Communications in Statistics - Simulation and Computation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lssp20

On Efficiency of Multilevel Splitting

D. I. Miretskiy [a] , W. R. W. Scheinhardt [a] & M. R. H. Mandjes [b]

[a] Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, The Netherlands

[b] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Amsterdam, The Netherlands

Version of record first published: 10 Feb 2012.

To cite this article: D. I. Miretskiy , W. R. W. Scheinhardt & M. R. H. Mandjes (2012): On Efficiency of Multilevel Splitting, Communications in Statistics - Simulation and Computation, 41:6, 890-904

To link to this article: http://dx.doi.org/10.1080/03610918.2012.625337

Communications in Statistics—Simulation and Computation®, 41: 890–904, 2012
Copyright © Taylor & Francis Group, LLC
ISSN: 0361-0918 print/1532-4141 online
DOI: 10.1080/03610918.2012.625337

On Efficiency of Multilevel Splitting

D. I. MIRETSKIY[1], W. R. W. SCHEINHARDT[1],
AND M. R. H. MANDJES[2]

[1]Faculty of Electrical Engineering, Mathematics, and Computer Science,
University of Twente, The Netherlands
[2]Korteweg-de Vries Institute for Mathematics, University of Amsterdam,
Amsterdam, The Netherlands

This article focuses on estimating rare events using multilevel splitting schemes. The event of interest is that a Markov process enters some rare set before another ("tabu") set. It is known that in this setting a large deviations analysis is not always sufficient for constructing asymptotically efficient importance sampling schemes; additional modifications to the change of measure suggested by large deviations are needed. As an alternative, we design an asymptotically efficient multilevel splitting scheme that relies on the large deviations analysis only. This property makes it more flexible and easier to implement than corresponding importance sampling schemes.

Keywords Asymptotic efficiency; Fast simulation techniques; Queueing networks; Server slowdown; Splitting method; Variance reduction.

Mathematics Subject Classification 60K25; 60J22; 65C05.

1. Introduction

Rare event analysis has been attracting continuous and growing attention over the past decades. It has many possible applications in different areas, e.g., queueing theory, insurance, engineering, etc. As explicit expressions are hard to obtain, and asymptotic approximations often lack error bounds, one often applies simulation methods to obtain performance measures of interest.

In this article, we are interested in the rare event when a Markov process first enters some rare set before a "tabu" set. Obviously, using standard Monte Carlo simulation for estimating the probability of such a rare event has an inherent problem: it is extremely time consuming to obtain reliable estimates since the number of samples needed to obtain an estimate of a certain predefined accuracy is inversely proportional to the probability of interest. Two important techniques to speed up simulations are Importance Sampling (IS) and Multilevel Splitting (MS).

Received November 30, 2009; Accepted June 11, 2010

Address correspondence to W. R. W. Scheinhardt, Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands; E-mail: w.r.w.scheinhardt@utwente.nl

IS prescribes to simulate the system under a *new* probability measure such that the event of interest occurs more frequently, and corrects the simulation output by means of likelihood ratios to retain unbiasedness. The likelihood ratios essentially capture the likelihood of the realization under the old measure with respect to the new measure. The choice of a "good" new measure is rather delicate; in fact only measures that are *asymptotically efficient* are worthwhile to consider. Usually, the theory of large deviations is used to find a good change of measure, but in more complex situations this is often not enough. For instance, in the context of queueing networks, the measure that is suggested by large deviations may lead to problems around the boundaries of the state space, where one or more queues are empty. It has recently been shown that this can be resolved by using a state-dependent change of measure, which is not a trivial task, especially for larger networks, see Dupuis et al. (2007), Dupuis and Wang (2009), and Miretskiy et al. (2010). We refer to Heidelberger (1995) for a more general background on IS and its pitfalls.

The other technique, MS, is conceptually easier, in the sense that one can simulate under the normal probability measure. When a sample path of the process is simulated, this is viewed as the path of a "particle". When the particle approaches the target set to a certain distance, the particle splits into a number of new particles, each of which is then simulated independently of each other and of the past. This process may repeat itself several times, hence the term *multilevel* splitting. Typically, the states where particles should be split are determined by selecting a number of level sets of a so-called *importance function*. Every time a particle (sample path) crosses the next level set of the importance function, it is split. The *splitting factor* (i.e., the number of particles that replaces the original particle) may depend on the current level.

The challenge in MS is to choose an importance function that will ensure that the probability of reaching the target set is roughly the same for all states that belong to the same level. Moreover, choosing the splitting factors appropriately is also important. Sample paths will hardly ever end up in the rare set if this factor is too small, while the number of particles (and consequently the simulation effort) will grow fast if this factor is too large. For an overview of the MS method, see Shahabuddin (1995).

There are not many examples of *asymptotically efficient* MS schemes for estimating general types of rare events in the present literature. Most articles deal either with effective heuristics for particular (queueing) models, usually providing good estimates without rigorous analysis; see e.g. Villén-Altamirano and Villén-Altamirano (2006); or with restrictive models, see e.g. Glasserman et al. (1996). The recent work in Dean and Dupuis (2009) does enable one to construct an asymptotically efficient MS scheme for estimating the probability of first entrance to a rare set, when the decay rate of the probability is known for all starting states. The authors used control-theoretic techniques to derive and prove their results.

In this work, we also provide a simple and asymptotically efficient MS scheme for estimating the probability of first entrance to some rare set. The scheme can be seen as part of the class of asymptotically efficient MS schemes developed in Dean and Dupuis (2009). However, since we are only interested in easy-to-implement (but still efficient) schemes, we use a fixed, pre-specified splitting factor R, to be used for all levels. This is in contrast to the setting in Dean and Dupuis (2009) where the splitting factor may vary between levels and is usually noninteger (which is then implemented by using a randomization procedure). We accompany the scheme with a proof of its asymptotic efficiency which is relatively easy, in the sense that it only uses probabilistic arguments and some simple bounds, thereby giving insight into why the scheme works so well.

