log L^{\boldsymbol{\theta}} = \Lambda_A(-\theta_0)\frac{\mathcal{K}-1}{N} + \frac{\theta_0}{N}\sum_{k=1}^{\mathcal{K}-1} A_k + \sum_{j=1}^{d}\left(\Lambda_{B^{(j)}}(\theta_j)\frac{k_j}{N} - \frac{\theta_j}{N}\sum_{k=1}^{k_j} B_k^{(j)}\right),$$

and so the first term of the right-hand side of (18) is greater than or equal to

$$\begin{aligned}
f(\boldsymbol{\theta}) = \mathbb{E}^{\boldsymbol{\theta}^*}&\left[\liminf_{N\to\infty}\left(\Lambda_A(-\theta_0)\frac{\mathcal{K}-1}{N} + \frac{\theta_0}{N}\sum_{k=1}^{\mathcal{K}-1} A_k\right)\middle| E\right] \\
&+ \sum_{j=1}^{d}\mathbb{E}^{\boldsymbol{\theta}^*}\left[\liminf_{N\to\infty}\left(\Lambda_{B^{(j)}}(\theta_j)\frac{k_j}{N} - \frac{\theta_j}{N}\sum_{k=1}^{k_j} B_k^{(j)}\right)\middle| E\right].
\end{aligned}$$

Observe that with probability 1 we have $\lim_{N\to\infty} 1/K_N \sum_{k=1}^{K_N-1} A_k = \mathbb{E}^{\boldsymbol{\theta}^*}[A]$ and $\lim_{N\to\infty} 1/k_j \sum_{k=1}^{k_j} B_k^{(j)} = \mathbb{E}^{\boldsymbol{\theta}^*}[B^{(j)}]$ for all $j = 1,\ldots,d$. Conditional on the event E, for which we have $\mathbb{P}^{\boldsymbol{\theta}^*}(E) > 0$, we can replace $\mathcal{K}$ by K_N and note that with probability 1 the liminf is a constant as $K_N - N \leq k_j \leq K_N$. Then applying Lemmas 4.1 and 4.2, we can remove the conditioning from all terms of $f(\boldsymbol{\theta})$ and change the liminf to a limit in the first term. Thus, we have

$$\begin{aligned}
f(\boldsymbol{\theta}) = &\left(\Lambda_A(-\theta_0) + \theta_0\mathbb{E}^{\boldsymbol{\theta}^*}[A]\right)\mathbb{E}^{\boldsymbol{\theta}^*}\left[\lim_{N\to\infty}\frac{K_N}{N}\right] \\
&+ \sum_{j=1}^{d}\left(\mathbb{E}^{\boldsymbol{\theta}^*}\left[\liminf_{N\to\infty}\left(\Lambda_{B^{(j)}}(\theta_j) - \theta_j\mathbb{E}^{\boldsymbol{\theta}^*}\left[B^{(j)}\right]\right)\frac{k_j}{N}\right]\right).
\end{aligned}$$

We now first show that a unique minimum of the above (and hence the tightest lower bound of (18)) is achieved at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ and conclude the proof by showing that $f(\boldsymbol{\theta}^*) =$

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

$\Lambda_A(-\theta^*)$. To find the minimum of $f(\boldsymbol{\theta})$, we note that we only have to consider $\boldsymbol{\theta}$ such that $\Lambda_{B^{(j)}}(\theta_j) - \theta_j \mathbb{E}^{\boldsymbol{\theta}^*}[B^{(j)}] \leq 0$ for all $j = 1, \ldots, d$, so that we can take this constant out of the liminf which then becomes limsup. It is not hard to see that such $\boldsymbol{\theta}$ exists (for example, by the convexity of $\Lambda_{B^{(j)}}(\theta_j)$ and by (9), we have for all $\boldsymbol{\theta}$ with $0 \leq \theta_j \leq \theta_j^*$ that $\Lambda_{B^{(j)}}(\theta_j) \leq \theta_j \mathbb{E}^{\boldsymbol{\theta}}[B^{(j)}] \leq \theta_j \mathbb{E}^{\boldsymbol{\theta}^*}[B^{(j)}]$). For all such $\boldsymbol{\theta}$ we can write

$$f(\boldsymbol{\theta}) = \left(\Lambda_A(-\theta_0) + \theta_0 \mathbb{E}^{\boldsymbol{\theta}^*}[A]\right) \mathbb{E}^{\boldsymbol{\theta}^*}\left[\lim_{N\to\infty} \frac{K_N}{N}\right] + \sum_{j=1}^{d} \left(\Lambda_{B^{(j)}}(\theta_j) - \theta_j \mathbb{E}^{\boldsymbol{\theta}^*}\left[B^{(j)}\right]\right) \mathbb{E}^{\boldsymbol{\theta}^*}\left[\limsup_{N\to\infty} \frac{k_j}{N}\right]. \tag{19}$$

We take partial derivatives of $f(\boldsymbol{\theta})$:

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \theta_0} = \left(-\mathbb{E}^{\boldsymbol{\theta}}[A] + \mathbb{E}^{\boldsymbol{\theta}^*}[A]\right) \mathbb{E}^{\boldsymbol{\theta}^*}\left[\lim_{N\to\infty} \frac{K_N}{N}\right],$$

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \theta_j} = \left(\mathbb{E}^{\boldsymbol{\theta}}\left[B^{(j)}\right] - \mathbb{E}^{\boldsymbol{\theta}^*}\left[B^{(j)}\right]\right) \mathbb{E}^{\boldsymbol{\theta}^*}\left[\limsup_{N\to\infty} \frac{k_j}{N}\right], \qquad j = 1, \ldots, d.$$

These partial derivatives are zero if and only if $\theta_j = \theta_j^*$, $j = 0, \ldots, d$, since all limsups exist and are strictly positive constants. Since the log-moment generating functions $\Lambda_A(-\theta_0)$ and $\Lambda_{B^{(j)}}(\theta_j)$, $j = 1, \ldots, d$, are strictly convex functions (unless their distributions are deterministic), the right-hand side of (19) is a strictly convex function (unless all distributions are deterministic, but this is ruled out by the non-triviality and stability assumption). Therefore, and because $\boldsymbol{\theta}^*$ is one of the values of $\boldsymbol{\theta}$ for which (19) holds, we are justified in concluding that $\boldsymbol{\theta}^*$ is indeed a global minimum. Hence $f(\boldsymbol{\theta})$ is minimal only for $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

