Fast Simulation of Overflow Probabilities in a Queue with Gaussian Input

A. B. DIEKER and M. MANDJES
CWI and University of Twente

In this paper, we study a queue fed by a large number n of independent discrete-time Gaussian processes with stationary increments. We consider the *many-sources* asymptotic regime, i.e., the buffer-exceedance threshold B and the service capacity C are scaled by the number of sources ($B \equiv nb$ and $C \equiv nc$).

We discuss four methods for simulating the steady-state probability that the buffer threshold is exceeded: the single-twist method (suggested by large-deviation theory), the cut-and-twist method (simulating timeslot by timeslot), the random-twist method (the twist is sampled from a discrete distribution), and the sequential-twist method (simulating source by source).

The asymptotic efficiency of these four methods is analytically investigated for $n \to \infty$. A necessary and sufficient condition is derived for the efficiency of the single-twist method, indicating that it is nearly always asymptotically inefficient. The other three methods, however, are asymptotically efficient. We numerically evaluate the four methods by performing a detailed simulation study, where it is our main objective to compare the three efficient methods in practical situations.

Categories and Subject Descriptors: G.3 [**Mathematics of Computing**]: Probability and Statistics

General Terms: Algorithms, Performance
Additional Key Words and Phrases: Asymptotic efficiency, Gaussian processes, importance sampling, large deviations, overflow probability, queueing theory

1. INTRODUCTION

Many systems in real life can be modeled as *queues*. The generic queueing model consists of (i) a (random) arrival process, and (ii) a resource, commonly characterized by its service speed C, and buffer space B. If the traffic arrival rate temporarily exceeds C, work is stored in the buffer, and, after some delay, served. Traffic that does not fit into the buffer is lost. Hence, queues are an appropriate tool for describing congestion phenomena.

Corresponding author: A. B. Dieker. Email: ton@cwi.nl.
A. B. Dieker's work was supported by the Netherlands Organization for Scientific Research (NWO) under grant 631.000.002.

Gaussian traffic. In this paper, we consider a queue fed by *Gaussian* traffic. We focus on stationary input, i.e., the distribution of the traffic offered in an interval only depends on the interval length. The study of Gaussian input is mainly motivated by its flexibility and parsimony: a broad range of correlation structures can be described by few parameters. Notably, Gaussian processes may exhibit 'power-law correlations' corresponding to long-range dependence; an example is fractional Brownian motion (fBm). Processes of this type can be used to accurately model network data traffic. The focus on Gaussian models can also be justified from the fact that in many practical situations a large number of independent sources are superimposed; by virtue of central-limit-type arguments, one can argue that the aggregate traffic converges to a Gaussian process, see, e.g., Dębicki and Palmowski [1999].

Asymptotics. It is notoriously hard to calculate the full buffer content distribution of a queue with Gaussian input; in fact, one has succeeded in this only for simple special cases (Brownian motion, Brownian bridge). However, some limiting regimes allow explicit analysis. The present paper focuses on the so-called *many-sources regime*. In this regime we suppose that there are n i.i.d. Gaussian sources, and that the queueing resources are scaled with n, i.e., $C \equiv nc$ and $B \equiv nb$. The probability that the steady-state buffer content exceeds level nb becomes small when n grows large. For fixed but large n, we study this *buffer-content probability* p_n in a discrete-time model.

Likhanov and Mazumdar [1999] find the asymptotics of p_n, i.e., they identify a function g such that $p_n g(n) \to 1$ as $n \to \infty$; notably, they find that p_n decays roughly exponentially in n. Based on these asymptotics, one could estimate p_n by $1/g(n)$. However, due to the lack of error bounds one does not know *a priori* whether these estimates are any good. More specifically, we do not have an $n_0 = n_0(\epsilon)$ such that, for all $n > n_0$, it holds that $|p_n g(n) - 1| < \epsilon$, where $\epsilon > 0$ is a (small) parameter. In fact, the derivation of $p_n g(n) \to 1$ indicates that $1/g(n)$ has the undesirable property that it tends to underestimate p_n, cf. Equations (2.1) and (3.4) in Likhanov and Mazumdar [1999].

Simulation. In absence of analytical results (or asymptotic results that are backed up by error bounds), one could resort to simulation. When simulating loss probabilities in queues with Gaussian input, essentially two problems arise. The first is that it is not straightforward to quickly simulate Gaussian processes, see for instance Dieker and Mandjes [2003]. Although 'exact' methods for generating (discrete versions of) Gaussian processes are in general quite slow, a sophisticated simulation technique becomes available by exploiting the stationarity of the sources [Davies and Harte 1987]. In the important case of fBm, this leads to a fast algorithm (order of $T \log T$ for a trace of length T) for generating fBm traces. An inherent difficulty with this algorithm is that the trace length should be specified before the simulation

is started.

The second problem of simulation is that it is typically hard to estimate small probabilities; we mainly focus on this difficulty in the remainder of this paper, as in our setting $p_n \to 0$ as $n \to \infty$. The general rule is that, for an estimate with a fixed relative precision, the number of runs needed is inversely proportional to the probability to be estimated. Hence it is impractical, or even impossible, to estimate a probability of less than, say, 10^{-9} with conventional Monte Carlo simulation. This problem could be circumvented by performing a 'fast simulation' using a technique that is known as importance sampling. In importance sampling, samples are drawn from a distribution under which the event under consideration occurs more frequently. An unbiased estimator is obtained by weighing the simulation output by likelihood ratios. Inherently, one has the freedom to choose the importance sampling measure, and the challenge is to find the measure that is in some sense 'most efficient'. A widely accepted efficiency criterion for discriminating between estimators is *asymptotic efficiency*, sometimes referred to as *asymptotic optimality* or *logarithmic efficiency*. The analysis in the present paper is based on this criterion.

Contributions. Estimators based on large-deviation theory are natural candidates for efficient simulation. In fact, they are asymptotically efficient in many settings; see Asmussen and Rubinstein [1995] and Heidelberger [1995] and references therein. However, Glasserman and Wang [1997] give examples showing that this need not always be the case. A main contribution of this paper is that we develop conditions for asymptotic efficiency (as $n \to \infty$) of the large-deviation estimator that would apply to our buffer-content probability. It turns out that this estimator is predominantly asymptotically *inefficient* for a wide range of Gaussian inputs, including fBm and (perhaps surprisingly) even standard Brownian motion.

As the large-deviation estimator is inefficient in practice, a different approach has to be taken. We present three other methods that can be proven to be asymptotically efficient. The first uses ideas of Boots and Mandjes [2002], and simulates timeslot by timeslot. The second method is a randomized version of the large-deviation estimator; it is based on the work of Sadowsky and Bucklew [1990]. A third method relies on a recent paper by Dupuis and Wang [2004], and simulates source by source. In the latter approach, the change of measure of the source under consideration depends on the traffic generated by the sources that have already been simulated. We present a detailed performance evaluation of these four approaches, both analytically and empirically.

Some related results on fast simulation of queues with Gaussian input have been reported by Michna [1999] and by Huang *et al.* [1999]. Michna focuses on fBm input under the so-called *large-buffer scaling* (rather than our many-sources setting), but does not consider asymptotic efficiency of his simulation scheme. The study of Huang *et al.* also relates to the large-buffer asymptotic regime for fBm input. They empirically assess the variance reduction of their proposed change of measure, but

do not formally derive properties of their estimator (such as asymptotic efficiency); in fact, Lemma 4.2 below entails that their estimator is unnatural from the point of view of asymptotic efficiency. We would like to stress that the present paper focuses *only* on the simulation of the buffer-content probability in the many-sources regime (with general Gaussian input, not necessarily fBm).

The many-sources regime has been generally accepted as a framework that is particularly suitable when studying large multiplexing systems, and in this sense it is a useful alternative to the large-buffer regime, see for instance the reflections in Kelly [1996] and Ganesh *et al.* [2004]. It is also remarked that the aggregate of multiple i.i.d. Gaussian sources is again Gaussian. This implies that in the timeslot-by-timeslot approach it suffices to simulate just the aggregate input process, rather than the individual sources. In the source-by-source approach, however, one explicitly exploits the fact that there are multiple sources to obtain an asymptotically efficient algorithm.

Organization. This paper is organized as follows. Section 2 formalizes the queueing framework used in this paper. It also discusses how the simulation horizon can be truncated, so that we can work with traces of prespecified length. Preliminaries on importance sampling are given in Section 3. Section 4 studies the asymptotic efficiency of the four simulation methods mentioned above from an analytical perspective. Section 5 contains a numerical evaluation of these methods, to assess their performance under practical circumstances. The paper concludes with a discussion in Section 6.

2. THE BUFFER-CONTENT PROBABILITY

The present section contains the description of our queueing model. In particular, we show that the buffer-content probability can be translated into an exceedance probability of the so-called free process on an infinite time interval, see Section 2.1. To simulate this exceedance probability, the infinite time interval needs to be truncated, where the neglected probability mass is below a tolerable level. Under the truncation, we can obviously work with Gaussian traces of prespecified length. The truncation issue is addressed in Section 2.2.

2.1 Description of the model — many-sources framework

Traffic model. We start by describing the traffic model. We consider n i.i.d. sources feeding into a buffered resource. The sources are assumed to be *stationary*, so that the distribution of the traffic generated in an interval $[s, s+t)$ only depends on the interval length t (and not on the 'position' s). We focus on a discrete-time system, i.e., time is indexed by $\mathbb{Z}$.

Define $A_n(\cdot)$ as the aggregate cumulative traffic process. More precisely, let $A_n(s,t)$ denote the traffic generated by the superposition of the n sources in the interval $\{s,\ldots,t\}$, with $s \le t$. For notational convenience, we set $A_n(t) := A_n(0,t)$,

so that $A_n(s,t) = A_n(t) - A_n(s)$. We also suppose that $A_n(0) = 0$; this can be assumed without loss of generality, since we are interested in steady-state behavior.

In this paper, we assume that the sources are *Gaussian*, so that the distribution of $A_n(\cdot)$ is completely determined by the mean input rate and the covariance structure. Let μ denote the mean input rate of a single source, i.e., $\mathbb{E}A_n(t) =: n\mu t$. Because the stationarity of the sources results in stationary increments of the process A_n, the covariance structure is determined by the variance function $\sigma^2(t) := \mathbb{V}\mathrm{ar}A_1(t)$. We suppose that $\sigma^2(t)t^{-\alpha} \to 0$ as $t \to \infty$ for some $\alpha \in (0,2)$; the Borel-Cantelli lemma then shows that $A_1(t)/t \to \mu$ almost surely, see the proof of Lemma 3 in [Dieker 2005].

It is readily deduced that the covariance of $A_n(\cdot)$ is given by $\Gamma_n(s,t) = n\Gamma(s,t)$, where

$$\Gamma(s,t) := \mathbb{C}\mathrm{ov}(A_1(s), A_1(t)) = \frac{\sigma^2(s) + \sigma^2(t) - \sigma^2(|s-t|)}{2}.$$

An important special case of Gaussian input is *fractional Brownian motion* (fBm), for which $\sigma^2(t)$ is (proportional to) t^{2H}.

Queueing model. We now turn to the queueing model. In this paper, we scale the queue's (deterministic) service rate with the number of sources: the queue drains at rate $C \equiv nc$. To ensure stability, we assume that $\mu < c$.