The rest of the article's structure is as follows. In Sec. 2, we describe the general setting of interest and, after a review of the MS method in more detail, we provide the MS scheme itself. The proof of asymptotic efficiency of the scheme is given in Sec. 3. Supporting numerical results for two specific models are presented in Sec. 4 and compared with results from IS on the same models; in fact, it turns out that MS can be a good alternative to IS for certain parameter settings.

2. MS Scheme

We consider a Markov process $\{Q_k\}$ on some discrete state space D^B and assume that $\{Q_k\}$ has a finite number of possible jump directions v_i that are the same for each state x in the interior of the state space; when the state space has boundaries, states on the boundary share the same jump directions, with the exception of those that point to states outside of the state space. For any value of the *rarity parameter* B we are interested in the probability that $\{Q_k\}$ hits the (rare) target set T^B before the "tabu" set A^B, starting from some state $s \notin T^B \cup A^B$.

To clarify the situation we provide a simple queueing example, in which $\{Q_k\}$ is the joint-queue length after the kth transition of the Markov chain that describes a tandem Jackson network. Then we may be interested in the event where, starting from some state, the queue of the second node reaches a level B before the entire system becomes empty. Then obviously, B is the rarity parameter (in the sense that the event becomes more rare as we choose larger values for B), and we have $T^B = \{x \in D^B : x_2 \geq B\}$ and $A^B = (0, 0)$.

It is convenient to scale the process $\{Q_k\}$ with the parameter B. The scaled process $X_k = Q_k/B$ then makes jumps of size v_i/B, and has state space D, which is the scaled version[1] of the state space D^B of $\{Q_k\}$. The target and tabu sets T^B and A^B are scaled in the same manner, their scaled versions being given by T and A.

For such (disjoint) sets A and T and some starting state $s \notin A \cup T$, we define the stopping time

$$\tau_B^s = \inf\{k > 0 : X_k \in T, X_j \notin A \ \forall j = 1, \ldots, k-1, X_0 = s\},$$

where $\tau_B^s = \infty$ if $\{X_k\}$ hits the set A before the set T. The probability of interest is now as follows:

$$p_B^s = \mathbb{P}\left(\tau_B^s < \infty\right).$$

Importantly, we will assume that this probability decays exponentially in B, with decay rate

$$\gamma(s) = -\lim_{B \to \infty} B^{-1} \log p_B^s. \tag{1}$$

In fact, we will even assume that this convergence is uniform in s.

[1] Formally the state space is a subset of D (namely a grid) for any finite B. As B grows large, the grid becomes denser, leading to D itself in the limit $B \to \infty$.

Assumption 2.1. For any $\epsilon > 0$, some $B^* > 0$ exists such that for all $s \notin A \cup T$ we have $|B^{-1} \log p_B^s + \gamma(s)| < \epsilon$ for $B > B^*$.

It is often not too difficult to obtain expressions for the function $\gamma(s)$, since it is the large-deviations rate function, and can therefore often be found by exploiting large deviations techniques; see, e.g., Dupuis and Ellis (1997). For instance, in Miretskiy et al. (2009, 2010) we were able to find $\gamma(s)$ for the models that we will study further in Sec. 4. What we believe *is* difficult to prove in general is uniform convergence (where uniformity is with respect to the starting state s of the scaled process), for which no general guidelines can be given. It is conceivable that under mild regularity conditions, a Laplace principle (and consequently Assumption 2.1) for random walks holds uniformly on compacts. We refer to Theorems 6.3.3 and 7.2.3 in Dupuis and Ellis (1997) for cases with continuous and discontinuous statistics respectively.

To apply MS one first needs to define a family of nested sets $\{L_k\}$, $k = 0, \dots, m$ such that

$$T = L_m \subset L_{m-1} \subset \cdots \subset L_1 \subset L_0 \subset D.$$

Here, L_0 is chosen such that the starting state s lies on its boundary, i.e., $s \in \ell_0 = \partial L_0$. Furthermore, the family $\{L_k\}$ should be chosen such that every state on the boundary of L_k has similar importance, i.e., the probability of reaching T before A should be approximately equal for all states x that lie in the same boundary $\ell_k = \partial L_k$, $k = 0, \dots, m$. The sets L_i are typically chosen as the level sets of some function $g(x)$, which is called the *importance function.* Given this family, we start at the initial state s (which belongs to ℓ_0) with exactly R_0 particles. We continue to simulate each of them until they either cross level ℓ_1 or hit the tabu set A. All particles that end up in A are to be terminated without any replacement. Every particle that crosses level ℓ_1 is to be replaced by R_1 independent replicas. We continue to simulate all the (new) particles until they cross the next level ℓ_2 or hit the tabu set A, and so on. At stage k we start with some number of particles in level ℓ_{k-1} and simulate them until they either cross ℓ_k or reach A. Then each particle that crossed ℓ_k is replaced by R_k independent copies, while all particles in A are terminated. We stop the procedure when the mth level (i.e., the target set T) is reached. Now we construct the estimator as follows:

$$\hat{p}_B = \frac{X}{R_0 \cdot R_1 \cdot \cdots \cdot R_{m-1}}, \tag{2}$$

where X is the number of particles that eventually reach the target set T before the tabu set A. The estimate of p_B^s is constructed by averaging a number of independent replications of $\hat{p}_B$.

We now choose the importance function to be the logarithmic decay rate in (1), i.e., we choose $g(x) = \gamma(x)$ and describe the Multilevel Splitting scheme as follows:

1. Choose some integer R to be the splitting factor for all levels.
2. Compute the number of levels $n_B := \lfloor B\gamma(s)/\log R \rfloor$.
3. Define levels $\ell_k := \left(x \in D : \gamma(s) - \gamma(x) = \frac{k}{B} \log R \right), \quad k = 0, \dots, n_B.$ (3)
4. Define $R' := \lfloor e^{B\gamma(s) - n_B \log R} \rfloor$, to be used as splitting factor at level ℓ_{n_B} only.