To show that $f(\boldsymbol{\theta}^*) = \Lambda_A(-\theta^*)$, we take $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ in (19) above. With $\theta_0^* = \theta_{j^*}^* = \theta^*$, and $\theta_j^* = 0$ for all other j, only two terms remain: one for the inter-arrival time A, involving $\mathbb{E}^{\boldsymbol{\theta}^*}[\lim_{N\to\infty} K_N/N]$, which is given in Lemma 4.1, and one for the bottleneck service time $B^{(j^*)}$, involving $\mathbb{E}^{\boldsymbol{\theta}^*}[\limsup_{N\to\infty} k_{j^*}/N]$, for which we can use $k_{j^*} = \mathcal{K} - N + \sum_{j=j^*+1}^{d} n_j + \mathbb{1}\{n_{j^*} > 0\}$. This leads to

$$f(\boldsymbol{\theta}^*) = \left(\Lambda_A(-\theta^*) + \theta^* \mathbb{E}^{\boldsymbol{\theta}^*}[A]\right) \frac{\mathbb{E}^{\boldsymbol{\theta}^*}\left[B^{(j^*)}\right]}{\mathbb{E}^{\boldsymbol{\theta}^*}\left[B^{(j^*)}\right] - \mathbb{E}^{\boldsymbol{\theta}^*}[A]} + \left(\Lambda_{B^{(j^*)}}(\theta^*) - \theta^* \mathbb{E}^{\boldsymbol{\theta}^*}\left[B^{(j^*)}\right]\right) \times \left(\frac{\mathbb{E}^{\boldsymbol{\theta}^*}[A]}{\mathbb{E}^{\boldsymbol{\theta}^*}\left[B^{(j^*)}\right] - \mathbb{E}^{\boldsymbol{\theta}^*}[A]} + \mathbb{E}^{\boldsymbol{\theta}^*}\left[\limsup_{N\to\infty} \frac{\sum_{j=j^*+1}^{d} n_j}{N}\right]\right).$$

Since queues $j^* + 1, \ldots, d$ are stable queues under the $\boldsymbol{\theta}^*$-tilt and we have, by assumption, that $\Lambda_A(-\theta^*) + \Lambda_{B^{(j^*)}}(\theta^*) = 0$, we find $f(\boldsymbol{\theta}^*) = \Lambda_A(-\theta^*)$. Thus, (17) holds when $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$. ■

Remark 4.2: Note that the tilt $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ can still give an asymptotically efficient estimator, but this is not guaranteed.

Remark 4.3: In case of a single queue, Sadowsky showed that the $\boldsymbol{\theta}^*$-tilt *is* the unique change of measure that is asymptotically efficient, see [9, Theorem 3]; however, he assumes bounded support for the service time distribution, which we do not need.

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

5. NECESSARY CONDITIONS FOR ASYMPTOTIC EFFICIENCY WHEN $d = 2$

Having found that the $\boldsymbol{\theta}^*$-tilt is the only exponential state-independent change of measure that can possibly give an asymptotically efficient estimator, we show that additional conditions are needed for this change of measure to actually give an asymptotically efficient estimator. In this section, we assume that we have two queues in tandem ($d = 2$). First, we derive the conditions in Section 5.1, then we zoom in to the Markovian case and compare with earlier work in Section 5.2.

5.1. Derivation of necessary conditions

To work with (16), we first rewrite the likelihood $L^{\boldsymbol{\theta}}$ as given in (15), using (5) and (6). Taking $d = 2$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ (with $\theta_0^* = \theta_{j^*}^* = \theta^*$, for which we have $M_A(-\theta^*)M_{B^{(j^*)}}(\theta^*) = 1$, and $\theta_{3-j^*}^* = 0$), we find

$$L^{\boldsymbol{\theta}^*} = \frac{M_A(-\theta^*)^{\mathcal{K}-1-k_{j^*}}}{e^{-\theta^*\left(\sum_{k=1}^{\mathcal{K}-1} A_k - \sum_{k=1}^{k_{j^*}} B_k^{(j^*)}\right)}}. \tag{20}$$

To rewrite the denominator of (20), we note the following relation for I_j, the idle time of queue j during the busy cycle, when $\mathcal{K} = K_N$,

$$I_j = \sum_{k=1}^{\mathcal{K}-1} A_k - \sum_{k=1}^{k_j} B_k^{(j)} + \bar{B}^{(j)},$$

where $\bar{B}^{(j)}$ is the residual service time of the customer in service (if any) in queue j just before the overflow level N is reached; in the event that queue j is empty just before N is reached (which is unlikely when $j = j^*$), we set $\bar{B}^{(j)} = 0$. Combining with (16) and (20) we have asymptotic efficiency when

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\boldsymbol{\theta}^*} \left[\frac{M_A(-\theta^*)^{2(\mathcal{K}-1-k_{j^*})}}{e^{-2\theta^*\left(I_{j^*} - \bar{B}^{(j^*)}\right)}} \mathbb{1}\{\mathcal{K} = K_N\} \right] \le 2\Lambda_A(-\theta^*).$$

For the numerator, we distinguish between two cases, depending on which queue is the bottleneck. When this is queue 1 ($j^* = 1$), we have $\mathcal{K} - 1 - k_{j^*} = n_1 - \mathbb{1}\{n_1 > 0\}$, so that we have asymptotic efficiency if and only if

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\boldsymbol{\theta}^*} \left[M_A(-\theta^*)^{2(n_1 - \mathbb{1}\{n_1>0\})} e^{2\theta^*\left(I_1 - \bar{B}^{(1)}\right)} \mathbb{1}\{\mathcal{K} = K_N\} \right] \le 2\Lambda_A(-\theta^*). \tag{21}$$

When queue 2 is the bottleneck queue ($j^* = 2$), we have $\mathcal{K} - 1 - k_{j^*} = n_1 + n_2 - \mathbb{1}\{n_2 > 0\}$ $= N - 1 - \mathbb{1}\{n_2 > 0\}$, so that the condition is

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\boldsymbol{\theta}^*} \left[M_A(-\theta^*)^{-2\mathbb{1}\{n_2>0\}} e^{2\theta^*(I_2 - \bar{B}^{(2)})} \mathbb{1}\{\mathcal{K} = K_N\} \right] \le 0, \tag{22}$$

where we used that $\limsup_{N\to\infty} 1/N \log M_A(-\theta^*)^{2(N-1)}$ equals the right-hand side in (16), $2\Lambda_A(-\theta^*)$.

In the sequel we will give necessary conditions for these inequalities to hold by considering specific sample paths which are very unlike the "typical" paths that lead to overflow. The advantage of this approach, which is also used in Glasserman and Kou [7] for the Markovian case, is that the chosen unlikely paths are easy to analyze, and the process spends much

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

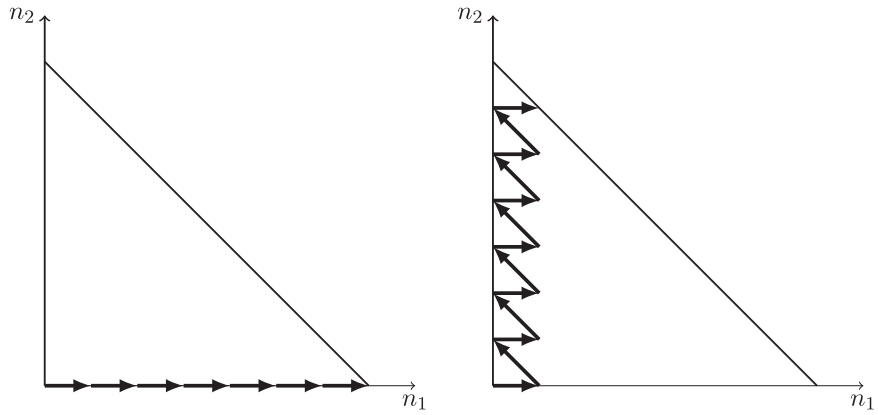

FIGURE **2.** Sample paths considered in the proofs of Theorems 5.1 (left, $j^* = 2$) and 5.2 (right, $j^* = 1$).

time on the boundaries of the state space which we know is problematic for asymptotic efficiency, at least in the Markovian case.