We are interested in the steady-state probability p_n of the buffer content exceeding some prespecified level, which we again scale with the number of sources: $B \equiv nb > 0$. We first express this probability in terms of the aggregate cumulative arrival process $A_n(\cdot)$, as follows. Let $Q_n(t)$ be the buffer content in the n-scaled model at time $t \in \mathbb{Z}$:

$$Q_n(t) := \sup_{s \in \{0,\ldots,t\}} [A_n(s,t) - nc(t-s)],$$

see for instance Equation (2.1) of Norros [1994]. The steady-state probability p_n of the buffer content exceeding nb reads

$$p_n := \lim_{t\to\infty} P(Q_n(t) > nb) = \lim_{t\to\infty} P\left(\sup_{s\in\{0,\ldots,t-1\}} [A_n(s,t) - nc(t-s)] > nb \right).$$

For any given $t = 1, 2, \ldots$, $\{A_n(s,t) : 0 \le s \le t-1\}$ has the same distribution as $A_n(t-s) : 0 \le s \le t-1\}$; this is called *time-reversibility* of $A_n(\cdot)$. Therefore, p_n can be rewritten as

$$p_n = \lim_{t\to\infty} P\left(\sup_{s\in\{1,\ldots,t\}} [A_n(s) - ncs] > nb \right) = P\left(\sup_{t\in\{1,2,\ldots\}} [A_n(t) - nct] > nb \right). \tag{1}$$

Indeed, Equation (1) constitutes a quite remarkable, and perhaps slightly counter-intuitive, result: the steady-state buffer content of the queueing process (which is reflected at 0, and hence takes values in $[0,\infty)$) *has the same distribution as* the

supremum of the free process (i.e., the process $A_n(t) - nct$, which is *not* reflected at 0). This useful duality between the buffer-content probability p_n and exceedance probabilities of the free process is discussed in greater detail in, for instance, Section V.4 of Asmussen [2000].

In view of the duality of Representation (1), we can conclude that there are essentially two ways of simulating the buffer-content probability:

—In the first place, one could estimate the buffer-content probability from the evolution of the *reflected process* (i.e., the buffer-content process). However, standard simulation approaches to do this have their intrinsic difficulties: a regenerative approach fails in the context of Gaussian inputs (notice that busy periods are dependent!), whereas in a 'batch-means' approach the estimator could suffer from the relatively strong dependencies between the batches (particularly when the Gaussian process is long-range dependent).

—As an alternative, one could estimate the buffer-content probability from sample paths of the *free process* $A_n(t) - nct$. Every run is an independent sample of this free process, and the corresponding estimator is the fraction of runs in which nb is exceeded (for some $t \in \mathbb{N}$). This approach clearly overcomes the above-mentioned problems arising when estimating p_n from the reflected process.

A practical difficulty of the latter approach, however, relates to the infinite 'simulation horizon' involved (it needs to be verified whether the free process exceeds nb for some $t \in \mathbb{N}$), but this issue can be addressed, see the next subsection. Motivated by these arguments, we have chosen to use in this paper the representation of the buffer-content probability as an exceedance probability of the free process; in other words: we estimate p_n relying on the right-hand side of (1).

We remark that the probability p_n of the steady-state buffer content being larger than nb in a system with *infinite* buffer is often used as an approximation for the loss probability in a system with *finite* buffer nb.

We emphasize that the behavior of the probability p_n in discrete time is essentially different from continuous time. The buffer-content probability in continuous time is obtained by replacing $\mathbb{N}$ by $\mathbb{R}_+$ in Eq. (1). Notably, the asymptotics of the buffer-content probability in continuous time differ qualitatively from those in (1), see Dębicki and Mandjes [2003]. A further discussion of this issue is relegated to Section 6.

2.2 The simulation horizon

Representation (1) shows that the buffer-content probability equals an exceedance probability on an *infinite* time horizon. Hence, to estimate p_n through simulation, we first have to truncate $\mathbb{N}$ to $\{1, \ldots, T\}$, for some finite T, while still controlling the error made. That can be done as follows.

Suppose we approximate p_n by

$$p_n^T := P\left(\sup_{t\in\{1,\ldots,T\}} [A_n(t) - nct] > nb\right). \tag{2}$$

This is evidently a probability smaller than p_n, but the larger T the smaller the error. We now analyze how large T should be. Define $\tau_n := \inf\{t \in \mathbb{N} : A_n(t) - nct > nb\}$, so that $p_n = P(\tau_n < \infty)$. As we propose to approximate p_n by $P(\tau_n \leq T)$, we discard the contribution of $P(T < \tau_n < \infty)$. As in Boots and Mandjes [2002], we choose T such that

$$\frac{P(T < \tau_n < \infty)}{p_n} < \epsilon, \tag{3}$$

for some predefined, typically small, $\epsilon > 0$. When choosing ϵ small enough, the truncation is of minor impact. Clearly, the smaller ϵ, the larger the T required.

The requirement in (3) does not directly translate into an explicit expression for the simulation horizon T as a function of ϵ and n. Following Boots and Mandjes [2002], this problem is tackled by establishing tractable bounds on $P(T < \tau_n < \infty)$ and p_n: with a lower bound on p_n and an upper bound on $P(T < \tau_n < \infty)$, we can choose T so large that $P(T < \tau_n < \infty)/p_n < \epsilon$. We write

$$I_t := \frac{(b + (c-\mu)t)^2}{2\sigma^2(t)}.$$

A lower bound on p_n. Obviously, for any $t \in \mathbb{N}$, application of (1) entails

$$\begin{aligned} p_n &\geq P(A_n(t) > nb + nct) \\ &= \int_{\sqrt{n}\frac{b+(c-\mu)t}{\sigma(t)}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) dx \\ &\geq \frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{nI_t} + \sqrt{nI_t + 2}} e^{-nI_t}, \end{aligned} \tag{4}$$

where the last inequality is a standard bound for the standard normal cumulative density function (see Mitrinović [1970, p. 177–181] for related inequalities and references).

In order to find the best possible lower bound, we compute $t^* := \arg\inf_{t\in\mathbb{N}} I_t$ and use the lower bound (4) for $t = t^*$. The existence of t^* is guaranteed by the assumption that $\sigma^2(t)t^{-\alpha} \to 0$ as $t \to \infty$ for some $\alpha \in (0, 2)$. In case t^* is unique, it is usually referred to as the 'most probable' exceedance epoch: given that the free process $A_n(t) - nct$ exceeds nb, it is most likely that it happens at epoch t^*; see for instance Wischik [2001b].

An upper bound on $P(T < \tau_n < \infty)$. By a Chernoff-bound argument, we have

$$P(T < \tau_n < \infty) = \sum_{t=T+1}^{\infty} P(\tau_n = t) \leq \sum_{t=T+1}^{\infty} P(A_n(t) - nct > nb) \leq \sum_{t=T+1}^{\infty} e^{-nI_t}. \tag{5}$$

In the present generality, it is difficult to further bound this sum. We could proceed by focusing on a specific correlation structure, such as fBm for which $\sigma^2(t) = t^{2H}$, where $H \in (0,1)$. Instead, we focus on the somewhat more general situation that the variance function can be bounded (from above) by a polynomial: $\sigma^2(t) \leq Ct^{2H}$, for some $H \in (0,1)$ and $C \in (0,\infty)$. For instance, if $\sigma^2(\cdot)$ is regularly varying (see, e.g., Bingham *et al.* [1989]) with index α, then $\sigma^2(t)$ can be bounded from above (for t sufficiently large) by $t^{\alpha+\delta}$, for $\delta > 0$; see Prop. 1.5.1 of Bingham *et al.* [1989]. Obviously, it is desirable to choose the horizon as small as possible under the restriction that (3) holds; for this, C and H should be chosen as small as possible.

Under $\sigma^2(t) \leq Ct^{2H}$ we can bound (5) as follows:

$$\sum_{t=T+1}^{\infty} e^{-nI_t} \leq \sum_{t=T+1}^{\infty} \exp\left(-n\frac{(c-\mu)^2}{2C}t^{2-2H}\right) \leq \int_T^{\infty} \exp\left(-n\frac{(c-\mu)^2}{2C}t^{2-2H}\right) dt.$$

It turns out that we have to consider the cases $H \leq 1/2$ and $H > 1/2$ separately. For $H \leq 1/2$, the following bound is readily found (its proof is deferred to Appendix A.1.1). Set $C_0 := (c-\mu)^2/(2C)$ and $q := 1/(2-2H)$ for notational convenience.

LEMMA 2.1. *In case* $H \leq 1/2$*, we have*

$$\int_T^{\infty} \exp\left(-nC_0t^{1/q}\right) dt \leq \frac{q}{C_0 n} \exp\left(-nC_0T^{1/q}\right). \tag{6}$$

We now focus on $H > 1/2$ (and hence $q > 1$). Let m be the largest natural number such that $q - 1 - m \in (0,1]$. Moreover, we define

$$\gamma_q := q - 1 - m, \text{ and } \beta_q := \frac{(q-1)\cdots(q-m)}{\gamma_q^m e^{\gamma_q}}. \tag{7}$$

These quantities play a central role in the following lemma, which is proven in Appendix A.1.2.

LEMMA 2.2. *In case* $H > 1/2$*, we have*

$$\int_T^{\infty} \exp\left(-nC_0t^{1/q}\right) dt \leq \frac{q\beta_q}{C_0^q(n-\gamma_q)} \exp\left(-(n-\gamma_q)C_0T^{1/q}\right).$$

By combining the upper bounds and the lower bound, we derive the following corollary:

COROLLARY 2.3. *For* $H \leq 1/2$*, let* $T(n)$ *be the smallest integer larger than*

$$\left(-\frac{1}{nC_0}\log\left[\frac{1}{q\sqrt{\pi}}\frac{nC_0\epsilon}{\sqrt{nI_{t^*}}+\sqrt{nI_{t^*}+2}}e^{-nI_{t^*}}\right]\right)^q,$$

and for $H > 1/2$ let $T(n)$ be the smallest integer larger than

$$\left(-\frac{1}{nC_0}\log\left[\frac{1}{q\beta_q\sqrt{\pi}}\frac{(n-\gamma_q)C_0^q\epsilon}{\sqrt{nI_{t^*}}+\sqrt{nI_{t^*}+2}}e^{-nI_{t^*}}\right]\right)^q.$$

Then the error as defined in (3) *does not exceed ϵ.*
Moreover, $\overline{T} := \lim_{n\to\infty} T(n) = (I_{t^*}/C_0)^{1/(2-2H)}$.

Recall that t^* could be interpreted as the most likely epoch at which the supremum in (1) is attained. Hence, it is not surprising that $\overline{T} > t^*$:

$$\frac{I_{t^*}}{C_0} = \frac{(b+(c-\mu)t^*)^2}{2\sigma^2(t^*)}\Big/\frac{(c-\mu)^2}{2C} > (t^*)^{2-2H} = (t^*)^{1/q}. \tag{8}$$

2.3 Hurst parameter

In this subsection, we investigate the influence of the Hurst parameter on the simulation horizon. This is of special interest since the computational effort to obtain estimates with the cut-and-twist method (see Section 4.2) is extremely sensitive to this horizon.

As already observed, the limiting value (as $n \to \infty$) of the simulation horizon is given by $(I_{t^*}/C_0)^{1/(2-2H)}$, which equals by definition

$$T = T(H) = \left(\inf_{t\in\mathbb{N}} \frac{b+(c-\mu)t}{(c-\mu)t^H}\right)^{1/(1-H)}.$$

Assuming that the infimum is taken over the whole real halfline, we see that $T(H)$ can be approximated by

$$\tilde{T}(H) := \frac{b}{c-\mu}\frac{H^{H/(H-1)}}{1-H}.$$

Clearly, $\tilde{T}(H)$ has a pole at $H = 1$, but it is insightful to plot $\tilde{T}$ as a function of H and see how quickly it tends to infinity. Set $b/(c-\mu) = 1$. In Figure 1, we have plotted this function and its derivative.

It is intuitively clear that $\tilde{T}(H)$ increases in H. The higher H is, the more long-term correlations are present, and more time is needed until unusual behavior is diminished. In practice, it will hardly be possible to simulate the probability with relative error at most ϵ if $H > 0.95$, cf. (3).

3. PRELIMINARIES ON RARE-EVENT SIMULATION

This section provides some background on the simulation of (small) probabilities. Section 3.1 reviews the concept of *importance sampling*, one of the standard techniques in rare-event simulation. The key metric for evaluating simulation approaches is the so-called *asymptotic effiency*, as defined in Section 3.2.