The idea of the scheme is as follows: different states x in the same level have the same decay rate for their corresponding probabilities p_B^x, and the different levels are defined such that the total decay rate $\gamma(s)$ is "evenly spread"; in other words, the distances between consecutive levels are equal in terms of decay rate. The corresponding probability of crossing the next level is roughly equal to $1/R$ due to the choice of n_B in step 2, so that on average only one particle out of R will cross the next level. Finally, since level n_B is in general not the boundary of the target set T (due to the rounding in step 2), and the probability to reach T from this level is larger than $1/R$, we can do with the lower splitting factor R' at level n_B.

3. Asymptotic Efficiency

Clearly, some multilevel splitting schemes perform better than others, in terms of the variance of the resulting estimator (2) obtained under some time constraints. *Asymptotic efficiency* (or *asymptotic optimality*) of a MS scheme effectively says that the *work-normalized* variance (which is the product of the variance and the expected computational effort per simulation run; see, e.g., Glynn and Whitt, 1992) of the estimator behaves roughly like the square of its first moment. In other words, the work-normalized variance is equivalent to the variance resulting from a fixed computational budget. Keeping this concept in mind, we will call an estimator asymptotically efficient if

$$\liminf_{B\to\infty} \frac{\log\left(w(B)\mathbb{E}\hat{p}_B^2\right)}{\log \mathbb{E}\hat{p}_B} \geq 2, \tag{4}$$

where $w(B)$ represents the expected computational effort per replication of $\hat{p}_B$ (i.e., per simulation run). Having (1) in mind, we can rewrite the definition of asymptotic efficiency for an unbiased estimator (4) as follows:

$$\limsup_{B\to\infty} B^{-1}\log\left(w(B)\mathbb{E}\hat{p}_B^2\right) \leq -2\gamma(s), \tag{5}$$

An obvious consequence of the asymptotic efficiency of a multilevel splitting scheme is that the computational effort is substantially smaller compared to that of the standard Monte Carlo scheme as B grows large.

Notice that when we replace $w(B)$ by 1, we obtain the "classical" definition of asymptotic efficiency (as widely used in the study of IS schemes, where we only generate one sample path per simulation run). For the specific form of $w(B)$ we can make various choices. Here, we assume that the required time effort increases linearly in the starting level. That is, we assume it takes $k+1$ time units to simulate a sample path of a particle starting from level ℓ_k, since with high probability it will reach A before ℓ_{k+1}, which takes more time when k is large; see also Glasserman et al. (1996) for the motivation of this choice. The result is that we have

$$w(B) = \mathbb{E}\left[\sum_{k=0}^{n_B-1} R\times(k+1)\times\alpha(k) + R'\times(n_B+1)\times\alpha(n_B)\right], \tag{6}$$

where the random variable $\alpha(k)$ is the number of paths that have crossed level ℓ_k, but not ℓ_{k+1}.

From now on, in order to simplify the notation, we omit the dependence on B in the notation n_B for the number of levels. Also we rewrite the estimator in (2) as follows:

$$\hat{p}_B = \frac{1}{R^n R'} \sum_{i=1}^{R^n R'} I_i. \tag{7}$$

Here, we use that we have the same splitting factor R at each level, except for the last one where we have R'. Furthermore the I_i are indicator random variables for each of the $R^n R'$ *possible* particles that may be simulated: $I_i = 1$ if the ith particle hits the target set T before the tabu set A, and $I_i = 0$ otherwise. At first sight, it may seem that the number of particles needed to obtain this estimator grows exponentially in n, and consequently in B. However, this is not the case, since we only need to simulate a few of all possible $R^n R'$ particles till the end. Suppose for instance that from the initial R particles only one crosses ℓ_1 before A is reached, then the maximum number of possible particles to be simulated further is already reduced from $R^n R'$ to $R^{n-1} R'$.

In order to prove that (5) holds for our scheme, we first analyze the second moment of the estimator, for which we have the following result.

Lemma 3.1. *Under Assumption* 2.1 *the logarithm of the second moment of the estimator in* (7) *satisfies*:

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbb{E}\hat{p}_B^2 \leq -2\gamma(s).$$

Proof. We first write

$$\frac{1}{B} \log \mathbb{E}\hat{p}_B^2 = \frac{1}{B} \log \frac{1}{R^{2n} R'^2} + \frac{1}{B} \log \mathbb{E}\left[\sum_{i=1}^{R^n R'} I_i\right]^2. \tag{8}$$

It is not difficult to see that the first term in the right-hand side of (8) has the following behavior:

$$\lim_{B\to\infty} \frac{1}{B} \log \frac{1}{R^{2n} R'^2} = -2\gamma(s), \tag{9}$$

thanks to step 2 in (3). The second term is somewhat difficult to analyze.

$$\begin{aligned}
\frac{1}{B} \log \mathbb{E}\left[\sum_{i=1}^{R^n R'} I_i\right]^2 &= \frac{1}{B} \log \mathbb{E}\left(\sum_{i=1}^{R^n R'} I_i^2 + \sum_{i=1}^{R^n R'} \sum_{j=1, j\neq i}^{R^n R'} I_i I_j\right)\\
&= \frac{1}{B} \log \left(\sum_{i=1}^{R^n R'} \mathbb{E} I_i + \sum_{i=1}^{R^n R'} \sum_{j=1, j\neq i}^{R^n R'} \mathbb{E} I_i I_j\right)\\
&= \frac{1}{B} \log \left(R^n R' p_B^s + \sum_{i=1}^{R^n R'} \sum_{j=1, j\neq i}^{R^n R'} \mathbb{E}\left[I_i | I_j = 1\right] p_B^s\right)\\
&= \frac{1}{B} \log \left(R^n R' p_B^s \left[1 + \sum_{i=2}^{R^n R'} \mathbb{E}\left[I_i | I_1 = 1\right]\right]\right)
\end{aligned}$$

$$= \frac{1}{B}\log\left(R^n R' p_B^s\right) + \frac{1}{B}\log\left(1 + \sum_{i=2}^{R^n R'} \mathbb{E}\left[I_i | I_1 = 1\right]\right). \qquad (10)$$