The specific paths we will consider are illustrated in Figure 2, and will be used in the proofs of the following theorems. After stating the theorems we will consider the Markovian case, also comparing with De Boer [3] and Glasserman and Kou [7].

We start with the necessary condition for asymptotic efficiency when queue 2 is the bottleneck queue since this is the easiest case, and show what it looks like for some special cases, including the $M|M|1$ tandem queue case.

THEOREM 5.1: *Consider* 2 $GI|GI|1$ *queues in tandem and suppose queue 2 is the bottleneck queue* ($j^* = 2$). *Under Assumption* 2.1 *a necessary condition for asymptotic efficiency of the* $\boldsymbol{\theta^*}$*-tilt is*

$$\limsup_{N\to\infty} \frac{1}{N} \log \left(\int_0^\infty e^{2\theta^* x} [1 - F_{B^{(1)}}(x)] \, dF^{\boldsymbol{\theta^*}}_{A,N-1}(x) \right) \leq 0, \tag{23}$$

where $F^{\boldsymbol{\theta^*}}_{A,N-1}(x)$ *is the* $(N-1)$*-fold convolution of* $F^{\boldsymbol{\theta^*}}_A(x)$, *the probability distribution function of* A *under tilt* $\boldsymbol{\theta^*}$.

More specifically, when the service times of the first queue are exponentially distributed with rate μ_1, *this condition becomes*

$$M_A(\theta^* - \mu_1) \leq M_A(-\theta^*), \tag{24}$$

and for an $M|M|1$ *tandem queue with arrival rate* λ *and service rates* μ_1 *and* μ_2 *(with* $\mu_1 > \mu_2$*), the condition becomes*

$$2\mu_2 \leq 2\lambda + \mu_1. \tag{25}$$

PROOF: Consider the specific sample path with no service completions before level N is reached, that is, the path that moves $N-1$ steps to the right from $(1,0)$ to $(N,0)$, see Figure 2, left panel. Based on this path we find a lower bound on the left-hand side of (22).

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

Since for this path $\mathbb{1}\{n_2 > 0\} = 0$, $\bar{B}^{(2)} = 0$ and $I_2 = \sum_{k=1}^{N-1} A_k$, it follows that

$$\begin{aligned}
&\mathbb{E}^{\boldsymbol{\theta}^*}\left[M_A(-\theta^*)^{-2\mathbb{1}\{n_2>0\}} e^{2\theta^*(I_2-\bar{B}^{(2)})}\mathbb{1}\{\mathcal{K}=K_N\}\right] \\
&\quad \geq \mathbb{E}^{\boldsymbol{\theta}^*}\left[e^{2\theta^*\sum_{k=1}^{N-1}A_k}\mathbb{1}\left\{\sum_{k=1}^{N-1}A_k < B_1^{(1)}\right\}\right] \\
&\quad = \int_0^\infty e^{2\theta^* x}[1-F_{B^{(1)}}(x)]dF^{\boldsymbol{\theta}^*}_{A,N-1}(x),
\end{aligned}$$

where the inequality follows because we only consider one possible path to reach the overflow level. Thus the general necessary condition for asymptotic efficiency in (23) follows.

When $B^{(1)} \sim \exp(\mu_1)$, the argument of the logarithm of the left-hand side in (23) reduces to

$$\begin{aligned}
\int_0^\infty e^{2\theta^* x}[1-F_{B_1^{(1)}}(x)]dF^{\boldsymbol{\theta}^*}_{A,N-1}(x) &= \int_0^\infty e^{(2\theta^*-\mu_1)x}dF^{\boldsymbol{\theta}^*}_{A,N-1}(x) \\
&= \left[M_A^{\boldsymbol{\theta}^*}(2\theta^*-\mu_1)\right]^{N-1},
\end{aligned}$$

so the necessary condition for asymptotic efficiency becomes

$$M_A^{\boldsymbol{\theta}^*}(2\theta^*-\mu_1) \leq 1,$$

which, using (7), leads to (24). Finally, (25) follows immediately from (24) by noting that $\theta^* = \mu_2 - \lambda$ for an $M|M|1$ tandem queue with $j^* = 2$. ■

Next, we provide a necessary condition for an asymptotically efficient estimator when queue 1 is the bottleneck queue. Here, we will assume that the service times of the second queue are exponentially distributed (with rate μ_2) in order to give a useful expression for the necessary conditions. We also consider some special cases, including the $M|M|1$ tandem queue.

THEOREM 5.2: *Consider 2 $GI|GI|1$ queues in tandem and suppose queue 1 is the bottleneck queue ($j^* = 1$) and that the service times of queue 2 are exponentially distributed with rate μ_2. Under Assumption 2.1 a necessary condition for asymptotic efficiency of the $\boldsymbol{\theta}^*$-tilt is*

$$\int_0^\infty e^{(2\theta^*-\mu_2)x}\int_0^x e^{-2\theta^* y}\, dF^{\boldsymbol{\theta}^*}_{B^{(1)}}(y)dF^{\boldsymbol{\theta}^*}_A(x) \leq M_A(-\theta^*)^2, \tag{26}$$

where, as before, $F^{\boldsymbol{\theta}^*}_A(x)$ and $F^{\boldsymbol{\theta}^*}_{B^{(1)}}(y)$ denote the probability distribution functions of A and $B^{(1)}$ under the tilt $\boldsymbol{\theta}^*$.
More specifically, when both queues have exponential services with rates μ_1 and μ_2 respectively, this condition becomes

$$\frac{\mu_1-\theta^*}{\theta^*+\mu_1}\left[M_A(\theta^*-\mu_2) - M_A(-(\mu_1+\mu_2))\right] \leq M_A(-\theta^*)^3, \tag{27}$$

and for an $M|M|1$ tandem queue with arrival rate λ and service rates μ_1 and μ_2 (with $\mu_1 < \mu_2$), the condition becomes

$$\frac{\mu_1}{2\mu_1-\lambda}\left[\frac{1}{2\lambda+\mu_2-\mu_1} - \frac{1}{\lambda+\mu_1+\mu_2}\right] \leq \frac{\lambda}{\mu_1^2}. \tag{28}$$

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

PROOF: We determine a lower bound on the expected value in (21) by considering the sample path that alternates between 0 and 1 customers in queue 1, with no departures from queue 2, until level N is reached; that is, the path moves from $(1,0)$ to $(0,1),(1,1),(0,2),(1,2),(0,3),\ldots,$ to $(1,N-1)$, see Figure 2, right panel. It is not hard to see that on this path we have $B_k^{(1)} < A_k, k = 1,\ldots,N-1$, and also $\sum_{k=1}^{N-1} A_k < B_1^{(1)} + B_1^{(2)}$. Obviously every $B_k^{(1)}$ should be smaller than $B_1^{(1)} + B_1^{(2)}$ as well, but this condition is implied by the above.