3.1 Importance sampling

Importance sampling is a variance reduction technique in which samples are drawn from a distribution under which the rare event occurs relatively frequently. The

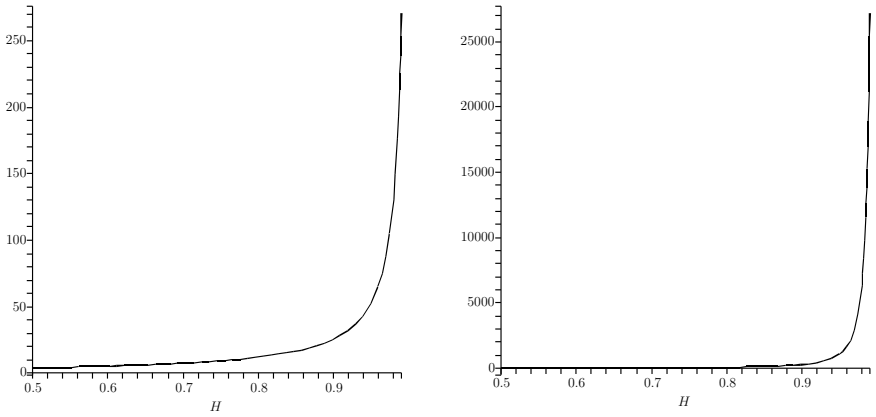

Fig. 1. $\tilde{T}(H)$ as a function of H (left panel) and its derivative (right panel).

simulation output is weighed by so-called likelihood ratios, keeping track of the difference between the original and new measures, thus obtaining unbiased estimates.

More formally, suppose that we are given a probability measure ν on some measurable space $(\mathcal{X}, \mathcal{B})$, and that we are interested in the simulation of the ν-probability of a given event $A \in \mathcal{B}$, where $\nu(A)$ is typically small. The idea of importance sampling is to sample from a different distribution on $(\mathcal{X}, \mathcal{B})$, say λ, under which A occurs more frequently. This is done by specifying a measurable function $d\lambda/d\nu : \mathcal{X} \to [0, \infty]$ and by setting

$$\lambda(B) := \int_B \frac{d\lambda}{d\nu} d\nu.$$

Since λ must be a probability measure, $d\lambda/d\nu$ should integrate to unity with respect to ν.

Assuming the equivalence of the measures ν and λ, set $d\nu/d\lambda := (d\lambda/d\nu)^{-1}$ and note that

$$\nu(A) = \int_A \frac{d\nu}{d\lambda} d\lambda = \int_{\mathcal{X}} \mathbf{1}_A \frac{d\nu}{d\lambda} d\lambda,$$

where $\mathbf{1}_A$ denotes the indicator function of A. We refer to $d\nu/d\lambda$ as the *likelihood ratio* (or simply likelihood). The importance sampling estimator $\widehat{\nu_\lambda(A)}$ of $\nu(A)$ is found by drawing N independent samples $X^{(1)}, \ldots, X^{(N)}$ from λ:

$$\widehat{\nu_\lambda(A)} := \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}_{\{X^{(k)} \in A\}} \frac{d\nu}{d\lambda}(X^{(k)}). \tag{9}$$

It is clear that $\widehat{\nu_\lambda(A)}$ is an unbiased estimator, i.e., $\mathbb{E}_\lambda \widehat{\nu_\lambda(A)} = \nu(A)$. However, one has the freedom to choose the distribution λ; a good choice results in an estimator with small variance. In particular, it is of interest to find the change of measure that

minimizes this variance. Since $\widehat{\nu_\lambda(A)}$ is by construction unbiased, it is equivalent to minimize the second moment

$$\int_A \left(\frac{d\nu}{d\lambda}\right)^2 d\lambda = \int_{\mathcal{X}} \mathbf{1}_A \left(\frac{d\nu}{d\lambda}\right)^2 d\lambda.$$

It is not difficult to see that a zero-variance estimator is found by letting λ be the conditional distribution of ν given A, see, e.g., Heidelberger [1995]. However, the resulting estimator is infeasible for simulation purposes, since then $d\nu/d\lambda$ depends on the *unknown* quantity $\nu(A)$. This motivates the use of another optimality criterion, *asymptotic efficiency*.

3.2 Asymptotic efficiency

In order to compare simulation techniques the notion of *asymptotic efficiency* was introduced. Consider a family of probability measures $\{\nu_n\}$ on $(\mathcal{X}, \mathcal{B})$. Suppose we associate to each ν_n an importance sampling distribution λ_n on $(\mathcal{X}, \mathcal{B})$; in Section 4, we study several choices for λ_n.

Let $X_{\lambda_n}^{(1)}, \ldots, X_{\lambda_n}^{(N)}$ be N i.i.d. samples from λ_n. We define the importance sampling estimator of $\nu_n(B)$ as in (9):

$$\widehat{\nu_{\lambda_n}(B)}_N := \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}_{\left\{X_{\lambda_n}^{(k)} \in B\right\}} \frac{d\nu_n}{d\lambda_n}\left(X_{\lambda_n}^{(k)}\right). \tag{10}$$

The *relative error* of the importance sampling estimator is defined as

$$\eta_N(\lambda_n, B) := \frac{\sqrt{\mathbb{V}\mathrm{ar}_{\lambda_n}\left(\widehat{\nu_{\lambda_n}(B)}_N\right)}}{\nu_n(B)} = \frac{\sqrt{\mathbb{E}_{\lambda_n}\left(\widehat{\nu_{\lambda_n}(B)}_N\right)^2 - \nu_n(B)^2}}{\nu_n(B)}; \tag{11}$$

here the notation $\mathbb{V}\mathrm{ar}_{\lambda_n}(\cdot)$ and $\mathbb{E}_{\lambda_n}(\cdot)$ indicates integration with respect to λ_n. Notice that the relative error, i.e., the square root of (11), is proportional to the width of a confidence interval relative to the (expected) estimate itself; hence, it measures the variability of the importance sampling estimator. Let $N_{\lambda_n}^* := \inf\{N \in \mathbb{N} : \eta_N(\lambda_n, B) \leq \eta_{\max}\}$ be the number of samples needed for a prespecified relative error. For asymptotic efficiency we require that this number vanishes on an exponential scale. Asymptotic efficiency is sometimes referred to as *asymptotic optimality*, *logarithmic efficiency*, or *weak efficiency*. The following notion was introduced by Sadowsky [1991].

Definition 3.1. An importance sampling family $\{\lambda_n\}$ is called *asymptotically efficient* if

$$\limsup_{n\to\infty} \frac{1}{n} \log N_{\lambda_n}^* = 0,$$

for some given maximal relative error $0 < \eta_{\max} < \infty$.

It is important to observe that studying an importance sampling family for $n \to \infty$ does not necessarily yield knowledge on the performance of the corresponding

estimator for a *given* n. For instance, if one family induces a bounded relative error (in n) and another yields an unbounded relative error which still vanishes on an exponential scale, the first is clearly preferable. According to the above criterion, both families are then called asymptotically efficient, and hence the asymptotic efficiency criterion disguises the nice property of bounded relative error for the first family. In the context of the simulation methods that we study in this paper, we further clarify this point in Section 5.5.

We also note that, under a weak condition on the sets B, asymptotic efficiency is equivalent to $\limsup_{n\to\infty} E_n \geq 2$, with

$$E_n := \frac{\log \int_B \left(\frac{d\nu_n}{d\lambda_n}\right)^2 d\lambda_n}{\log \nu_n(B)}; \tag{12}$$

see criterion (2) of Asmussen and Binswanger [1997], and Dieker and Mandjes [2005] for more details. For a given n, we refer to E_n as the *relative efficiency*.

4. SIMULATION METHODS

Using the bounds of Section 2.2, the simulation horizon can be truncated. We therefore focus in the sequel of the paper on the simulation of this 'truncated' buffer-content probability $p_n^{T(n)}$ defined in (2).

As argued in Section 3.2, asymptotic efficiency corresponds to the performance of simulation methods for large n. Notice that, by virtue of Corollary 2.3, we can safely set $T(n) = \lceil \overline{T} \rceil$ for n large enough; for ease denote $T := \lceil \overline{T} \rceil$. Conclude that we can restrict ourselves to assess asymptotic efficiency of methods for estimating p_n^T.

In this paper, we concentrate on four methods for simulating p_n^T. The first, which we refer to as the *single (exponential) twist* method, is the simplest of the four. We present explicit conditions on the covariance structure of the Gaussian sources under which the method is asymptotically efficient. It appears that for important cases the method does *not* yield asymptotic efficiency. Therefore, we also discuss three asymptotically efficient alternatives: the first solves the theoretical difficulties by simulating timeslot by timeslot (which we therefore call *cut-and-twist*), the second by randomization of the twist (*random twist*), and the third by simulating source by source (*sequential twist*).

4.1 The single-twist method

Large-deviation theory suggests an importance sampling distribution based on an exponential change of measure ('twist'). In a considerable number of simulation settings this alternative distribution has shown to perform well — in some cases it is asymptotically efficient, see for instance Asmussen [1989], Bucklew *et al.* [1990], Collamore [2002], Lehtonen and Nyrhinen [1992a], Lehtonen and Nyrhinen [1992b], Siegmund [1976]. However, one has to be careful, as a successful application of such an exponential twist critically depends on the specific problem at hand, see

e.g. Dieker and Mandjes [2005], Dupuis and Wang [2004], Glasserman and Wang [1997]. Before deriving conditions for asymptotic optimality of the exponential twist in the setup of the present paper, we first provide more background.

We denote

$$\begin{aligned}\mathcal{O}_T &:= \{x \in \mathbb{R}^T : \exists t \in \{1,\ldots,T\} : x_t + \mu t \geq b + ct\} \\ &= \bigcup_{t\in\{1,\ldots,T\}} \bigcup_{\{y: y+\mu t \geq b+ct\}} \{x \in \mathbb{R}^T : x_t = y\},\end{aligned} \tag{13}$$

such that

$$p_n^T = \nu_n^{(T)}(\mathcal{O}_T),$$

with $\nu_n^{(T)}$ denoting the distribution of the centered (i.e., zero mean) process

$$\{A_n(t)/n - \mu t : t = 1,\ldots,T\}.$$

The following lemma, which is proven in Appendix A.2, states that $\nu_n^{(T)}(\mathcal{O}_T)$ decays exponentially in n. We let $\Gamma^{(T)}$ denote the covariance matrix of $\{A_1(t) - \mu t : t = 1,\ldots,T\}$, i.e., $\Gamma^{(T)} := \{\Gamma(s,t) : s,t = 1,\ldots,T\}$. All proofs for this subsection are given in Appendix A.2.

LEMMA 4.1. *We have*

$$\lim_{n\to\infty} \frac{1}{n} \log \nu_n^{(T)}(\mathcal{O}_T) = -\frac{1}{2} x^{*\prime} \left(\Gamma^{(T)}\right)^{-1} x^* = -I_{t^*}, \tag{14}$$

where $t^* := \arg\inf_{t\in\mathbb{N}} I_t$*, and the vector* $x^* \in \mathbb{R}^T$ *is given by*

$$x_t^* := \frac{b + (c-\mu)t^*}{\sigma^2(t^*)} \Gamma(t^*, t). \tag{15}$$

Recall that time epoch t^* can be thought of as the most likely epoch that the free process $A_n(t) - nct$ exceeds nb: as n grows, the probability of exceeding nb vanishes, but given that it occurs, with overwhelming probability it occurs at t^*. Likewise, x^* can informally be interpreted as the *most likely path to exceedance*; note that indeed $x_{t^*}^* = b + (c-\mu)t^*$. It is important to realize that x^* is piecewise linear only in the case of (scaled) Brownian input (i.e., $\sigma^2(t) = Ct$ for some $C > 0$); in general x^* is a 'curved' path.