In view of (1) and step 2 of (3), it is clear that the first term in the last line of (10) tends to 0 as B grows to infinity. For the second term, we condition on the level where particles 1 and k had their last common ancestor. Thus, for random states $S_i \in \ell_i$, $i = 1, \ldots, n$, and $S_0 \equiv s$ we have

$$\begin{aligned}
&\frac{1}{B}\log\left(1 + \sum_{i=2}^{R^n R'} \mathbb{E}\left[I_i | I_1 = 1\right]\right) \\
&\quad = \frac{1}{B}\log\left(1 + \sum_{i=0}^{n-1}(R-1)R^{n-i-1}R'\mathbb{E}p_B^{S_i} + (R'-1)\mathbb{E}p_B^{S_n}\right) \\
&\quad \leq \frac{1}{B}\log\left(1 + \sum_{i=0}^{n} R^{n-i}R'\mathbb{E}p_B^{S_i}\right). \qquad (11)
\end{aligned}$$

Now choose for any $\epsilon > 0$ the value of B large enough, such that for any fixed realization $s_i \in \ell_i$ of S_i, we have $p_B^{s_i} < e^{-B\gamma(s_i)+\epsilon B}$; obviously, this is possible due to Assumption 2.1. Then, since $\gamma(s_i) = \gamma(s) - (i/B)\log R$ (by step 3 of (3)), and also $R' \leq e^{B\gamma(s)-n\log R}$ (by step 4 of (3)), it is easy to see that $R^{n-i}R'p_B^{s_i} < e^{\epsilon B}$. Therefore, returning to (11), we find

$$\begin{aligned}
\frac{1}{B}\log\left(1 + \sum_{i=2}^{R^n R'} \mathbb{E}\left[I_i | I_1 = 1\right]\right) &\leq \frac{1}{B}\log\left(1 + (n+1)e^{\epsilon B}\right) \leq \frac{1}{B}\log\left((2+n)e^{\epsilon B}\right) \\
&= \frac{1}{B}\log(2+n) + \epsilon,
\end{aligned}$$

which converges to zero when B grows to infinity (by step 2 of (3)). Thus, taking into account expressions (8) and (9) we obtain the result.

Now let us analyze the expected computational effort.

Lemma 3.2. *When Assumption* 2.1 *is satisfied the logarithm of the expected workload* (6) *grows subexponentially in B, i.e.,*

$$\lim_{B\to\infty} \frac{1}{B}\log w(B) = 0.$$

Proof. We use similar arguments as in the proof of Lemma 3.1: for all ϵ there exists B^* such that for any i we have that for all $B > B^*$ it holds that

$$\mathbb{E}\alpha(i) = R^i p_B^{s,\ell_i} < e^{\epsilon B},$$

where p_B^{s,ℓ_i} is the probability that a sample path hits level ℓ_i, but does not hit level ℓ_{i+1}, starting from the initial state $s \in \ell_0$. Now for B large enough we obtain the following:

$$w(B) \le R\sum_{i=0}^{n-1}(i+1)R^i p_B^{s,\ell_i} + R'(n+1)R^n p_B^{s,\ell_n} \le R(n+1)^2 e^{\epsilon B}.$$

Taking the logarithm and sending B to infinity we obtain the following:

$$\lim_{B\to\infty}\frac{1}{B}\log w(B) \le \epsilon, \quad \forall\epsilon > 0,$$

which completes the proof.

Combining the statements of Lemmas 3.1 and 3.2 now immediately leads to the main result.

Theorem 3.1. *Under Assumption* 2.1 *the Multilevel Splitting algorithm* (3) *is asymptotically efficient.*

4. Applications

In this section, we illustrate the efficiency of the MS scheme by applying it to some specific queueing networks, namely to a two-node tandem Jackson network and a system with *server slowdown*, also known as a system with *backpressure*; see Van Foreest et al. (2005), which can be seen as a generalization of the standard Jackson network. In both cases, the rare event that we consider is that the second (downstream) queue fills up to a large level B before the entire system empties, starting from an arbitrary state s. The corresponding probability is denoted by p_B^s.

4.1. *Tandem Network*

Here, we consider a standard tandem Jackson network with arrival rate λ, and service rates μ_1 and μ_2 for the first (upstream) and the second (downstream) stations, respectively.

In Miretskiy et al. (2010) we determined the decay rate $\gamma(s)$ by minimizing certain large-deviations cost functions. When $\mu_2 < \mu_1$, the outcome is that

$$\gamma(s) = \begin{cases} -(1-s_1-s_2)\log(\lambda/\mu_2), & \text{if } s_1 \le \alpha_1(1-s_2), \\ -s_1\log(\tilde{\lambda}(s)/\lambda) - (1-s_2)\log(\tilde{\mu}_2(s)/\mu_2), & \text{if } \alpha_1(1-s_2) < s_1 < \alpha_1^{-1}(1-s_2), \\ 0, & \text{if } s_1 \ge \alpha_1^{-1}(1-s_2), \end{cases} \tag{12}$$

where $\alpha_1 = (\mu_1 - \mu_2)/(\mu_1 - \lambda)$, and $\tilde{\lambda}(s)$ and $\tilde{\mu}_2(s)$ must be determined by solving

$$\begin{cases} \tilde{\lambda} = \tilde{\mu}_1 - \dfrac{s_1}{1-s_2}(\tilde{\mu}_1 - \tilde{\mu}_2) \\ \tilde{\lambda} + \tilde{\mu}_1 + \tilde{\mu}_2 = \lambda + \mu_1 + \mu_2 \\ \tilde{\lambda}\tilde{\mu}_1\tilde{\mu}_2 = \lambda\mu_1\mu_2 \\ \tilde{\lambda} \le \tilde{\mu}_1 \text{ and } \tilde{\mu}_1 > \tilde{\mu}_2 \\ \tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2 > 0 \end{cases} \tag{13}$$