Also on this path we have $n_1 = 0$, $\bar{B}^{(1)} = 0$, $\mathcal{K} = K_N = N$ and $k_1 = N-1$, so

$$\begin{aligned}
&\mathbb{E}^{\boldsymbol{\theta^*}}\left[M_A(-\theta^*)^{2(n_1-\mathbb{1}\{n_1>0\})}e^{2\theta^*(I_1-\bar{B}^{(1)})}\mathbb{1}\{\mathcal{K}=K_N\}\right]\\
&\quad\geq \mathbb{E}^{\boldsymbol{\theta^*}}\left[e^{2\theta^*\left(\sum_{k=1}^{N-1}A_k-\sum_{k=1}^{N-1}B_k^{(1)}\right)}\mathbb{1}\{B_k^{(1)}<A_k,\,\forall k=1,\ldots,N-1\}\right.\\
&\qquad\left.\times\,\mathbb{1}\left\{\textstyle\sum_{k=1}^{N-1}A_k<B_1^{(1)}+B_1^{(2)}\right\}\right]\\
&\quad\geq \mathbb{E}^{\boldsymbol{\theta^*}}\left[e^{2\theta^*\left(\sum_{k=1}^{N-1}A_k-\sum_{k=1}^{N-1}B_k^{(1)}\right)}\left(\prod_{k=1}^{N-1}\mathbb{1}\{B_k^{(1)}<A_k\}\right)\mathbb{1}\left\{\textstyle\sum_{k=1}^{N-1}A_k<B_1^{(2)}\right\}\right], \quad \textbf{(29)}
\end{aligned}$$

where the first inequality follows because there are more paths that reach the overflow level than just the one we consider here. Next, since $B^{(2)}$ has the memoryless property, we can write

$$\mathbb{1}\left\{\sum_{k=1}^{N-1}A_k<B_1^{(2)}\right\}\stackrel{d}{=}\prod_{k=1}^{N-1}\mathbb{1}\left\{A_k<B_{1,k}^{(2)}\right\},$$

where the $B_{1,k}^{(2)}$ are i.i.d. copies of $B_1^{(2)}$, independent of all else and $\stackrel{d}{=}$ denotes an equality in distribution. Note that, by assuming $j^* = 1$, the service times of queue 2 remain exponentially distributed with rate μ_2 under the $\boldsymbol{\theta^*}$-tilt, and hence still have the memoryless property. As a consequence, the right-hand side of (29) can be written as

$$\begin{aligned}
&\mathbb{E}^{\boldsymbol{\theta^*}}\left[\prod_{k=1}^{N-1}e^{2\theta^*\left(A_k-B_k^{(1)}\right)}\mathbb{1}\{B_k^{(1)}<A_k\}\mathbb{1}\{A_k<B_{1,k}^{(2)}\}\right]\\
&\quad=\left(\int_0^\infty\int_0^x e^{2\theta^*(x-y)}\left[1-F_{B^{(2)}}(x)\right]\,dF_{B^{(1)}}^{\boldsymbol{\theta^*}}(y)dF_A^{\boldsymbol{\theta^*}}(x)\right)^{N-1},
\end{aligned}$$

where in the last step the independence of A_i and $B_i^{(1)}$ is used. The general necessary condition for asymptotic efficiency in (26) now follows from applying (21) and $B^{(2)} \sim \exp(\mu_2)$.

When $B^{(1)} \sim \exp(\mu_1)$ (and $B^{(2)} \sim \exp(\mu_2)$ as before), the left-hand side of (26) reduces to

$$\begin{aligned}
&(\mu_1-\theta^*)\int_0^\infty e^{(2\theta^*-\mu_2)x}\int_0^x e^{-(\theta^*+\mu_1)y}dydF_A^{\boldsymbol{\theta^*}}(x)\\
&\quad=\frac{\mu_1-\theta^*}{\theta^*+\mu_1}\left[M_A^{\boldsymbol{\theta^*}}(2\theta^*-\mu_2)-M_A^{\boldsymbol{\theta^*}}(-(\mu_1+\mu_2-\theta^*))\right].
\end{aligned}$$

from which (27) follows by using (7). Finally (28) follows from (27) by noting that $\theta^* = \mu_1 - \lambda$ for an $M|M|1$ tandem queue with $j^* = 1$. ■

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

5.2. Comparison of necessary conditions for the $M|M|1$ tandem queue

In this section, we will make a comparison with earlier papers in the Markovian case. Since these papers always consider simulation in discrete time, we will first explain how this relates to our current work.

5.2.1. Continuous-time vs. discrete-time models. In the current paper, we represent the GI/GI/1 queueing systems in continuous time, and we *simulate in continuous time*, by which we mean that we tilt the (typically continuous) distributions of the A_k and $B_k^{(j)}$. Alternatively, if all distributions are exponential, the system state can also be represented by a discrete-time Markov chain, embedded at transition epochs, which can also be simulated. We will refer to this as *simulation in discrete time*.

Parekh and Walrand [8] consider both simulation in continuous and in discrete time. For the single Markovian queue, they show that their heuristic for both continuous and discrete-time leads to the same change of measure, namely an interchange of the arrival and service rates (or probabilities). In the same way, any exponential change of measure in the discrete-time Markov chain (changing transition probabilities) can easily be shown to be equivalent to an exponential change of measure in the corresponding continuous-time Markov chain (changing transition rates).

We will now compare all known conditions for asymptotic efficiency from De Boer [3] and Glasserman and Kou [7], who apply simulation in discrete time, and the current paper. For ease of comparison, we will normalize the (continuous time) rates such that $\lambda + \mu_1 + \mu_2 = 1$, so that on the interior of the state space they coincide with the (discrete time) transition probabilities .

5.2.2. Queue 2 is bottleneck. First, we consider the case in which queue 2 is the bottleneck, which now means that $\mu_1 > \mu_2$. With the normalization, our necessary condition in Theorem 5.1 becomes $\mu_1 + 4\mu_2 \leq 2$, and since $\mu_1 > \mu_2$ it follows in particular that a necessary condition for asymptotic efficiency is $\mu_2 < 2/5$. This is stricter than $\mu_2 \leq \sqrt{2} - 1$, which was obtained in Glasserman and Kou [7], simulating in discrete time. Thus, if $\mu_2 \in [2/5, \sqrt{2} - 1]$ our estimator *cannot* be asymptotically efficient, while the estimator in [7] *could* be asymptotically efficient. Although this situation seems very unlikely, we cannot rule out the possibility since the two estimators are different.

5.2.3. Queue 1 is bottleneck. Next, we look at the case in which queue 1 is the bottleneck, which for the $M|M|1$ tandem queue means that $\mu_1 < \mu_2$. For this case, importance sampling has never been studied analytically before, because in the Markovian network both servers are interchangeable without changing the probability of overflow, see [10]. Nevertheless in [3] it is shown by numerical computations that in terms of asymptotic efficiency both servers are not interchangeable. Before we continue to summarize all known necessary conditions for the $M|M|1$ tandem queue in a figure, we present the "missing" result in Glasserman and Kou [7], namely a necessary condition for asymptotic efficiency of the estimator for simulation in discrete time, when queue 1 is the bottleneck queue. The proof is completely analogous to their approach for the other case, except that we consider the path in the right panel of Figure 2 (in discrete time), rather than the left panel.