We can now introduce the family $\{\lambda_n^{(T)}\}$ of exponentially-twisted probability measures. The probability mass assigned to a Borel set $A \subset \mathbb{R}^T$ under this new distribution is

$$\lambda_n^{(T)}(A) = \int_A \exp\left(n \frac{b + (c-\mu)t^*}{\sigma^2(t^*)} x_{t^*} - nI_{t^*}\right) \nu_n^{(T)}(dx). \tag{16}$$

Observe that

$$(\Gamma^{(T)})^{-1} x^* = \frac{b + (c-\mu)t^*}{\sigma^2(t^*)} \begin{pmatrix} \Gamma(1,1) & \cdots & \Gamma(1,T) \\ \vdots & & \vdots \\ \Gamma(T,1) & \cdots & \Gamma(T,T) \end{pmatrix}^{-1} \begin{pmatrix} \Gamma(t^*,1) \\ \vdots \\ \Gamma(t^*,T) \end{pmatrix}$$

$$= \frac{b+(c-\mu)t^*}{\sigma^2(t^*)} e_{t^*},$$

where e_i is a T-dimensional vector of zeros, except for a one on the i-th position. As a result, cf. (14),

$$\frac{1}{2} x^{*\prime} \left(\Gamma^{(T)}\right)^{-1} x^* = \frac{b+(c-\mu)t^*}{2\sigma^2(t^*)} x^*_{t^*} = I_{t^*}.$$

Also, the density on $\mathbb{R}^T$ corresponding to $\lambda_n^{(T)}$ reduces to

$$\exp\left(n\frac{b+(c-\mu)t^*}{\sigma^2(t^*)} x_{t^*} - nI_{t^*}\right) \frac{1}{(\sqrt{2\pi})^T |\Gamma^{(T)}/n|^{1/2}} \exp\left(-\frac{n}{2} x' \left(\Gamma^{(T)}\right)^{-1} x\right)$$
$$= \frac{1}{(\sqrt{2\pi})^T |\Gamma^{(T)}/n|^{1/2}} \exp\left(-\frac{n}{2}(x-x^*)' \left(\Gamma^{(T)}\right)^{-1} (x-x^*)\right);$$

here it is used that

$$x' \left(\Gamma^{(T)}\right)^{-1} x^* = \frac{b+(c-\mu)t^*}{\sigma^2(t^*)} x_{t^*}.$$

In other words: the new measure $\lambda_n^{(T)}$ corresponds to the distribution of a Gaussian process with mean vector $\{x_t^* : t = 1, \ldots, T\}$, and covariance matrix $\Gamma^{(T)}/n$. Remark that the mean vector of the new measure is different from the old mean (in fact, the new Gaussian process does *not* correspond to stationary sources anymore), whereas the covariances under the old and new measure coincide. Since samples from $\lambda_n^{(T)}$ tend to follow the most likely path x^* for large n, we say that this exponential twist is in accordance with the large-deviation behavior of Lemma 4.1.

In the Gaussian setting, the above calculation shows that exponentially twisting amounts to changing the mean vector; see, e.g., Huang *et al.* [1999] and Dieker and Mandjes [2005]. Indeed, the above calculations show that the covariance structure remains unchanged, while only the mean changes. Huang *et al.* [1999] (see their Eq. (16)) and Michna [1999] propose to take a straight path as the mean vector, as opposed to the 'curved' most-likely path x^*. The following lemma shows, however, that x^* is in fact the 'best' way to change the mean (i.e., the only candidate that possibly yields asymptotic efficiency).

LEMMA 4.2. *Any mean vector different from x^* does not yield asymptotic efficiency.*

Lemma 4.2 further motivates the verification of the asymptotic efficiency of the twisted distribution $\lambda_n^{(T)}$, and the following theorem is therefore the main result of this subsection. It presents sufficient and necessary conditions for asymptotic efficiency of the estimator determined by (10), where $\lambda_n^{(T)}$ is given by (16).

We recently came across a related theorem by Baldi and Pacchiarotti [2004]. An important difference is that these authors study the continuous-time buffer-content probability. We wish to remark, however, that our method can be extended to cover continuous time by applying standard theorems for large deviations of Gaussian

measures on Banach spaces, see for instance Deuschel and Stroock [1989] or Dieker [2005]. However, we believe that discrete time is more natural in a simulation framework; see also Section 6. Another difference is the proof technique; Baldi and Pacchiarotti [2004] use recent insights into certain Gaussian martingales, while we take a direct approach.

THEOREM 4.3. *Importance sampling under a 'single exponential twist' is asymptotically efficient for simulating p_n^T if and only if*

$$\inf_{t\in\{1,\ldots,T\}} \frac{b+(c-\mu)t+x_t^*}{\sigma(t)} = 2\,\frac{b+(c-\mu)t^*}{\sigma(t^*)}. \tag{17}$$

Clearly,

$$h_{t^*} = 2\,\frac{b+(c-\mu)t^*}{\sigma(t^*)}, \quad \text{where} \quad h_t := \frac{b+(c-\mu)t+x_t^*}{\sigma(t)};$$

hence Theorem 4.3 states that the change of measure is asymptotically efficient if and only if $h_t \geq h_{t^*}$ for all $t \in \{1,\ldots,T\}$.

In the above we represented time by the natural numbers $\mathbb{N}$, i.e., we used a grid with mesh 1. The same techniques can be used to prove a similar statement for any arbitrary simulation grid. In the following intermezzo, we analyze the impact of making the grid more fine-meshed.

Intermezzo: refining the simulation grid. Consider simulation on the grid $m\mathbb{N}\cap[0,T]$ for some grid mesh $m>0$. One can repeat the analysis in the proof of Theorem 4.3 (see the appendix) to deduce that the infimum in (17) should then be taken over $m\mathbb{N}\cap[0,T]$. Thus, by refining the grid, the left hand side of (17) can be made arbitrarily close to the infimum over $[0,T]$. This motivates an analysis of the function $g:\mathbb{R}_+\to\mathbb{R}_+$ given by $g(t) := [b+(c-\mu)t+\bar{x}^*(t)]/\sigma(t)$, where $\bar{x}^*$ denotes the continuous-time analogue of (15):

$$\bar{x}^*(t) = \frac{b+(c-\mu)t^*}{2\sigma^2(t^*)}\left[\sigma^2(t^*)+\sigma^2(t)-\sigma^2(|t-t^*|)\right].$$

Hence, there is asymptotic optimality for any grid on $[0,T]$ if and only if $g(t)\geq g(t^*)$ for all $t\in[0,T]$. Suppose that σ^2 is twice continuously differentiable with first and second derivative denoted by $\dot{\sigma}^2$ and $\ddot{\sigma}^2$ respectively. Necessary conditions for $\inf_{t\in[0,T]} g(t) \geq g(t^*)$ are then $\dot{g}(t^*)=0$ and $\ddot{g}(t^*)>0$. Therefore we compute

$$\lim_{t\uparrow t^*}\dot{g}(t) = \frac{1}{2}\frac{b+(c-\mu)t^*}{\sigma^3(t^*)}\dot{\sigma}^2(0),$$

so that $\dot{\sigma}^2(0)>0$ implies that exponential twisting becomes asymptotically *inefficient* as the grid mesh m tends to zero. For the complementary case $\dot{\sigma}^2(0)=0$, we can certainly find an 'inefficient' grid mesh if $\lim_{t\uparrow t^*}\ddot{g}(t)<0$. After some calculations, one obtains

$$\lim_{t\uparrow t^*}\ddot{g}(t) = \frac{1}{4}\frac{b+(c-\mu)t^*}{\sigma^3(t^*)}\left[\frac{[\dot{\sigma}^2(t^*)]^2}{\sigma^2(t^*)}-\ddot{\sigma}^2(t^*)-\ddot{\sigma}^2(0)\right], \tag{18}$$

which is negative if $[\dot{\sigma}^2(t^*)]^2 < \sigma^2(t^*)[\ddot{\sigma}^2(t^*) + \ddot{\sigma}^2(0)]$.

Having these conditions at our disposal, we can study some specific cases and ask whether the single exponential twist becomes inefficient as the mesh tends to zero. For instance, suppose that the input traffic $A_1(t)$ is a fractional Brownian motion (fBm) with Hurst parameter $H \in (0,1)$, i.e., $\sigma^2(t) = t^{2H}$. Note that a special case is Brownian motion, which corresponds to $H = 1/2$. If $H \leq 1/2$, one has $\dot{\sigma}^2(0) > 0$ and a single exponential twist is therefore asymptotically inefficient for grid meshes m small enough, in line with the results of Baldi and Pacchiarotti [2004]. Moreover, if $H > 1/2$, it follows from (18) and $\ddot{\sigma}^2(0) = \infty$ that $\lim_{t\uparrow t^*} \ddot{g}(t) < 0$, so that we here have inefficiency as well.

From the above we also see that it could be that the exponential twist is asymptotically optimal for some grid mesh m, but loses the optimality at some finer threshold grid mesh m^*.

Intuition behind (in-)efficiency of exponential twist. Having seen that a single exponential twist can be asymptotically inefficient, one may wonder *why* this occurs. To this end, consider the likelihood term $d\nu_n^{(T)}/d\lambda_n^{(T)}$ following from (16):

$$\exp\left(-n\frac{b+(c-\mu)t^*}{\sigma^2(t^*)}\,x_{t^*} + nI_{t^*}\right) = \exp\left(-n\frac{b+(c-\mu)t^*}{\sigma^2(t^*)}\,(x_{t^*} - x^*_{t^*}) - nI_{t^*}\right),$$

where x_{t^*} corresponds to the value of $A_n(t^*)/n - \mu t^*$, with mean $x^*_{t^*} = b + (c-\mu)t^*$ under $\lambda_n^{(T)}$. For asymptotic optimality, this likelihood ratio should be 'small' for realizations in the set $\mathcal{O}_T$. If there is exceedance at time t^*, then clearly

$$\frac{d\nu_n^{(T)}}{d\lambda_n^{(T)}} \leq e^{-nI_{t^*}} \tag{19}$$

(use $x_{t^*} \geq b + (c-\mu)t^*$). However, if exceedance occurs at any other time epoch, the likelihood ratio can take any (positive) value. Obviously, an extremely high value has a dramatic effect on the variance of the estimator, but the probability of such an extreme value might be low. Summarizing, condition (17) gives a criterion to check whether high values for the likelihood are probable enough to affect (the exponential decay of) the variance of the estimator.

4.2 The cut-and-twist method

In Section 4.1 we have seen that the likelihood may explode while simulating p_n^T with a single exponential twist. This can be overcome by partitioning the event $\mathcal{O}_T$ into disjoint sub-events, and simulating these individually. To this end, write

$$p_n^T = \nu_n^{(T)}\left(\bigcup_{t\in\{1,\ldots,T\}} \mathcal{O}_T(t)\right) = \sum_{t\in\{1,\ldots,T\}} \nu_n^{(T)}(\mathcal{O}_T(t)),$$

where $\mathcal{O}_T(t)$ corresponds to the event that exceedance occurs *for the first time* at time t:

$$\mathcal{O}_T(t) := \{x \in \mathbb{R}^T : x_t + \mu t \geq b + ct;\, \forall s \in \{1,\ldots,t-1\} : x_s + \mu s < b + cs\}.$$

Hence, the problem reduces to simulating T probabilities of the type $\nu_n^{(T)}(\mathcal{O}_T(t))$. This partitioning approach is also taken by Boots and Mandjes [2002], where this idea was exploited for a queue fed by (discrete-time) on-off sources.

The resulting simulation algorithm, to be called cut-and-twist method, works as follows. Define the exponentially twisted measure ${}^t\lambda_n^{(T)}$ as in (16), but with t instead of t^*, and estimate the probability $\nu_n^{(T)}(\mathcal{O}_T(t))$ with the importance-sampling distribution ${}^t\lambda_n^{(T)}$. An estimate of p_n^T is found by summing the estimates over $t \in \{1, \ldots, T\}$.

Before discussing the efficiency of this method, we note for the sake of clarity that the estimator equals

$$\frac{1}{N}\sum_{k=1}^{N}\sum_{t\in\{1,\ldots,T\}} \mathbf{1}_{\left\{X_t^{(k)}\in\mathcal{O}_T(t)\right\}}\frac{d\nu_n^{(T)}}{d\lambda_n^{(T)}}\left(X_t^{(k)}\right), \tag{20}$$

where $X_t^{(1)}, \ldots, X_t^{(N)}$ is an i.i.d. sample from ${}^t\lambda_n^{(T)}$, and the samples $X_t^{(\cdot)}$, $t = 1, \ldots, T$ are also independent.

The following theorem is proven in Appendix A.3. Its proof is based on the property that the method is such that, when estimating $\nu_n^{(T)}(\mathcal{O}_T(t))$, for any $x \in \mathcal{O}_T(t)$ the corresponding likelihood is uniformly bounded by $\exp(-nI_t)$, cf. (19).