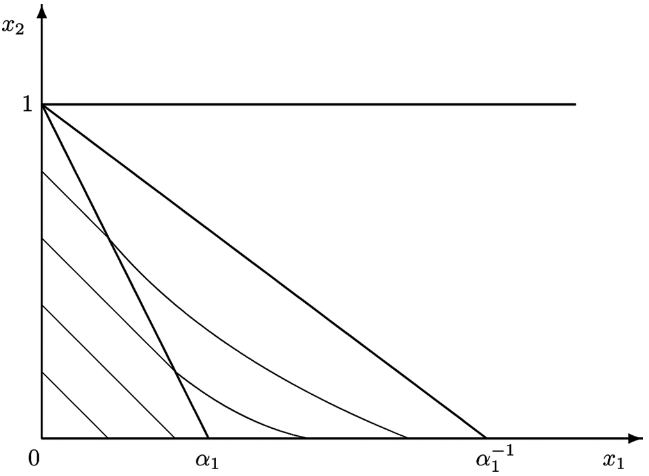

Figure 1. Splitting levels in the tandem network when $\mu_2 < \mu_1$.

In Fig. 1, we present the level curves of $\gamma(s)$, which we use as splitting levels ℓ_i in our MS simulations. Note that this figure represent the state space of the *scaled* queue-length process, i.e., x_1 and x_2 are the contents of the first and second buffers, scaled with the parameter B.

As an aside, we mention that $\tilde{\lambda}(s)$, $\tilde{\mu}_1(s)$, and $\tilde{\mu}_2(s)$ can be interpreted as the parameter values according to which the system temporarily behaves, conditional that overflow occurs, when it starts from a state s "in the middle triangle" (i.e., from a state s with $\alpha_1(1-s_2) < s_1 < \alpha_1^{-1}(1-s_2)$). The (scaled) typical path that the process follows is then simply a straight line from state s to overflow at state (0,1). Similarly, when s satisfies $s_1 \geq \alpha_1^{-1}(1-s_2)$, the typical path just runs straight from s to $(s_1 - \alpha_1^{-1}(1-s_2), 1)$, while for $s_1 \leq \alpha_1(1-s_2)$ the path first runs from s to $(0, s_2 + \alpha_1^{-1}s_1)$, and then continues along the vertical axis to (0,1).

When the first server is the bottleneck, we see a rather different picture. In this case,

$$\gamma(s) = \begin{cases} -\log(\lambda/\mu_2) + s_1\log(\lambda/\mu_1) & \text{if } f(s) \leq 0 \\ -s_1\log(\tilde{\lambda}(s)/\lambda) - (1-s_2)\log(\tilde{\mu}_2(s)/\mu_2) & \text{if } f(s) > 0 \text{ and } s_1 < \alpha_2^{-1}(1-s_2) \\ -(1-s_2)\log(\mu_1/\mu_2) & \text{if } s_1 \geq \alpha_2^{-1}(1-s_2) \end{cases}, \tag{14}$$

where $\alpha_2 = (\mu_2 - \mu_1)/(\mu_2 - \lambda)$ and $\tilde{\lambda}(s)$ and $\tilde{\mu}_2(s)$ are again found from (13). Furthermore, the function f is given by

$$f(s) = -\log(\lambda/\mu_2) + s_1\log(\tilde{\lambda}(s)/\mu_1) + (1-s_2)\log(\tilde{\mu}_2(s)/\mu_2).$$

In Fig. 2, we present the level curves of $\gamma(s)$ for the case $\mu_1 \leq \mu_2$. For states s with $f(s) > 0$, the typical path to overflow is quite straightforward (namely straight to (0,1), unless $s_1 \geq \alpha_2^{-1}(1-s_2)$, in which case it runs straight to $(s_1 - \alpha_2^{-1}(1-s_2), 1)$). On the other hand, when $f(s) < 0$, the typical path is rather remarkable, namely it first goes straight to $(s_1 + s_2(\mu_2-\lambda)/(\mu_1-\lambda), 0)$, i.e., emptying the second

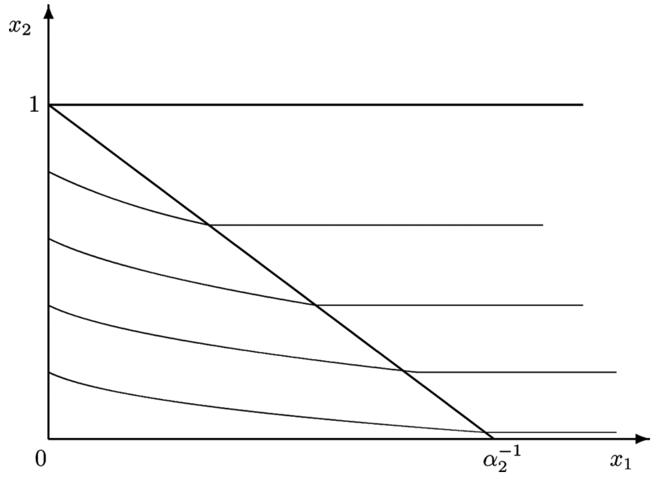

Figure 2. Splitting levels in the tandem network when $\mu_1 \leq \mu_2$.

queue, then runs along the horizontal axis to $(\alpha_2, 0)$, and then straight to (0,1). However, this behavior is not visible in Fig. 2, since the zero level curve of f (which we did not plot) lies below the lowest level curve of γ, and therefore the effect described above is not relevant for our splitting procedure (whereas the IS approach does take this into account; see Fig. 2 in Miretskiy et al., 2010).