PROPOSITION 5.3: *For an $M|M|1$ tandem queue, simulated in discrete time, with arrival rate λ and service rates μ_1 and μ_2 such that $\lambda < \mu_1 < \mu_2$ (queue 1 is bottleneck), a necessary*

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

condition for asymptotic efficiency of the corresponding estimator is

$$\mu_1^3(\mu_1+\mu_2) \leq \lambda(\lambda+\mu_2)^2(\lambda+\mu_1+\mu_2).$$

PROOF: In this case, the change of measure (here denoted as $\mathbb{Q}$) prescribes to interchange λ and μ_1. The definition for asymptotic efficiency is (cf. the continuous-time analog in Definition 4.1),

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\mathbb{Q}}\left[L^2 \mathbb{1}\{\mathcal{K} = K_N\}\right] \leq \log \frac{\lambda^2}{\mu_1^2}. \tag{30}$$

Here L is the likelihood ratio of the path as simulated in discrete time, that is, $L = \Pi_i \mathbb{P}(t_i)/\mathbb{Q}(t_i)$ where the product is taken over all transitions t_i on the path, and $\mathbb{P}(t_i)$ and $\mathbb{Q}(t_i)$ are the probabilities of t_i under the original and changed measure respectively. In order to get a lower bound on $\mathbb{E}^{\mathbb{Q}}[L^2\mathbb{1}\{\mathcal{K} = K_N\}]$, we consider the path in the right panel of Figure 2 (in discrete time) and note the following.

For each transition, t_i which is an arrival to queue 1, the contribution $\mathbb{P}(t_i)/\mathbb{Q}(t_i)$ to the likelihood ratio is

$$\frac{\frac{\lambda}{\lambda+\mu_2}}{\frac{\mu_1}{\mu_1+\mu_2}} = \frac{\lambda}{\mu_1}\frac{\mu_1+\mu_2}{\lambda+\mu_2}.$$

In order to reach the overflow level, there are $N-1$ arrivals to queue 1 (as we start with one customer in queue 1).
For each departure from queue 1, except for the first one, the contribution to the likelihood ratio is

$$\frac{\frac{\mu_1}{\lambda+\mu_1+\mu_2}}{\frac{\lambda}{\lambda+\mu_1+\mu_2}} = \frac{\mu_1}{\lambda}.$$

The contribution to the likelihood ratio of the first departure from queue 1 is

$$\frac{\frac{\mu_1}{\mu_1+\lambda}}{\frac{\lambda}{\lambda+\mu_1}} = \frac{\mu_1}{\lambda}.$$

In total, there are $N-1$ departures from queue 1. Therefore, the total likelihood ratio for this path is

$$\left(\frac{\lambda}{\mu_1}\frac{\mu_1+\mu_2}{\lambda+\mu_2}\right)^{N-1}\left(\frac{\mu_1}{\lambda}\right)^{N-1} = \left(\frac{\mu_1+\mu_2}{\lambda+\mu_2}\right)^{N-1}.$$

Similarly, the probability of this path under the change of measure $\mathbb{Q}$, is $(\mu_1/\mu_1+\mu_2)^{N-1}(\lambda/\lambda+\mu_1+\mu_2)^{N-2}\lambda/\mu_1+\lambda$. Hence we have that

$$\mathbb{E}^{\mathbb{Q}}\left[L^2\mathbb{1}\{\mathcal{K}=K_N\}\right] \geq \left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{N-1}\left(\frac{\lambda}{\lambda+\mu_1+\mu_2}\right)^{N-2}\frac{\lambda}{\mu_1+\lambda}\left(\frac{\mu_1+\mu_2}{\lambda+\mu_2}\right)^{2(N-1)},$$

so that the left-hand side of (30) is at least

$$\log\left(\frac{\mu_1}{\lambda+\mu_2}\frac{\lambda}{\lambda+\mu_1+\mu_2}\frac{\mu_1+\mu_2}{\lambda+\mu_2}\right).$$

Solving (30) concludes the proof. ■

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

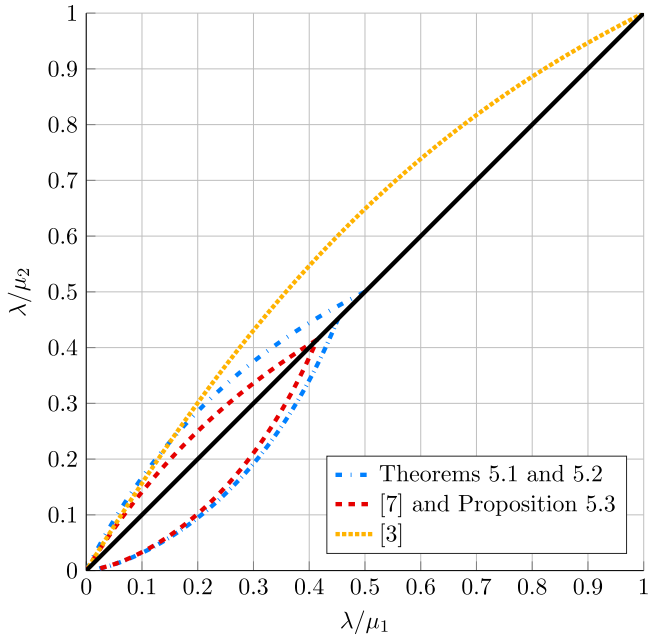

FIGURE **3.** Summary of results for the $M|M|1$ tandem queue with arrival rate λ and service rates μ_1 and μ_2 for queues 1 and 2, respectively.

5.2.4. Comparison. We are now ready to summarize all necessary conditions from [3], [7] and this section for the $M|M|1$ tandem queue (with the convention that $\lambda + \mu_1 + \mu_2 = 1$) in Figure 3. For each estimator this figure shows for which parameter settings the estimator is *certainly not* asymptotically efficient, and for which settings it *could* be:

- Between the dash-dotted (blue) lines the change of measure as discussed in the current paper (simulated in continuous time) does not give an asymptotically efficient estimator according to Theorems 5.1 and 5.2.
- Between the dashed (red) lines the change of measure as discussed in [7] (simulated in discrete time) does not give an asymptotically efficient estimator according to [7] and Proposition 5.3.
- Between the dotted (yellow) and solid (black) line the change of measure as discussed in [7] (simulated in discrete time) does not give an asymptotically efficient estimator according to [3].

When we compare the areas where asymptotic efficiency is certainly not attained, as derived from considering some unlikely path, that is, the area between the blue dash-dotted lines (simulated in continuous time), and the area between the red dashed lines (simulated in discrete time), we see that the first is largest. The method used by De Boer [3] gives an even bigger area for the discrete-time estimator, but this approach is different and only for the case where queue 2 is the bottleneck. Unfortunately this method cannot be used for simulation in continuous time.

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

6. NUMERICAL RESULTS

In this section, we give an example of the conditions that have been shown in the previous section. In order to easily show both bottleneck cases in one figure, we consider a tandem queue with exponentially distributed service times. Also we compare our results for the $M|M|1$ tandem queue with the results obtained by De Boer [3].