THEOREM 4.4. *The cut-and-twist method is asymptotically efficient for estimating* p_n^T.

This method is asymptotically optimal, but it has the obvious drawback that it may take a substantial amount of time to simulate the T probabilities individually.

Summarizing, in this approach the exceedance event is split into disjoint events that correspond to exceedance (for the first time) at time t. The main advantage of this splitting is that every of these individual events can be 'controlled' now (the corresponding likelihoods are even bounded, see the proof of Theorem 4.4), whereas the single-twist method suffers from the (potentially large) likelihoods that correspond to exceedance at time epochs different from the most likely time t^*, as was noted in Section 4.1. As a result, the single-twist method is *not* necessarily asymptotically efficient, while cut-and-twist *is*.

4.3 The random-twist method

An approach closely related to the cut-and-twist method was proposed by Sadowsky and Bucklew [1990]. In the method of Sadowsky and Bucklew [1990], a *random* $\mathcal{T}$ is drawn in each simulation run according to some (arbitrary) distribution $Q = \{q_t : t = 1, \ldots, T\}$ with q_t *strictly positive* for any $t \in \{1, \ldots, T\}$; subsequently, one does a simulation run under the measure ${}^{\mathcal{T}}\lambda_n^{(T)}$ (as defined in Section 4.2).

The likelihood ratio becomes

$$\left[\sum_{t=1}^{T} q_t \exp\left(n\frac{b+(c-\mu)t}{\sigma^2(t)}\,x_t - nI_t\right)\right]^{-1}.$$

Note that this likelihood ratio depends on the whole path $x_1, \ldots, x_T$, as opposed to the previous two methods.

The following result follows from Theorem 2(a) of Sadowsky and Bucklew [1990].

THEOREM 4.5 (SADOWSKY-BUCKLEW). *The random-twist method is asymptotically efficient for estimating* p_n^T.

Remarkably, the asymptotic efficiency does not depend on the specific choice of the q_t, as long as they are strictly positive. Hence, a drawback of the method is that it is unclear how the distribution Q is best chosen. For instance, if Q is 'almost' degenerate in t^*, then the method is similar to the single-twist method; therefore, it may suffer in practice from the same problems as discussed in Section 4.2. The theorem indicates that this effect eventually vanishes (when n grows large), but this could require extremely large n.

It is not the aim of this paper to investigate the impact of the choice of Q on the quality of the estimates; in the sequel, we suppose that the q_t correspond to a truncated Poisson distribution with mean t^*, i.e.,

$$q_t = \frac{(t^*)^t/t!}{\sum_{k=1}^T (t^*)^k/k!}, \quad t = 1, \ldots, T.$$

The reason for this choice is that the Poisson distribution is nicely spread around its mean value. In addition, it is straightforward to sample from a Poisson distribution, so that one can sample from Q with a simple acceptance-rejection procedure.

4.4 The sequential-twist method

Recently, Dupuis and Wang [2004] introduced an intuitively appealing approach to rare-event simulation. We now give a brief description of their method in the setting of the present paper, although the method is known to work in a considerably more general setting. Consider a sequence $\bar{A}_1, \bar{A}_2, \ldots$ of centered i.i.d. random vectors in $\mathbb{R}^T$, where the $\bar{A}_j$ are distributed as $\{A_1(t) - \mu t : t = 1, \ldots, T\}$; as a consequence, the vectors $\bar{A}_j$ have distribution $\nu_1^{(T)}$. Note that p_n^T can be written as

$$P\left(\frac{1}{n}\sum_{i=1}^n \bar{A}_i \in \mathcal{O}_T\right),$$

with $\mathcal{O}_T$ defined in (13), and hence

$$p_n^T = \int_{\{(x^{(1)},\ldots,x^{(n)}):\frac{1}{n}\sum_{i=1}^n x^{(i)} \in \mathcal{O}_T\}} \nu_1^{(T)}(dx^{(1)}) \cdots \nu_1^{(T)}(dx^{(n)}). \tag{21}$$

Instead of twisting ν_n^T as in the previous methods, the sequential-twist method twists *each copy* of $\nu_1^{(T)}$ (i.e., each source) in Equation (21) differently, exploiting the fact that the sources behave stochastically independently. Recall that exponential twisting for Gaussian random variables corresponds to shifts in the mean (and no change in the covariance structure).

This gives rise to the following sequential approach. Suppose $\bar{A}_1, \ldots, \bar{A}_j$ (i.e., source 1 up to j) are already generated, and we are about to twist the traffic generated by source $j+1$ (for $j = 0, \ldots, n-1$). We aim to find the 'cheapest' way to reach the exceedance set $\mathcal{O}_T$ *given* $\bar{A}_1, \ldots, \bar{A}_j$. Hence, we do not change the measure if already $\frac{1}{n}\sum_{i=1}^{j} \bar{A}_i \in \mathcal{O}_T$ (under this condition reaching $\mathcal{O}_T$ is not 'hard' anymore, as $\mathbb{E}\bar{A}_j(t) = 0$); otherwise we change the mean of the distribution of $\bar{A}_{j+1}$ to μ_{j+1} (recall this is a vector in $\mathbb{R}^T$), where

$$\mu_{j+1} := \underset{\{y \in \mathbb{R}^T : \frac{1}{n}\sum_{i=1}^{j} \bar{A}_i + \frac{1}{n}\sum_{i=j+1}^{n} y \in \mathcal{O}_T\}}{\arg\inf} y' \left(\Gamma^{(T)}\right)^{-1} y;$$

here an empty sum is interpreted as zero. The following lemma gives a useful explicit expression for μ_{j+1}. The proof is given in Appendix A.4.

LEMMA 4.6. *Define for* $j = 0, \ldots, n-1$,

$$t^*_{j+1} := \underset{t \in \{1,\ldots,T\}}{\arg\inf} \frac{nb + n(c-\mu)t - \sum_{i=1}^{j} \bar{A}_i(t)}{(n-j)\sigma(t)}, \tag{22}$$

and denote the corresponding infimum by J_{j+1}. *Then we have*

$$\mu_{j+1} = \frac{J_{j+1}}{\sigma(t^*_{j+1})} \Gamma(\cdot, t^*_{j+1}).$$

Observe that for $j = 0$ the formula reduces to the large-deviation most probable path, which is to be expected since then no information is available on the previously generated sources. The reader may check that the resulting likelihood ratio is

$$\prod_{j=1}^{n} \exp\left(-\frac{J_j}{\sigma(t^*_j)} \bar{A}_j(t^*_j) + \frac{1}{2} J_j^2 \right).$$

An estimator is obtained by performing N independent runs, and computing the estimate using (9).

The conditions for the following theorem of Dupuis and Wang [2004] are checked in Appendix A.4.

THEOREM 4.7 (DUPUIS-WANG). *The sequential-twist method is asymptotically efficient for estimating* p_n^T.

Informally speaking, the idea behind the sequential-twist approach is that, by adapting the mean μ_j of every next source j in the way described above, the set $\mathcal{O}_T$ is reached close to its most likely point, thus avoiding large likelihood ratios. Apparently, as claimed in the above theorem, the resulting estimator is asymptotically optimal.

A drawback of this approach is that all sources should be generated individually; one does not simulate the aggregate input process, as in the previous methods. However, the sequential approach can also be used with less than n Gaussian vectors while retaining the property of asymptotic efficiency. This is done by twisting

source *batches* instead of individual sources. Let M be a *batch size* such that $n/M \in \mathbb{N}$, where M does not depend on n. Define $\bar{A}_i^{(M)} := \frac{1}{M}\sum_{j=1}^{M} \bar{A}_{j+(i-1)M}$. It is important that M does not depend on n. We refer to this approach as the *batch sequential-twist method*; since

$$P\left(\frac{1}{n}\sum_{i=1}^{n} \bar{A}_i \in \mathcal{O}_T\right) = P\left(\frac{1}{n/M}\sum_{i=1}^{n/M} \bar{A}_i^{(M)} \in \mathcal{O}_T\right),$$

Theorem 4.7 also yields the asymptotic efficiency of the batch sequential-twist estimator for any fixed M.

Although the sequential-twist method and its batch counterpart are both asymptotically efficient, some practical issues arise when M is made (too) large. The relative efficiency then converges much slower to 2, so that we might not even be close to efficiency for reasonable n. This issue is addressed empirically in Section 5.4.

5. EVALUATION

In this section, we evaluate the four methods of Section 4 as follows. First, we discuss some issues related to our implementation of the methods. Based on this, we come to preliminary conclusions on the time complexity of each of the methods. In Section 5.2, we check empirically that our simulations support the claims of Theorems 4.3, 4.4, 4.5, and 4.7. After this analysis, the reliability of the methods is studied by refining the simulation grid; for this, we also take the computational effort into account. Further empirical insight into the batch sequential-twist method is gained by studying the influence of the batch size on the relative efficiency and the relative error.

While the preceding sections are applicable to general Gaussian processes with stationary increments (satisfying certain conditions), in this section we focus on the important case of fractional Brownian motion.

5.1 Implementation and time complexity

Simulation of fractional Brownian motion is highly nontrivial. As the simulation grid is equispaced, it is best to simulate the (stationary!) incremental process, often called fractional Gaussian noise. When T is a power of two, the fastest available algorithm for simulating T points of a fractional Gaussian noise is the method of Davies and Harte [1987]. In this approach, the covariance matrix is embedded in a so-called circulant matrix, for which the eigenvalues can easily be computed. The Fast Fourier Transform (FFT) is then used for maximum efficiency; the computational effort is of order $T \log T$ for a sample size of length T. For more details on this method, we refer to Dietrich and Newsam [1997] and Wood and Chan [1994].

Although we use this fast algorithm, the required number of traces per simulation run still highly influences the speed of the methods. The single-twist method and the random-twist method only need one trace (of length T) for each simulation

run, while n such traces are needed for the sequential-twist method. For the cut-and-twist method, traces of length $t = 1, \ldots, T$ are needed. These considerations indicate that it depends on the parameter values which method performs best.

We first address the impact of the number of sources n. A clear advantage of the cut-and-twist method and the random-twist method is that the required simulation time depends just mildly on n (due to the fact that $T(n) \to T$), as opposed to sequential-twist (where the simulation time is roughly proportional to n).

As for the influence on the simulation horizon, we have already observed that T is large when either $b/(c-\mu)$ or H is large, see Section 2.3. This badly affects the cut-and-twist method, since such a sample is needed for each time epoch (of which there are T). The random-twist method only needs a single fBm trace, but the computation of the likelihood ratio is of the order T. Sequential-twist calculates the best twist n times, which amounts to computing the infimum in (22); this computation is of order T.

We raise one further implementation issue, which plays a role for all of the methods. Once we have calculated the simulation horizon T, we round it off to the smallest power of two T' with $T' \geq T$, *and we use this new horizon* T'. Similarly, since traces of length $t = 1, \ldots, T'$ are needed for the cut-and-twist method, t is rounded off, for every t.

5.2 Empirical validation of the theory

In Section 4, we studied whether the four discussed simulation methods are asymptotically efficient. In the present subsection, our aim is to validate these theoretical results by performing simulation experiments. By doing so, we gain insight into the quality of the methods.

The parameters are chosen as follows: $b = 0.3$, $c - \mu = 0.1$, $H = 0.8$, $M = 1$, $\epsilon = 0.05$, and $\eta_{\max} = 0.1/1.96$. Recall from Section 3.2 that the simulation is stopped when the relative error drops below $\eta_{\max}$. It is left to the reader to check that condition (17) does not hold, i.e., that the single-twist estimator is not asymptotically efficient. The choice $H = 0.8$ is supported by several measurement studies, see for instance Leland *et al.* [1994], and $\eta_{\max}$ is chosen such that the width of the confidence interval is 20% of the estimated probability. Reduction of this value has a significant impact on the simulation time, and the present value yields typical results within a reasonable time frame.