We now discuss Assumption 2.1 in more detail. In the case when $\mu_2 < \mu_1$, there is no problem, since we use the assumption only for states s that are in one of the splitting levels ℓ_k, $k = 0, \ldots, n_B$, which are all subsets of the "triangle" $\{x \in \mathbb{R}_+^2 : x_2 \leq 1 - \alpha_1 x_1\}$; see Fig. 1. Since this is a compact set, the convergence in (1), which was shown in Miretskiy et al. (2010), immediately implies uniform convergence.

On the other hand, when $\mu_1 \leq \mu_2$ we have a different situation since the level sets are now unbounded, see Fig. 2, so in principle we need to prove uniform convergence on $\mathbb{R}_+^2$, which is difficult. In order to deal with this problem, we slightly change the probability of interest by truncating the state space to be $[0, 1) \times [0, K]$, where K is some sufficiently large constant. Keeping in mind the structure of the typical path to overflow, one may conclude that the probability almost stays the same for the original and the truncated state spaces, when K is large enough. This means that, although we do not have a formal proof of the validity of Assumption 2.1 for the original system, we can still apply the scheme to the adapted case in which compactness ensures uniform convergence.

We now provide numerical results for the tandem Jackson network. In Tables 1–3 we present estimates of p_B^s obtained by the MS scheme in (3) for different starting states s and parameter settings, accompanied by their 95% confidence intervals and relative errors. To enable comparison, we took the same parameter settings as in the asymptotically efficient IS scheme developed in Miretskiy et al. (2009). In that article, the IS simulations were performed until the relative error of the estimator reached a pre-specified value RE(IS)= 10^{-2} or RE(IS) $= 5 \cdot 10^{-2}$. In order to make a fair comparison we use the resulting computation times from those IS simulations as a time budget for our current MS scheme, resulting in the relative errors as reported in the tables; the corresponding (fixed) relative error RE(IS) is reported below the tables in the caption. (Note that the estimates of p_B^s themselves

Table 1
MS for tandem network with $(\lambda, \mu_1, \mu_2) = (0.3, 0.36, 0.34)$, $R = 4$, $\mathrm{RE(IS)} = 10^{-2}$

s	B	p_B^s	RE	Time
$(0, 0)$	20	$5.98 \cdot 10^{-2} \pm 7.27 \cdot 10^{-4}$	$0.60 \cdot 10^{-2}$	3
	50	$1.51 \cdot 10^{-3} \pm 3.69 \cdot 10^{-5}$	$1.24 \cdot 10^{-2}$	16
	100	$2.92 \cdot 10^{-6} \pm 1.06 \cdot 10^{-7}$	$1.85 \cdot 10^{-2}$	76

Table 2
MS for tandem network with $(\lambda, \mu_1, \mu_2) = (0.1, 0.55, 0.35)$, $R = 4$, $\mathrm{RE(IS)} = 10^{-2}$

s	B	p_B^s	RE	Time
$(0.6B, 0)$	20	$2.03 \cdot 10^{-5} \pm 2.59 \cdot 10^{-6}$	$6.50 \cdot 10^{-2}$	0.3
	50	$3.27 \cdot 10^{-12} \pm 5.25 \cdot 10^{-13}$	$8.19 \cdot 10^{-2}$	0.8
	100	$1.86 \cdot 10^{-23} \pm 5.48 \cdot 10^{-24}$	$15.09 \cdot 10^{-2}$	1.5

Table 3
MS for tandem network with $(\lambda, \mu_1, \mu_2) = (0.3, 0.33, 0.37)$, $R = 8$, $\mathrm{RE(IS)} = 5 \cdot 10^{-2}$

s	B	p_B^s	RE	Time
$(0, 0)$	20	$3.33 \cdot 10^{-2} \pm 3.67 \cdot 10^{-4}$	$0.56 \cdot 10^{-2}$	14
	50	$7.16 \cdot 10^{-5} \pm 2.05 \cdot 10^{-6}$	$1.46 \cdot 10^{-2}$	103
	100	$1.95 \cdot 10^{-9} \pm 8.31 \cdot 10^{-11}$	$2.17 \cdot 10^{-2}$	427

that were obtained in Miretskiy et al. (2009) are not quoted here, since they highly resemble the values we obtained here, as should be the case).

Typically, the relative errors obtained via the MS and IS methods are comparable, especially when the parameters are such that IS is hard to apply (see Tables 1 and 3). When B is not so large, MS can outperform IS in this respect.

4.2. *Slowdown Network*

We now proceed with an extension of the previous model, the slowdown network. The slowdown mechanism is designed to offer the downstream queue some sort of protection against frequent overflows and works as follows: as long as the number of jobs in the downstream queue is smaller than some pre-specified threshold, the server of the upstream queue works in the “normal” regime at rate μ_1, but when the number of jobs in the second queue is above the threshold, the first server works in a “slow” regime, at rate $\mu_1^+ < \mu_1$. This property is of significant practical interest, as a related mechanism has been proposed, e.g., in the design of Metro Ethernet (Malhotra et al., 2009; Noureddine and Tobagi, 1999). Again, we are interested in p_B^s, where the slowdown threshold scales with B; in all simulations we choose the threshold to be $0.8B$.