In Figure 4, we give an example to show that the necessary conditions for asymptotic efficiency are not always satisfied. In Tables 1 and 2, we show some simulation results for parameters as in Figures 3 and 4. In these tables RE denotes the relative error, that is, 1.96 times the standard deviation of the estimator divided by the mean of the estimator, and AE is given by

$$AE = \frac{\log \frac{1}{S} \sum_{i=1}^{S} L^2(i) I^2(i)}{\log \frac{1}{S} \sum_{i=1}^{S} L(i) I(i)},$$

where S is the total number of simulations, $L(i)$ is the likelihood ratio in simulation i, and $I(i)$ indicates whether level N has been reached in simulation i or not. This value should be 2 in case of asymptotic efficiency as N goes to infinity, see also Definition 4.1 and the text below it.

In Table 1, we present the results in case of a two node $M|M|1$ tandem queue, where the parameters are chosen such that the second queue is the bottleneck, and the necessary conditions for asymptotic efficiency given by De Boer [3] and Glasserman and Kou [7] *are* satisfied, while the conditions in Theorem 5.1 are *not* satisfied. In Table 2, we give results for a tandem queue with uniform arrivals and exponential service times at both queues, where the first queue is the bottleneck. Again, the parameters are such that the necessary conditions in Theorem 5.2 are not satisfied.

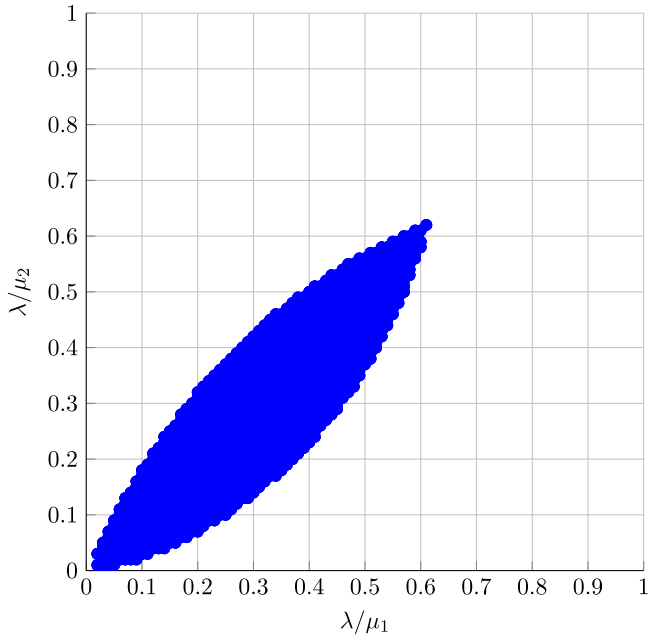

FIGURE 4. A tandem queue with $A \sim U[0, 2]$. Here $\lambda = 1/\mathbb{E}[A]$. The colored area shows for which parameter values the necessary conditions for asymptotic efficiency are *not* satisfied.

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

TABLE **1.** Simulation results for a two node $M|M|1$ tandem queue with $\lambda = 0.04$, $\mu_1 = 0.6$, and $\mu_2 = 0.36$. The number of simulations is 10^6.

N	p_n	RE	AE
100	7.61236×10^{-095}	0.0251669	1.97642
120	6.15669×10^{-114}	0.0101619	1.98723
140	5.05552×10^{-133}	0.0100353	1.98915
160	4.19910×10^{-152}	0.0165730	1.98771
180	3.49391×10^{-171}	0.0153721	1.98946
200	2.81984×10^{-190}	0.0161061	1.99031
220	2.32997×10^{-209}	0.0095608	1.99332
240	1.90921×10^{-228}	0.0153529	1.99212
260	1.58646×10^{-247}	0.0132837	1.99323
280	1.29246×10^{-266}	0.0096219	1.99474
300	1.07142×10^{-285}	0.0129584	1.99421

TABLE **2.** Simulation results when $A \sim U[0,2]$, $B^{(1)} \sim \exp(3)$ and $B^{(2)} \sim \exp(5.5)$. The number of simulations is 10^6.

N	p_n	RE	AE
100	7.51653×10^{-068}	0.0709869	1.95355
120	1.86154×10^{-081}	0.0286627	1.97111
140	4.58798×10^{-095}	0.0236072	1.97706
160	1.14667×10^{-108}	0.0272791	1.97879
180	2.92085×10^{-122}	0.0317422	1.98008
200	7.75733×10^{-136}	0.1271680	1.97317
220	1.90697×10^{-149}	0.1065780	1.97666
240	4.50666×10^{-163}	0.0548901	1.98217
260	1.08825×10^{-176}	0.0261362	1.98720
280	2.77072×10^{-190}	0.0254623	1.98824
300	6.72588×10^{-204}	0.0211566	1.98981

These tables suggest that the estimators are asymptotically efficient, as AE tends to 2 when N goes to infinity, although, in fact, they are not since they do not satisfy the conditions in Theorems 5.1 and 5.2. We can explain this in the following way. Firstly, in the proofs of Theorems 5.1 and 5.2 we considered very unlikely paths. So it is likely that these paths did not occur during these simulations and therefore it still seems that the estimator is asymptotically efficient.

Secondly, from Figures 5 and 6 we can indeed see that there are (very) unlikely paths with a large contribution to the likelihood ratio.

These figures show AE for three fixed values of N against the number of simulation runs S. We see that, even though for increasing N the value of AE seems to increase to 2 (as in Tables 1 and 2), the value of AE is clearly decreasing as the number of simulations increases. Therefore, the estimator cannot be asymptotically efficient. Moreover, the big jumps are caused by (rare) paths that have a large contribution to the likelihood ratio and they suggest that there exist paths that are even more unlikely to occur. Those paths

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

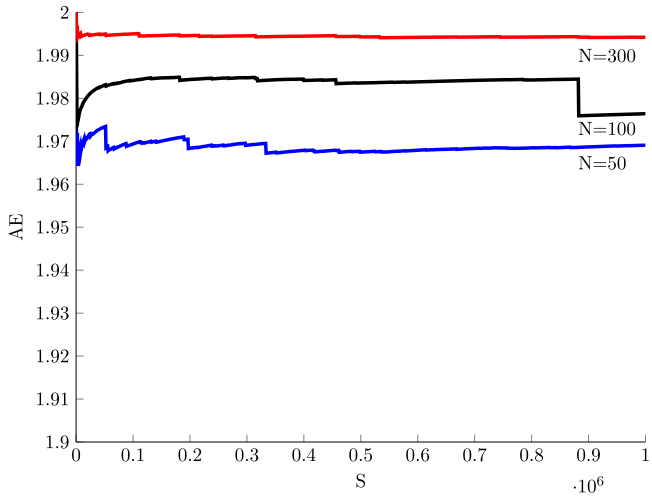

FIGURE **5.** A possible explanation of why it seems that the estimator in Table 1 is asymptotically efficient, while we proved that it is not.