We first study the asymptotic efficiency of the simulation methods by varying n and analyzing the number of simulation runs N_n^* needed to achieve the required relative error. In the left panel of Figure 2, we have plotted $\log N_n^*$ for $n = 100, 150, \ldots, 500$ and all four simulation methods, and in addition the 'naive' direct Monte Carlo estimator. The confidence intervals are not plotted, since they are completely determined by the estimates themselves and the value of $\eta_{\max}$. Note that under asymptotic efficiency, $\log N_n^*$ should be (ultimately) sublinear. Therefore, the plot supports Theorem 4.4, Theorem 4.7, and the fact that the naive

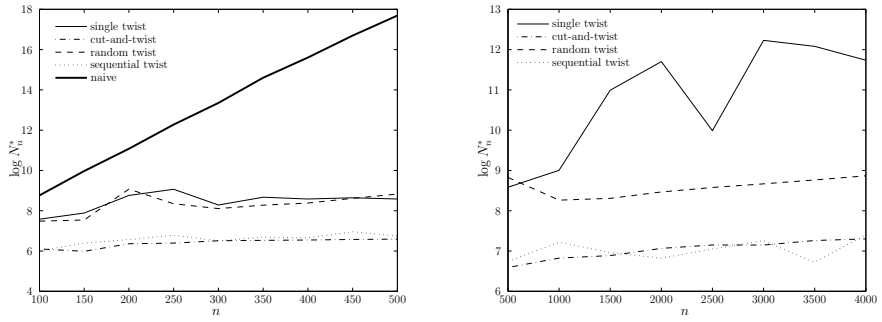

Fig. 2. Empirical verification of the asymptotic efficiency of the simulation methods.

	$n = 300$	rel. eff.	$n = 1000$	rel. eff.
naive	6.12×10^{-4}	—	—	—
single twist	4.84×10^{-4}	1.68	1.03×10^{-10}	1.87
cut-and-twist	5.95×10^{-4}	1.86	1.32×10^{-10}	1.94
random twist	5.50×10^{-4}	1.70	1.38×10^{-10}	1.89
sequential twist	6.39×10^{-4}	1.86	1.41×10^{-10}	1.93
'exact'	5.8×10^{-4}		1.38×10^{-10}	

Table I. Two of the estimates corresponding to Figure 2.

estimator is inefficient (in fact, the number of runs grows exponentially, in line with p_n decaying exponentially). However, it is not immediate from the left panel of Figure 2 that single twist is asymptotically inefficient (cf. Theorem 4.3), and that random twist is asymptotically efficient (cf. Theorem 4.5). Although the irregular behavior indicates that this might indeed be the case, we find more convincing evidence by increasing n further. This is done in the right panel of Figure 2.

It is interesting to see some of the estimated probabilities that correspond to Figure 2. We give these for two different values of n in Table I. To obtain a benchmark, we also performed a very long simulation, see the row labeled 'exact'. It was obtained with the cut-and-twist method, where the simulation is stopped as soon as both ends of the confidence interval give the same value when rounded off to one digit for $n = 300$, and to two digits for $n = 1000$.

The unstable behavior of the single-twist method (also reflected in a low value of the relative efficiency in Table I) has been explained theoretically through the interpretation of a possible failure of the exponential twist, see Section 4. As noted there, the supremum is attained at time epoch t^* in a 'typical' simulation run, but it might also happen at some other epoch $t \neq t^*$. Although such a realization is (relatively) rare, it has an impact on both the estimate and the estimated variance. Since these two estimated quantities determine whether the simulation is stopped, it may occur that the number of these 'rare' realizations is too low, so that the

simulation is stopped too early and the buffer-content probability is underestimated. Likewise, random twist can lead to underestimation, but this effect vanishes when n grows large, in line with the theory of Section 4.3. The table shows that random twist has a low relative efficiency for small n.

It is interesting to see that the sequential-twist method and cut-and-twist method have comparable performance, both in terms of relative efficiency (Table I) and number of simulation runs (Figure 2).

Hence, cut-and-twist and sequential twist seem to perform best, although the random-twist method improves considerably as n grows. However, these methods are also the slowest (i.e., the effort per experiment is highest). To obtain a more realistic comparison, one should consider CPU time, rather than the number of experiments. This is done in the next subsection.

5.3 Simulation grid

While the observations in the previous subsection were predicted by theory, we now perform experiments that relate to the CPU time needed, for which no theory is available. This analysis provides further insight into the performance of the methods in practice, both in terms of reliability and speed. As a first step, we investigate the influence of the grid mesh on the estimated probability.

We evaluate, with again $\bar{A}_n \in \mathbb{R}^T$ denoting the centered version of A_n,

$$ {}^{\alpha}p_n := P\left(\sup_{t \in \{\alpha, 2\alpha, \ldots\}} \bar{A}_n(t) - n(c-\mu)t > nb \right) \tag{23}$$

for a range of $\alpha \geq 0$, in such a way that the simulation grid becomes finer. For instance, one can take $\alpha = 1, 1/2, 1/4, 1/8$; ${}^{\alpha}p_n$ then increases as α is made smaller, as we only add grid points, and hence the supremum of the free process becomes larger. Therefore, as a sanity check, we can test the reliability of the simulation methods by checking whether the estimates indeed increase when making the grid finer.

Before we can compare the estimated probabilities for different α, we first study the impact of α on the simulation horizon. We denote this simulation horizon, as a function of α, by T^{α}. We now verify whether also ${}^{\alpha}p_n^{T^{\alpha}}$ (defined in a self-evident manner) should increase when decreasing α. Since $\bar{A}_n(t)$ is a centered fractional Brownian motion by assumption, one readily verifies that $\bar{A}_n(\alpha t)$ has the same distribution as $\alpha^H \bar{A}_n(t)$. This so-called self-similarity property now yields that (23) equals

$$P\left(\sup_{t \in \mathbb{N}} \alpha^H \bar{A}_n(t) - n\alpha(c-\mu)t > nb \right) = P\left(\sup_{t \in \mathbb{N}} \bar{A}_n(t) - n\alpha^{1-H}(c-\mu)t > n\alpha^{-H}b \right).$$

The above equation entails that a grid mesh α is equivalent to a unit grid mesh if

b and $c-\mu$ are replaced by $b_\alpha := \alpha^{-H} b$ and $c_\alpha := \alpha^{1-H}(c-\mu)$. Note that then

$$I^\alpha_{t^*} := \inf_{t\in\{\alpha,2\alpha,\ldots\}} \frac{(b+(c-\mu)t)^2}{2t^{2H}} = \inf_{t\in\mathbb{N}} \frac{(b+\alpha(c-\mu)t)^2}{2\alpha^{2H}t^{2H}},$$

so that the (limiting) simulation horizon then becomes (see (8))

$$T^\alpha = \frac{I^\alpha_{t^*}}{c_\alpha^2/2} = \inf_{t\in\mathbb{N}} \frac{(b/\alpha+(c-\mu)t)^2}{c^2t^{2H}},$$

which is monotonic in α and tends to infinity as $\alpha \downarrow 0$. We conclude that the monotonicity is preserved: ${}^\alpha p_n^{T^\alpha}$ increases when $\alpha \downarrow 0$, just like ${}^\alpha p_n$ does.

In order to investigate whether the estimates indeed decrease in α, we perform some simulations with parameters $n = 150$, $b = 0.9$, $c-\mu = 0.3$, $H = 0.8$, $M = 1$, and $\epsilon = 0.05$. To obtain each of the estimates, we stop the simulation after exactly five minutes of CPU time. It would be desirable to do the simulations for grid sizes $2^0, 2^1, 2^2, 2^3, 2^4, \ldots$, but this quickly becomes computationally too intensive. Therefore, we focus on four sets of grids; $1/\alpha = 1, 2, 4, 8$, $1/\alpha = 3, 6, 12$, $1/\alpha = 3, 9$, and $1/\alpha = 5, 10$.

In Figure 3, we have plotted these four sets using the four different methods. The dotted lines correspond to the boundaries of the confidence intervals. Roughly speaking, each of the plots shows the expected monotonicity, with the only exception of the single-twist method. This is in line with Theorem 4.3. The behavior of the single-twist confidence intervals also differs from the other methods, but it is interesting to compare these intervals for the other three (efficient) methods.

The widths of the confidence intervals do not seem to grow proportionally to the estimates, most prominently for $1/\alpha = 12$. Although the stopping criterion (CPU time) is proportional to the number of runs, the CPU time per run varies, since the time horizon T depends on α. For instance, fBm traces of length 1024 are generated if $\alpha = 1/10$, while this increases to 2048 for $\alpha = 1/12$. This is reflected in the plots (especially in the cut-and-twist plot, as anticipated in Section 5.1). Clearly, the random-twist confidence intervals are the smallest, but we have to keep in mind that the probabilities may be underestimated, in view of the previous subsection.

5.4 Batch size for the sequential-twist method

The aim of the present subsection is to investigate the influence of the parameter M in the batch sequential-twist method. That is, the n sources are divided into batches of size M and each of the batches is considered a single source (using the fact that the sum of independent Gaussian vectors is again Gaussian). The advantage over the 'normal' sequential-twist method is that one needs to sample just n/M sources per simulation run, rather than n. On the other hand, however, this 'lumping procedure' limits the flexibility of the method (due to the fact that the probability measure is adapted less often).

If we let the simulation run for a specified time period (here five minutes), there are two effects as M increases. First, the relative efficiency decreases; in analogy

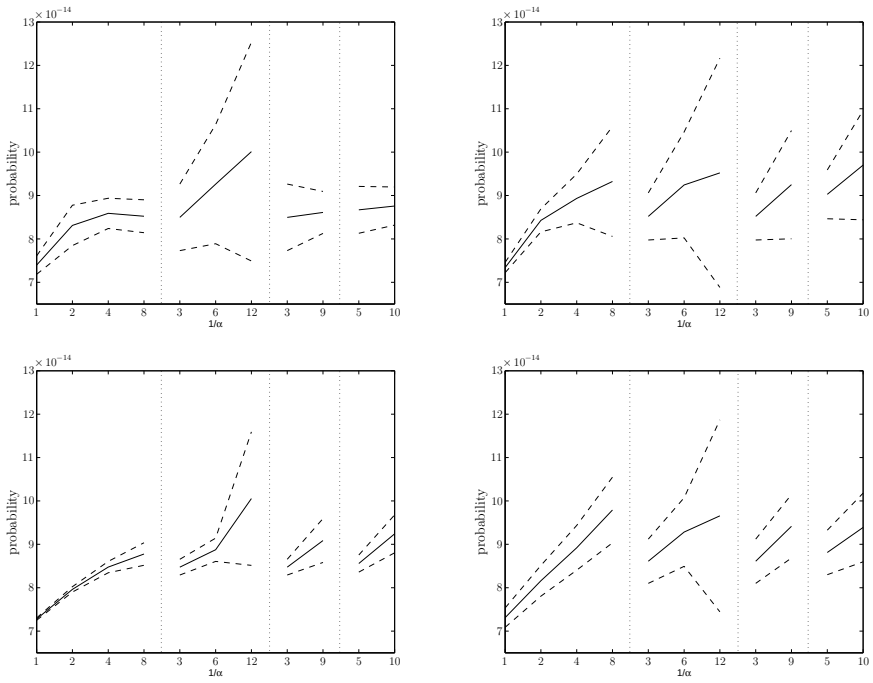

Fig. 3. The influence of the grid mesh on the probability for single twist, cut-and-twist, random twist, and sequential twist. The solid lines represent the estimates, while the dashed lines correspond to confidence intervals.

with the single-twist method, this may cause underestimation of the probability of interest. In the second place, we observe from our experiments that the relative error decreases. It is the aim of this subsection to study these two opposite effects. The values of the parameters are the same as in Section 5.2, except for the value of M, which now varies.

We measure efficiency by means of the (estimated) relative efficiency. We set $n = 3840$ and estimate the relative efficiency for $M = 2, 4, 6, 8, 10, 12$. The resulting plot is given in Figure 4. From the plot, it is not so clear that an increase in M makes the simulation less efficient, although the relative error seems to decrease. Therefore, we also investigate what happens if $M = 80, 160, 240, 320, 480, 640, 960$; the relative efficiency (relative error) is then estimated as 1.972 (0.0164), 1.969 (0.0135), 1.966 (0.0125), 1.960 (0.0140), 1.956 (0.0139), 1.953 (0.0137), and 1.950 (0.0127) respectively. These values indeed indicate that the simulation becomes less efficient as M increases, while the relative error decreases.