Table 4
MS for slowdown network with $(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.1, 0.7, 0.15, 0.2)$, $R = 10$, $\text{RE(IS)} = 10^{-2}$

s	B	p_B^s	RE	Time
(0, 0)	20	$3.73 \cdot 10^{-7} \pm 4.91 \cdot 10^{-8}$	$6.58 \cdot 10^{-2}$	1
	50	$1.45 \cdot 10^{-16} \pm 5.71 \cdot 10^{-17}$	$19.6 \cdot 10^{-2}$	2
	100	$4.99 \cdot 10^{-32} \pm 2.58 \cdot 10^{-32}$	$25.8 \cdot 10^{-2}$	6
$(0.7B, 0)$	20	$6.09 \cdot 10^{-3} \pm 2.03 \cdot 10^{-4}$	$1.70 \cdot 10^{-2}$	1
	50	$3.40 \cdot 10^{-6} \pm 3.97 \cdot 10^{-7}$	$5.95 \cdot 10^{-2}$	10
	100	$2.89 \cdot 10^{-11} \pm 9.72 \cdot 10^{-12}$	$17.1 \cdot 10^{-2}$	37
$(1.5B, 0)$	20	$5.25 \cdot 10^{-1} \pm 4.37 \cdot 10^{-3}$	$0.42 \cdot 10^{-2}$	1
	50	$1.35 \cdot 10^{-1} \pm 2.37 \cdot 10^{-3}$	$0.89 \cdot 10^{-2}$	2
	100	$1.05 \cdot 10^{-2} \pm 2.31 \cdot 10^{-4}$	$1.12 \cdot 10^{-2}$	21

The decay rate function $\gamma(s)$, the level curves of which we use as splitting levels, was described in Miretskiy et al. (2009). The structure of $\gamma(s)$ again depends on which buffer is the bottleneck, which in this case leads to three different possibilities: $\mu_2 < \mu_1^+ < \mu_1$, $\mu_1^+ \leq \mu_2 < \mu_1$ and $\mu_1^+ < \mu_1 \leq \mu_2$. The function has a similar form as in (12) and (14), only there is now a different prescript for arguments s with $s_2 < \theta$ and for arguments with $s_2 \geq \theta$. The latter now involves quantities $\tilde{\lambda}^+$, $\tilde{\mu}_1^+$, and $\tilde{\mu}_2^+$, which satisfy a system that is rather similar to (13). We do not provide further details on the shape of $\gamma(s)$ for each of the three cases here, as this would not add any substantial insight; the interested reader can consult Miretskiy et al. (2009) for the relevant expressions and derivations. Finally, Assumption 2.1 in the context of the slowdown network can be treated in the same way as for the tandem Jackson network.

We present some numerical studies of the slowdown system in Tables 4–7. Again we present estimates of p_B^s, accompanied by their 95% confidence intervals and relative errors. As was the case for the tandem Jackson network we compare the outcomes of the MS scheme with the results of the (also asymptotically efficient) IS scheme developed in Miretskiy et al. (2008). Again, we use the computation time of the IS simulations as time budget for MS. Note that the relative error obtained

Table 5
MS for slowdown network with $(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.3, 0.36, 0.32, 0.34)$, $R = 4$, $\text{RE(IS)} = 10^{-2}$

s	B	p_B^s	RE	Time
(0, 0)	20	$5.68 \cdot 10^{-2} \pm 9.07 \cdot 10^{-4}$	$0.81 \cdot 10^{-2}$	2
	50	$1.25 \cdot 10^{-3} \pm 4.45 \cdot 10^{-5}$	$1.40 \cdot 10^{-2}$	18
	100	$1.81 \cdot 10^{-6} \pm 1.91 \cdot 10^{-7}$	$5.27 \cdot 10^{-2}$	49
$(0.35B, 0)$	20	$2.02 \cdot 10^{-1} \pm 1.42 \cdot 10^{-3}$	$0.35 \cdot 10^{-2}$	2
	50	$1.36 \cdot 10^{-2} \pm 3.71 \cdot 10^{-4}$	$1.39 \cdot 10^{-2}$	9
	100	$1.29 \cdot 10^{-4} \pm 7.01 \cdot 10^{-6}$	$2.77 \cdot 10^{-2}$	25

Table 6
MS for slowdown network with $(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.3, 0.36, 0.35, 0.34)$, $R = 4$, $\mathrm{RE(IS)} = 10^{-2}$

s	B	p_B^s	RE	Time
$(0, 0)$	20	$5.91 \cdot 10^{-2} \pm 9.35 \cdot 10^{-4}$	$0.80 \cdot 10^{-2}$	2
	50	$1.46 \cdot 10^{-3} \pm 3.88 \cdot 10^{-5}$	$1.33 \cdot 10^{-2}$	21
	100	$2.71 \cdot 10^{-6} \pm 7.17 \cdot 10^{-8}$	$1.42 \cdot 10^{-2}$	121
$(0.35B, 0)$	20	$2.10 \cdot 10^{-1} \pm 2.52 \cdot 10^{-3}$	$0.54 \cdot 10^{-2}$	2
	50	$1.54 \cdot 10^{-2} \pm 4.05 \cdot 10^{-4}$	$1.34 \cdot 10^{-2}$	11
	100	$2.21 \cdot 10^{-4} \pm 8.89 \cdot 10^{-6}$	$2.05 \cdot 10^{-2}$	35

Table 7
MS for slowdown network with $(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.25, 0.35, 0.28, 0.4)$, $R = 10$, $\mathrm{RE(IS)} = 5 \cdot 10^{-2}$

s	B	p_B^s	RE	Time
$(0, 0)$	20	$1.14 \cdot 10^{-4} \pm 6.20 \cdot 10^{-6}$	$2.71 \cdot 10^{-2}$	2
	50	$4.11 \cdot 10^{-11} \pm 7.15 \cdot 10^{-12}$	$8.69 \cdot 10^{-2}$	7
	100	$7.80 \cdot 10^{-22} \pm 2.81 \cdot 10^{-22}$	$18.0 \cdot 10^{-2}$	42
$(0.35B, 0)$	20	$6.35 \cdot 10^{-4} \pm 5.24 \cdot 10^{-5}$	$4.21 \cdot 10^{-2}$	1
	50	$2.61 \cdot 10^{-9} \pm 3.63 \cdot 10^{-10}$	$7.09 \cdot 10^{-2}$	5
	100	$4.61 \cdot 10^{-18} \pm 9.22 \cdot 10^{-19}$	$10.2 \cdot 10^{-2}$	25

via the IS scheme is always 10^{-2}, with the exception of Table 6 where the relative error is $5 \cdot 10^{-2}$.