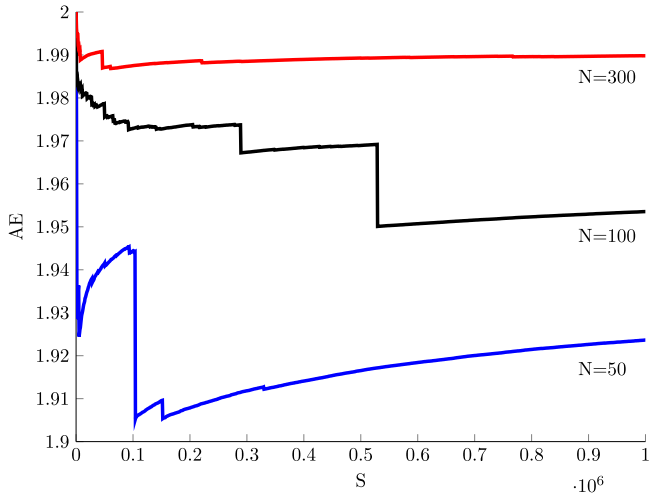

FIGURE **6.** Similar possible explanation as in Figure 5, but corresponding to the situation as in Table 2.

probably have an even larger contribution to the likelihood ratio such that the estimator is not asymptotically efficient.

Next, we compare our results for the $M|M|1$ tandem queue with the results obtained numerically by De Boer [3]. Both our and [3]'s results concern P&W changes of measure, but ours in continuous time and [3]'s in discrete time; cf. Section 5.2.1. In order to compare all these results, we use the convention that $\lambda + \mu_1 + \mu_2 = 1$ and we transform Figure 3 from [3], see Figure 7, such that λ/μ_1 and λ/μ_2 are along the x-axis and y-axis, respectively (which has been used more often throughout this paper).

What we see in Figure 7 is that when queue 1 is the bottleneck queue our necessary condition is within the blue area. When queue 2 is the bottleneck queue it seems that our necessary condition does not coincide with the numerical results of [3]. This means that there are

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

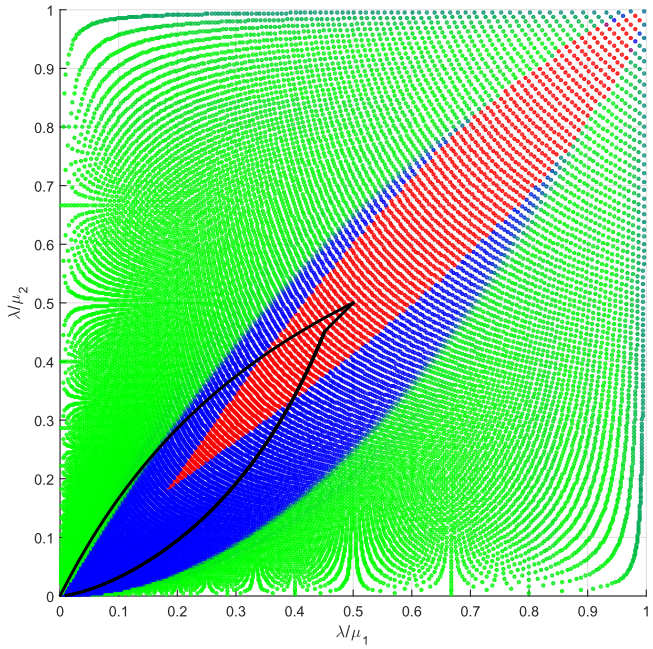

FIGURE **7.** Figure 3 of [3], displayed with different x-axis and y-axis, together with our necessary conditions from Theorems 5.1 and 5.2. The green part has bounded relative error, the blue part does not give an asymptotically efficient estimator but has finite variance and the red part has infinite variance. Between the black lines, our necessary conditions are not satisfied.

certain parameter choices where our estimator is not asymptotically efficient, while the numerical results of [3] tell that the estimator considered there has bounded relative error. This can be explained by the fact that in [3] the queueing system is simulated in discrete time, while we simulate it in continuous time.

7. CONCLUSIONS

Parekh and Walrand [8] introduced a method to estimate the probability that the total number of customers in a queueing network reaches some level N in a busy cycle using simulation, but unfortunately for a network of $GI|GI|1$ queues it is not clear how to do so. Frater and Anderson [6] found this change of measure for the $GI|GI|1$ 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. Fortunately, there is another way to *explicitly* find the change of measure for the $GI|GI|1$ tandem queue, based on the decay rate determined in Buijsrogge et al. [2]. In the present paper, we have shown that these two methods result in the same change of measure for the $GI|GI|1$ tandem queue for all cases where they are properly defined.

Also, we proved that this change of measure is the only exponential change of measure that can possibly result in an asymptotically efficient estimator. In other words, for a state-independent change of measure one should only consider using the P&W change of measure. However, using this change of measure does not guarantee asymptotic efficiency.

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

We have identified some additional necessary conditions for this change of measure to be asymptotically efficient in case of a two node tandem queue.

For future research, it seems useful to look for sufficient conditions for asymptotic efficiency, examine the tightness of the necessary conditions or maybe to improve the likelihood ratio with respect to Remark 4.1. However, it may be better to focus on state-dependent change of measures as these could be asymptotically efficient in the whole parameter space.

Acknowledgments

This work is supported by the Netherlands Organization for Scientific Research (NWO), project number 613.001.105.

References

1. Buijsrogge, A., de Boer, P.T., & Scheinhardt, W.R.W. (2015). A note on a state-independent change of measure for the $G|G|1$ tandem queue. *Memorandum 2051, Department of Applied Mathematics, University of Twente*.
2. Buijsrogge, A., de Boer, P.T., Rosen, K., & Scheinhardt, W. (2017). Large deviations for the total queue size in non-Markovian tandem queues. *Queueing Systems* 85(3): 305–312.
3. de Boer, P.T. (2006). Analysis of state-independent importance-sampling measures for the two-node tandem queue. *ACM Transactions on Modeling and Computer Simulation* 16(3): 225–250.
4. Dupuis, P. & Wang, H. (2009). Importance sampling for Jackson networks. *Queueing Systems* 62(1): 113–157.
5. Dupuis, P., Sezer, A.D., & Wang, H. (2007). Dynamic importance sampling for queueing networks. *Annals of Applied Probability* 17(4): 1306–1346.
6. Frater, M.R. & Anderson, B.D.O. (1994). Fast simulation of buffer overflows in tandem networks of $GI|GI|1$ queues. *Annals of Operations Research* 49: 207–220.
7. Glasserman, P. & Kou, S.G. (1995). Analysis of an importance sampling estimator for tandem queues. *ACM Transactions on Modeling and Computer Simulation* 5(1): 22–42.
8. Parekh, S. & Walrand, J. (1989). A quick simulation method for excessive backlogs in networks of queues. *IEEE Transactions on Automatic Control* 34(1): 54–66.
9. Sadowsky, J.S. (1991). Large deviations theory and efficient simulation of excessive backlogs in a $GI|GI|m$ queue. *IEEE Transactions on Automatic Control* 36(12): 1383–1394.
10. Weber, R.R. (1979). The interchangeability of $\cdot|M|1$ queues in series. *Journal of Applied Probability* 16(3): 690–695.

APPENDIX A

Here we present the proofs of Lemma 4.1 and Lemma 4.2, which we copy for convenience, along with (the proof of) Lemma A.1.