Although the differences in the relative efficiency look small, one must keep in mind that this quantity relates to the *exponential* decay rate of the variance of the estimator. Therefore, small differences blow up exponentially, and we propose to choose M small to control the risk of underestimation.

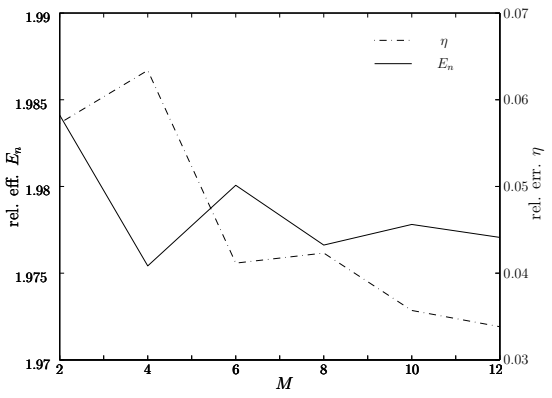

Fig. 4. The relative efficiency as a function of M for small M.

5.5 Concluding remarks on simulation methods

We end this section by giving a number of general conclusions on the presented simulation methods for estimating the buffer-content probability p_n.

—The single-twist method should not be used. It usually does not estimate p_n asymptotically efficiently, which makes the method slow or even unreliable (as it is not guaranteed that the variance of the corresponding estimator is finite).

—In our experiments the cut-and-twist method and the sequential-twist method performed roughly equally well, both in terms of relative efficiency and number of simulation runs. In both methods, there is no risk of underestimating the probability of interest. In order to choose between the sequential-twist method and the cut-and-twist method, the former is to be preferred if the time horizon T is excessively large, whereas the latter should be chosen when the number of sources is extremely large. We remark that T tends to be large when the system is heavily loaded (i.e., c being just slightly larger than μ), and, in the case of fBm, when H is close to 1 (recall Figure 1). It is not straightforward, however, to describe the trade-off in more detail, as it depends strongly on all parameters involved.

—The cut-and-twist and random-twist methods control the likelihood in the most explicit way. In fact, the proofs of their asymptotic efficiency reveal that the methods lead to a bounded relative error. Apart from its asymptotic efficiency, we do not know further optimality properties of the sequential-twist method.

—The random-twist method is usually faster than the cut-and-twist method and the sequential-twist method, but may suffer from underestimation. This effect is of course minimal when the contributions to p_n are concentrated in the immediate neighborhood of t^*; bounds on these contributions can be found with techniques in the same spirit as those used in Section 2.3.

—The relative error of the batch sequential-twist method decreases with the batch size. The relative efficiency, however, decreases as well, indicating an increasing risk of underestimation. We therefore advise that the batch size be chosen relatively small.

6. DISCUSSION

In this section, we stress two issues related to the findings of the present paper. First, we explain why the buffer-content probability in discrete time does not necessarily yield a good approximation for its continuous-time counterpart. We also make some remarks on the main assumption underlying our analysis: the Gaussianity of the sources.

Discrete time vs. continuous time. It is important to realize that the probability (1) behaves qualitatively different in continuous time, i.e., when $\mathbb{N}$ is replaced by $\mathbb{R}_+$. We illustrate this by recalling the asymptotics of (1) in both discrete and continuous time. Denote the probability in continuous time by $p_n^{\mathbb{R}_+}$.

In discrete time, there exists a constant $\mathcal{K}$ such that [Likhanov and Mazumdar 1999]

$$p_n \sim \frac{\mathcal{K}}{\sqrt{n}} \exp\left(-\frac{1}{2} n \frac{(b+(c-\mu)t^*)^2}{\sigma^2(t^*)}\right),$$

where t^* minimizes I_t over $\mathbb{N}$. However, in continuous time the asymptotics depend on the behavior of σ near zero. If $\sigma(t) \sim Ct^\gamma$ as $t \to 0$ for constants $C \in (0,\infty)$ and $\gamma \in (0,2)$, then, under suitable regularity assumptions, [Dębicki and Mandjes 2003]

$$p_n^{\mathbb{R}_+} \sim \mathcal{K}' n^{\frac{1}{\gamma}-1} \exp\left(-\frac{1}{2} n \frac{(b+(c-\mu)t^*)^2}{\sigma^2(t^*)}\right),$$

with t^* minimizer of I_t over $\mathbb{R}_+$, and for some constant $\mathcal{K}'$ (which involves the so-called *Pickands' constant* for which no explicit representation is available). Conclude that the polynomial term in the above asymptotic expansions is different. To our knowledge, reliable simulation methods for the continuous-time probability $p_n^{\mathbb{R}_+}$ do not exist.

Gaussian input. As pointed out in the Introduction, the study of a queue fed by Gaussian sources is often motivated by (central) limit theorems. The buffer-content probability of a Gaussian model may be a good approximation of the 'real' buffer-content probability, although in reality evidently network traffic cannot be Gaussian (as the Gaussian model allows negative traffic). It is important to realize that the accuracy of the approximation critically depends on the appropriateness of the imposed *scaling*. Therefore, this should first be studied before resorting to a Gaussian model; see the paper by Wischik [2001a] for a detailed discussion.

A. APPENDIX: PROOFS

In this appendix, we provide proofs of the assertions in this paper. We start in Appendix A.1 with the proofs related to the simulation horizon T, which apply to all methods discussed in Section 4. Appendices A.2 and A.3 deal with the single-twist method and cut-and-twist method respectively. The proof of Lemma 4.6 is given in Appendix A.4.

A.1 Upper bounds on $\int_T^\infty e^{-nC_0t^{1/q}}dt$

We distinguish the cases $q \le 1$ (Lemma 2.1) and $q > 1$ (Lemma 2.2).

A.1.1 *Proof of Lemma 2.1.* Since $q \le 1$ and $T \in \mathbb{N}$, we can bound the left hand side of (6) as follows:

$$\begin{aligned}\int_T^\infty \exp\left(-nC_0t^{1/q}\right)dt &= \frac{q}{C_0^q}\int_{C_0T^{1/q}}^\infty \exp(-ny)y^{q-1}dy\\ &\le \frac{q}{C_0^q}\left(C_0T^{1/q}\right)^{q-1}\int_{C_0T^{1/q}}^\infty \exp(-ny)dy\\ &= \frac{q}{C_0^q n}\left(C_0T^{1/q}\right)^{q-1}\exp\left(-nC_0T^{1/q}\right)\\ &\le \frac{q}{C_0 n}\exp\left(-nC_0T^{1/q}\right),\end{aligned}$$

as claimed.

A.1.2 *Proof of Lemma 2.2.* First note that $q > 1$, which is crucial throughout the proof. Recall that $m \ge 0$ denotes the largest integer such that $q-1-m \in (0,1]$. As before, we have by a simple substitution,

$$\int_T^\infty \exp\left(-nC_0t^{1/q}\right)dt = \frac{q}{C_0^q}\int_{C_0T^{1/q}}^\infty \exp(-ny)y^{q-1}dy. \tag{24}$$

The idea is to select $\beta, \gamma \in (0,\infty)$ such that

$$y^{q-1} \le \beta e^{\gamma y} \tag{25}$$

for all $y \in \mathbb{R}_+$. We now discuss how these parameters can be chosen.

If $q \in (1,2]$ (i.e., $m = 0$), then $p_q : y \mapsto y^{q-1}$ is concave. Since p_q is differentiable at 1 with derivative $q-1$, by Theorem 25.1 of Rockafellar [1970] we have for all $y \in \mathbb{R}_+$,

$$y^{q-1} \le 1 + (q-1)(y-1). \tag{26}$$

Similarly, since $y \mapsto \beta e^{\gamma y}$ is convex and differentiable at 1 with derivative $\beta\gamma e^\gamma$, we have for all $y \in \mathbb{R}_+$,

$$\beta e^{\gamma y} \ge \beta e^\gamma + \beta\gamma e^\gamma(y-1). \tag{27}$$

By comparing (26) to (27), we see that $y^{q-1} \le \beta e^{\gamma y}$ upon choosing $\gamma = q-1$ and $\beta = e^{-\gamma}$.

To find β, γ such that (25) holds for $q \in (m+1, m+2]$ where $m > 0$, the key observation is that this inequality is always satisfied for $y = 0$. Therefore, it suffices to choose β, γ such that the derivative of the left hand side of (25) does not exceed the right hand side. By applying this idea m times, one readily observes that it suffices to require that β, γ satisfy

$$\beta\gamma^m e^{\gamma y} \geq (q-1)\cdots(q-m)y^{q-m-1}.$$

Note that the right hand side of this expression is concave as a function of y since $q-m-1 \in (0,1]$, and that the left hand side is convex as a function of y. Therefore, we are in a similar situation as we were for $m = 0$. In this case, we choose β and γ such that

$$\begin{aligned} \beta\gamma^m e^{\gamma} &= (q-1)\cdots(q-m) \\ \beta\gamma^{m+1} e^{\gamma} &= (q-1)\cdots(q-m)(q-m-1). \end{aligned}$$

Note that β and γ as defined in (7) solve this system of equations uniquely. As before, Theorem 25.1 of Rockafellar [1970] is applied twice to see that for $y \in \mathbb{R}_+$,

$$\begin{aligned} & (q-1)\cdots(q-m)y^{q-m-1} \\ \leq\ & (q-1)\cdots(q-m) + (q-1)\cdots(q-m)(q-m-1)(y-1) \\ =\ & \beta\gamma^m e^{\gamma} + \beta\gamma^{m+1} e^{\gamma}(y-1) \leq \beta\gamma^m e^{\gamma y}. \end{aligned}$$

Now that we have found simple bounds on y^{q-1}, the assertion in the lemma follows upon combining these bounds with (24):

$$\begin{aligned} \int_T^\infty \exp\left(-nC_0 t^{1/q}\right) dt \ &\leq\ \frac{q\beta}{C_0^q} \int_{C_0 T^{1/q}}^\infty \exp(-(n-\gamma)y) dy \\ &=\ \frac{q\beta}{C_0^q (n-\gamma)} \exp\left(-(n-\gamma)C_0 T^{1/q}\right). \end{aligned}$$

A.2 Proofs for the single-twist method

The key ingredient in the proofs of this subsection is a large-deviation principle (LDP) known as Cramér's theorem. Therefore, we start by discussing this theorem in more detail. The proof of Lemma 4.2 is given last.

A.2.1 *Large deviations for multivariate Gaussian distributions.* The analysis in Section 4.1 relies on standard large-deviation techniques. The reader is referred to Dembo and Zeitouni [1998] for a rigorous introduction to the theory, or to Deuschel and Stroock [1989].

Recall that given some $T \in \mathbb{N}$, $\nu_n^{(T)}$ denotes the distribution of the centered process $\{A_n(t)/n - \mu t : t = 1, \ldots, T\}$. The covariance of $\nu_n^{(T)}$ is given by $\Gamma^{(T)}/n$, and this covariance defines an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norm $\|\cdot\|_{\mathcal{H}}$ on $\mathbb{R}^T$ as follows:

$$\langle x, y \rangle_{\mathcal{H}} := x' \left(\Gamma^{(T)}\right)^{-1} y, \quad \|x\|_{\mathcal{H}} := \sqrt{\langle x, x \rangle_{\mathcal{H}}}.$$

This inner product sometimes referred to as *Reproducing Kernel Hilbert Space* inner product or *Cameron-Martin* inner product.

As this paper deals with Gaussian random vectors, we state Cramér's theorem for the special case of Gaussian distributions. The theorem has been generalized to Gaussian measures on abstract spaces by Bahadur and Zabell [1979].

THEOREM A.1 (CRAMÉR). $\{\nu_n^{(T)}\}$ *satisfies the LDP in* $\mathbb{R}^T$ *with rate function* $I : x \to \frac{1}{2}\|x\|_{\mathcal{H}}^2$, *i.e.,*

(i) for any closed set $F \subset \mathbb{R}^T$*:* $\limsup_{n\to\infty} \frac{1}{n} \log \nu_n^{(T)}(F) \leq -\frac{1}{2} \inf_{x\in F} \|x\|_{\mathcal{H}}^2$;

(ii) for any open set $G \subset \mathbb{R}^T$*:* $\liminf_{n\to\infty} \frac{1}{n} \log \nu_n^{(T)}(G) \geq -\frac{1}{2} \inf_{x\in G} \|x\|_{\mathcal{H}}^2$.