Typically, the estimator has similar behavior as that observed for the case of the (standard) tandem network, i.e., MS performs best (in terms of relative error) when the system is highly loaded and B is not too large.

5. Discussion

In this article, we designed an asymptotically efficient MS scheme for estimating the probability of first entering some rare set before a tabu set. The scheme is easy to implement, and can always be used, as long as the logarithmic decay rate function of the probability of interest is known (as will typically follow from a large deviations analysis). We also provide a short and elegant proof of asymptotic efficiency of the proposed scheme under some mild assumption on the convergence of the decay rate. As an illustration, we applied the scheme to find the probability of overflow in the downstream buffer of a tandem Jackson network, and of a slowdown network.

We found that the scheme generally provides good estimates in reasonable time. Therefore, it can be a good alternative to IS schemes, especially when the system has high loads (in both nodes), or when the rarity parameter B is not extremely large, i.e., when the event of interest is not so rare. In such cases the relative error may even be lower than the one obtained via IS.

However, in some other cases the relative error of MS is quite high. This undesirable performance can be explained as follows. When the parameters of the network (λ, μ_1 (μ_1^+), μ_2) are clearly distinctive (i.e., when the server loads are low) then the IS scheme performs well, i.e., it requires a relatively small amount of time to obtain good estimates. On the other hand, this is the toughest case for MS since the queue-length process has a strong drift towards the origin. We argue that even in such cases, when MS is obviously outperformed by IS, MS may still be preferred over IS, since it is conceptually easier than IS. The main reason for this is that we do not need to resolve technical issues around the boundaries, as often needed in IS. This is a advantage over IS, especially when we want to simulate larger networks.

Finally, we like to mention that it may be possible to decrease the relative error of MS by fine-tuning the splitting factor R. We did not concentrate on this in the current article, since our primary goal was to design a simple MS scheme and to prove its asymptotic efficiency.

Acknowledgment

Part of this research has been funded by the Dutch BSIK/BRICKS project. This article is also the result of joint research in the 3 TU Centre of Competence Netherlands Institute for Research on ICT within the Federation of Three Universities of Technology in The Netherlands.

References

Dean, T., Dupuis, P. (2009). Splitting for rare event simulation: A large deviation approach to design and analysis. *Stochastic Processes and Their Applications* 19:562–587.

Dupuis, P., Ellis, R. S. (1997). *A Weak Convergence Approach to the Theory of Large Deviations*. New York: Wiley.

Dupuis, P., Sezer, A. D., Wang, H. (2007). Dynamic importance sampling for queueing networks. *Annals of Applied Probability* 17(4):1306–1346.

Dupuis, P., Wang, H. (2009). Importance sampling for Jackson networks. *Queueing Systems* 62(1–2):113–157.

Glasserman, P., Heidelberger, P., Shahabuddin, P., Zajic, T. (1996). Multilevel splitting for estimating rare event probabilities. *IBM Research Report* RC 20478.

Glynn, P. W., Whitt, W. (1992). An asymptotic efficiency of simulation estimators. *Operations Research* 40(3):505–520.

Heidelberger, P. (1995). Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation* 5(1):43–85.

Malhotra, R., Mandjes, M., Scheinhardt, W., van den Berg, H. (2009). A feedback fluid queue with two congestion control thresholds. *Mathematical Methods in Operations Research* 70:149–169.

Miretskiy, D. I., Scheinhardt, W. R. W., Mandjes, M. R. H. (2008). Simple and efficient importance sampling scheme for a tandem queue with server slow-down. *Proc. of RESIM 2008*. Rennes, France, pp. 38–50.

Miretskiy, D. I., Scheinhardt, W. R. W., Mandjes, M. R. H. (2009). State-dependent importance sampling for a slow-down tandem queue. *Annals of Operations Research* 189:299–329.

Miretskiy, D. I., Scheinhardt, W. R. W., Mandjes, M. R. H. (2009). Rare-event simulation for tandem queues: a simple and efficient importance sampling scheme. *Proc. of NET-COOP*. Eindhoven, The Netherlands, pp. 107–120.

Miretskiy, D. I., Scheinhardt, W. R. W., Mandjes, M. R. H. (2010). State-dependent importance sampling for a Jackson tandem network. *In ACM Transactions on Modeling and Computer Simulation* 20(3):Article 15.

Noureddine, W., Tobagi, F. (1999). Selective back-pressure in switched Ethernet LANs. *Proc. of Global Telecommunications Conference*. Vol. 2. Rio de Janeiro, Brazil, pp. 1256–1263.

Shahabuddin, P. (1995). Rare event simulation in stochastic models. *Proc. of the 1995 Winter Simulation Conference*. Arlington, VA, pp. 178–185.

Van Foreest, N. D., Mandjes, M. R. H., van Ommeren, J. C. W., Scheinhardt, W. R. W. (2005). A tandem queue with server slow-down and blocking. *Stochastic Models* 21(2–3):695–724.

Villén-Altamirano, M., Villén-Altamirano, J. (2006). On the efficiency of Restart for multidimensional state systems. *ACM Transactions on Modeling and Computer Simulation* 16(3):251–279.