Lemma 4.1. *Suppose we have d $GI|GI|1$ queues in tandem. Under the change of measure $\boldsymbol{\theta}^*$ for which $\mathbb{E}^{\boldsymbol{\theta}^*}[B^{(j^*)}] > \mathbb{E}^{\boldsymbol{\theta}^*}[A]$, we have for all N that $\mathbb{P}^{\boldsymbol{\theta}^*}(\mathcal{K} = K_N) \geq \mathbb{P}^{\boldsymbol{\theta}^*}(E) > 0$, where E is the event that the system never empties. Moreover, we have as $N \to \infty$ that $K_N/N \to \mathbb{E}^{\boldsymbol{\theta}^*}[B^{(j^*)}]/(\mathbb{E}^{\boldsymbol{\theta}^*}[B^{(j^*)}] - \mathbb{E}^{\boldsymbol{\theta}^*}[A])$ with probability* 1.

Proof of Lemma 4.1: Let $N_A(t)$ denote the number of arrivals to the system up to time t, $N_D(t)$ be the number of departures from the system up to time t and $N_{B^{(j^*)}}(t)$ be the number of departures from queue j^* at time t *if* its server would work continuously. Then by renewal theory, under the change of measure $\boldsymbol{\theta}^*$, with probability 1,

$$\lim_{t\to\infty} \frac{N_A(t) - N_D(t)}{t} \geq \lim_{t\to\infty} \frac{N_A(t) - N_{B^{(j^*)}}(t)}{t} = \frac{1}{\mathbb{E}^{\boldsymbol{\theta}^*}[A]} - \frac{1}{\mathbb{E}^{\boldsymbol{\theta}^*}\left[B^{(j^*)}\right]} > 0,$$

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

because $N_D(t) \leq N_{B^{(j^*)}}(t)$, since the number of departures of the system as a whole can never exceed the number of departures at the rightmost unstable queue *if* it would work continuously. Now if we would assume that $\mathbb{P}^{\boldsymbol{\theta}^*}(E) = 0$, that is, the system always empties, this would lead to a contradiction, since in the long run, the expected number of arrivals to the system would then equal the number of departures from the system. Hence, $\mathbb{P}^{\boldsymbol{\theta}^*}(E) > 0$. Clearly, we also have $\mathbb{P}^{\boldsymbol{\theta}^*}(\mathcal{K} = K_N) \geq \mathbb{P}^{\boldsymbol{\theta}^*}(E)$, because the event E implies that $\mathcal{K} = K_N$, so the first statement follows.

For the second statement, we let X_t be the total number of customers in the system at time t. In Lemma 3.2 it has been shown that queue j^* is the rightmost unstable queue under the $\boldsymbol{\theta}^*$-tilt. Accordingly, we write $X_t = X_t^{(1,\ldots,j^*)} + X_t^{(j^*+1,\ldots,d)}$, where $X_t^{(j^*+1,\ldots,d)}$ is the total number of customers in the (stable) queues $j^*+1,\ldots,d$ at time t, and $X_t^{(1,\ldots,j^*)}$ is the total number of customers in the (not necessarily stable) queues $1,\ldots,j^*$ at time t. Hence, with probability 1 under the $\boldsymbol{\theta}^*$-tilt,

$$\begin{aligned}\lim_{t\to\infty}\frac{X_t}{t} &= \lim_{t\to\infty}\frac{X_t^{(1,\ldots,j^*)}}{t} + 0\\ &= \lim_{t\to\infty}\frac{N_A(t) - N_{D^{(j^*)}}(t)}{t},\end{aligned}$$

where $N_{D^{(j^*)}}(t)$ is the number of departures from queue j^* up to time t. Since with probability 1, from a certain time onwards queue j^* does not empty, it follows that

$$\lim_{t\to\infty}\frac{N_A(t) - N_{D^{(j^*)}}(t)}{t} = \frac{1}{\mathbb{E}^{\boldsymbol{\theta}^*}[A]} - \frac{1}{\mathbb{E}^{\boldsymbol{\theta}^*}\left[B^{(j^*)}\right]}.$$

Note that X_t is the number of customers at (continuous) time t, but we need a discrete time result, so let $\tilde{X}_k$ be the total number of customers in queue $1,\ldots,d$ right after the k^{th} arrival and let t_k be the time of the arrival of customer k, then with probability 1

$$\lim_{k\to\infty}\frac{\tilde{X}_k}{k} = \lim_{k\to\infty}\frac{t_k}{k}\frac{X_{t_k}}{t_k} = \mathbb{E}^{\boldsymbol{\theta}^*}[A]\left(\frac{1}{\mathbb{E}^{\boldsymbol{\theta}^*}[A]} - \frac{1}{\mathbb{E}^{\boldsymbol{\theta}^*}\left[B^{(j^*)}\right]}\right).$$

But then also, when $N \to \infty$,

$$\frac{N}{K_N} = \mathbb{E}^{\boldsymbol{\theta}^*}[A]\left(\frac{1}{\mathbb{E}^{\boldsymbol{\theta}^*}[A]} - \frac{1}{\mathbb{E}^{\boldsymbol{\theta}^*}\left[B^{(j^*)}\right]}\right), \qquad \text{w.p. 1,}$$

as by definition $K_N = \min\{k : X_k = N\}$. This concludes the proof. ■

Lemma 4.2. *Consider a sequence* $\{X_N\}$ *of random variables that converges to a constant* c *with probability* 1 *as* $N \to \infty$ *and let* E *be an event with* $\mathbb{P}(E) > 0$. *Then* $\mathbb{E}[\lim_{N\to\infty} X_N \mid E] = \mathbb{E}[\lim_{N\to\infty} X_N] = c$.

Proof of Lemma 4.2: The lemma follows from elementary principles. Clearly, when $\mathbb{P}(\lim_{N\to\infty} X_N = c) = 1$, also $\mathbb{E}[\lim_{N\to\infty} X_N] = c$. To show that conditioning on some event E with $\mathbb{P}(E) > 0$ does not change this assertion, let event $F = \{\lim_{N\to\infty} X_N = c\}$ and note that $\mathbb{P}(F) = 1$ implies $\mathbb{P}(F \mid E) = \mathbb{P}(F \cap E)/\mathbb{P}(E) = 1$, and hence also $\mathbb{E}[\lim_{N\to\infty} X_N \mid E] = c$. ■

LEMMA A.1: $\mathbb{P}(B^{(j)} > A) = 0 \iff \theta^{(j)} = \infty$.

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press

PROOF: Suppose that $\mathbb{P}(B^{(j)} > A) = 0$, then $M_A(-\theta)M_{B^{(j)}}(\theta) = \mathbb{E}[e^{\theta(B^{(j)}-A)}] < 1$ for all $\theta > 0$. For the reverse statement, suppose that $\mathbb{P}(B^{(j)} - A > 0) > 0$. Then we also have $\mathbb{P}(B^{(j)} - A > \epsilon) > 0$ for some $\epsilon > 0$. Hence,

$$\mathbb{E}[e^{\theta(B^{(j)}-A)}] > \int_{\epsilon}^{\infty} e^{\theta x} dF_{B^{(j)}-A}(x) > e^{\theta\epsilon} P(B^{(j)} - A > \epsilon),$$

which goes to ∞ as $\theta \to \infty$. Therefore, $\mathbb{E}[e^{\theta(B^{(j)}-A)}]$ can only be smaller than or equal to 1 if $\theta < \infty$. Hence, $\theta^{(j)} < \infty$ if $\mathbb{P}(B^{(j)} - A > 0) > 0$ and so the reverse statement holds. ■

https://doi.org/10.1017/S0269964818000426 Published online by Cambridge University Press