PROOF. The proof can be found in Dembo and Zeitouni [1998, Thm. 2.2.30], noting that

$$\sup_{\theta\in\mathbb{R}^T} \left(\langle \theta, x\rangle - \log \int e^{\langle \theta, y\rangle} \nu^{(T)}(dy) \right) = \sup_{\theta\in\mathbb{R}^T} \left(\langle \theta, x\rangle - \frac{1}{2}\theta' \Gamma^{(T)} \theta \right),$$

which equals $\frac{1}{2} x' \left(\Gamma^{(T)}\right)^{-1} x = \frac{1}{2}\|x\|_{\mathcal{H}}^2$.

A.2.2 *Proof of Lemma 4.1.* Lemma 4.1 is an application of Cramér's theorem. We have to prove that

$$\lim_{n\to\infty} \frac{1}{n} \log \nu_n^{(T)}(\mathcal{O}_T) = -\frac{1}{2} \inf_{x\in\mathcal{O}_T} \|x\|_{\mathcal{H}}^2 = -\frac{1}{2}\|x^*\|_{\mathcal{H}}^2. \tag{28}$$

The second equality in (28) is due to Addie *et al.* [2002]. We therefore turn to the first equality. It is readily seen that $\mathcal{O}_T$ is closed in $\mathbb{R}^T$. Cramér's theorem gives an upper bound on the decay rate of $\nu_n^{(T)}(\mathcal{O}_T)$, as well as a lower bound on the decay rate of $\nu_n^{(T)}(\underline{\mathcal{O}}_T)$, where $\underline{\mathcal{O}}_T$ denotes the interior of $\mathcal{O}_T$. The first equality of (28) now follows upon combining these upper and lower bounds with the following lemma (applied for $r = 0$).

LEMMA A.2. *For all* $y \in \mathbb{R}^T$, *we have*

$$\inf_{x\in\underline{\mathcal{O}}_T} \|x+y\|_{\mathcal{H}}^2 = \inf_{x\in\mathcal{O}_T} \|x+y\|_{\mathcal{H}}^2 = \inf_{t\in\{1,\ldots,T\}} \frac{(b+(c-\mu)t+y_t)^2}{2\sigma^2(t)}.$$

PROOF. First note that the interior of the exceedance set is given by

$$\underline{\mathcal{O}}_T := \left\{ x = (x_1, \ldots, x_T) \in \mathbb{R}^T : x_t + \mu t > b + ct \text{ for some } t \in \{1, \ldots, T\} \right\}.$$

Also, evidently,

$$\inf_{x\in\underline{\mathcal{O}}_T} \|x+y\|_{\mathcal{H}}^2 = \inf_{x\in\underline{\mathcal{O}}_{T,y}} \|x\|_{\mathcal{H}}^2,$$

where

$$\underline{\mathcal{O}}_{T,y} := \left\{ x \in \mathbb{R}^T : x_t + \mu t > b + ct + y_t \text{ for some } t \in \{1, \ldots, T\} \right\}.$$

A similar reasoning that led to the second equality in (28) now yields the desired.

A.2.3 *Proof of Theorem 4.3.* As outlined in Section 2.2 of Dieker and Mandjes [2005], it is a consequence of Lemma A.2 (with $y = 0$) that the single exponential twist is asymptotically efficient if and only if

$$\limsup_{n\to\infty} \frac{1}{n} \log \int_{\mathcal{O}_T} \frac{d\lambda_n^{(T)}}{d\nu_n^{(T)}}(x)\lambda_n^{(T)}(dx) \leq -\frac{(b+(c-\mu)t^*)^2}{\sigma^2(t^*)} = -2I_{t^*}, \qquad (29)$$

cf. (12). In principle, the statement can be proven using Theorem 1 of Dieker and Mandjes [2005]. However, the argument can be given directly in this case. We apply Varadhan's Integral Lemma (Theorem 4.3.1 of Dembo and Zeitouni [1998]) to the left hand side of (29). In order to check the conditions for applying this lemma, we note that for $\gamma > 1$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \int_{\mathbb{R}^T} \exp\left(-n\gamma \frac{b+(c-\mu)t^*}{\sigma^2(t^*)} x_{t^*}\right) \nu_n^{(T)}(dx) = \gamma^2 \frac{[b+(c-\mu)t^*]^2}{2\sigma^2(t^*)} < \infty;$$

use that for a zero-mean normal random variable U (with variance σ^2) the moment generating function is $\mathbb{E}\exp(\theta U) = \exp(\theta^2\sigma^2/2)$. Formally, one proceeds by deriving lower and upper bounds for the integral on the left hand side of (29), but, in view of Lemma A.2, the resulting bounds coincide. We may therefore conclude that the lim sup is actually a proper limit; the reader is referred to Section 3 of Dieker and Mandjes [2005] for more details on this reasoning. Application of Varadhan's Lemma gives

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \log \int_{\mathcal{O}_T} \frac{d\nu_n^{(T)}}{d\lambda_n^{(T)}}(x)\nu_n^{(T)}(dx) \\
&= -\inf_{x\in\mathcal{O}_T} \left[\frac{1}{2}\|x\|_{\mathcal{H}}^2 + \frac{b+(c-\mu)t^*}{v(t^*)}x(t^*) - \frac{(b+(c-\mu)t^*)^2}{2v(t^*)}\right] \\
&= -\inf_{x\in\mathcal{O}_T} \left[\frac{1}{2}\|x\|_{\mathcal{H}}^2 + \langle x, x^*\rangle_{\mathcal{H}} - \frac{1}{2}\|x^*\|_{\mathcal{H}}^2\right] = -\left[\frac{1}{2}\inf_{x\in\mathcal{O}_T} \|x+x^*\|_{\mathcal{H}}^2\right] + \|x^*\|_{\mathcal{H}}^2 \\
&= -\frac{1}{2}\inf_{t\in\{1,\ldots,T\}} \frac{(b+(c-\mu)t+x_t^*)^2}{\sigma^2(t)} + \frac{(b+(c-\mu)t^*)^2}{\sigma^2(t^*)},
\end{aligned}$$

where the last equality is due to Lemma A.2. The claim follows by combining this with (29).

A.2.4 *Proof of Lemma 4.2.* Let λ_n^x denote a Gaussian measure on $\mathbb{R}^T$ with mean vector x and covariance $\Gamma^{(T)}/n$. It can be shown along the same lines of the proof of Theorem 4.3 that

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \log \int_{\mathcal{O}_T} \frac{d\nu_n^{(T)}}{d\lambda_n^x}(y)\nu_n^{(T)}(dy) &= -\left[\frac{1}{2}\inf_{y\in\mathcal{O}_T} \|y+x\|_{\mathcal{H}}^2\right] + \|x\|_{\mathcal{H}}^2 \\
&\geq -\frac{1}{2}\|x^*+x\|_{\mathcal{H}}^2 + \|x\|_{\mathcal{H}}^2 \\
&= \frac{1}{2}\|x^*-x\|_{\mathcal{H}}^2 - \|x^*\|_{\mathcal{H}}^2,
\end{aligned}$$

and this is strictly larger than $-\|x^*\|_{\mathcal{H}}^2$ if $x \neq x^*$, contradicting (29).

A.3 Proofs for the cut-and-twist method

In this subsection, we prove Theorem 4.4. Observe that for any $j = 1, 2, \ldots$, by definition of $\mathcal{O}_T(t)$,

$$\begin{aligned}
&\int_{\mathcal{O}_T(t)} \left(\frac{\nu_n^{(T)}}{{}^t\lambda_n^{(T)}}\right)^j d\ {}^t\lambda_n^{(T)} \\
&= \int_{\mathcal{O}_T(t)} \exp\left(nj\frac{(b+(c-\mu)t)^2}{2\sigma^2(t)} - nj\frac{b+(c-\mu)t}{\sigma^2(t)}x_t\right) d\ {}^t\lambda_n^{(T)} \\
&\leq \exp\left(-nj\frac{(b+(c-\mu)t)^2}{2\sigma^2(t)}\right) = e^{-njI_t}.
\end{aligned}$$

As an aside, we mention that this gives (by choosing $j = 1$), cf. Section 2.2,

$$p_n^T = \sum_{t=1}^T \nu_n^{(T)}(\mathcal{O}_T(t)) \leq \sum_{t=1}^T e^{-nI_t}.$$

The second moment of the cut-and-twist estimator follows from (20):

$$\begin{aligned}
&\frac{1}{N}\int_{\mathbb{R}^T} \left(\sum_{t\in\{1,\ldots,T\}} \mathbf{1}_{\{x_t\in\mathcal{O}_T(t)\}} \frac{d\nu_n^{(T)}}{d\ {}^t\lambda_n^{(T)}}(x_t)\right)^2 d\ {}^1\lambda_n^{(T)}(x_1)\cdots d\ {}^T\lambda_n^{(T)}(x_T) \\
&= \frac{1}{N}\sum_{t\in\{1,\ldots,T\}} \int_{\mathcal{O}_T(t)} \left(\frac{\nu_n^{(T)}}{{}^t\lambda_n^{(T)}}\right)^2 d\ {}^t\lambda_n^{(T)} \\
&\quad + \frac{1}{N}\sum_{\substack{s,t\in\{1,\ldots,T\}\\ s\neq t}} {}^s\lambda_n^{(T)}(\mathcal{O}_T(t))\cdot\ {}^t\lambda_n^{(T)}(\mathcal{O}_T(t)),
\end{aligned}$$

and therefore it is bounded by

$$\frac{1}{N}\left[\sum_{t\in\{1,\ldots,T\}} \exp\left(-n\frac{(b+(c-\mu)t)^2}{2\sigma^2(t)}\right)\right]^2 \leq \frac{1}{N}T^2\exp(-2nI_{t^*}),$$

where the last inequality is due to the definition of $t^* = \arg\inf_t I_t$. Now take logarithms, divide by n, and let $n\to\infty$ to see that the relative efficiency equals 2, cf. (12).

A.4 Proofs for the sequential-twist method

A.4.1 *Proof of Lemma 4.6.* We have to prove that

$$\operatorname*{arg\,inf}_{\{y\in\mathbb{R}^T;\frac{1}{n}\sum_{i=1}^j \bar{A}_i+(1-j/n)y\in\mathcal{O}_T\}} \|y\|_{\mathcal{H}}^2 = \frac{J_{j+1}}{\sigma(t_{j+1}^*)}\,\Gamma(\cdot,t_{j+1}^*).$$

From Lemma 4.1, we know that the infimum equals J_{j+1}^2. It is not hard to see that μ_{j+1} attains this value (by strict convexity of $\|\cdot\|_{\mathcal{H}}$, the minimizing argument is even unique).

A.4.2 *Proof of Theorem 4.7.* The two assumptions in Condition 2.1 of Dupuis and Wang [2004] hold: since we are in a multivariate Gaussian setup we obviously have an everywhere finite moment generating function, and Lemma A.2 implies that

$$\inf_{x\in\overline{\mathcal{O}_T}} x' \left(\Gamma^{(T)}\right)^{-1} x = \inf_{x\in\mathcal{O}_T^o} x' \left(\Gamma^{(T)}\right)^{-1} x.$$

The claim is Theorem 2.1 of Dupuis and Wang [2004].

REFERENCES

Addie, R., Mannersalo, P., and Norros, I. 2002. Most probable paths and performance formulae for buffers with Gaussian input traffic. *European Transactions on Telecommunications 13*, 183–196.

Asmussen, S. 1989. Risk theory in a Markovian environment. *Scand. Actuar. J.*, 69–100.

Asmussen, S. 2000. *Ruin probabilities*. World Scientific Publishing Co. Inc.

Asmussen, S. and Binswanger, K. 1997. Simulation of ruin probabilities for subexponential claims. *ASTIN Bulletin 27*, 297–318.

Asmussen, S. and Rubinstein, R. Y. 1995. Steady state rare events simulation in queueing models and its complexity properties. In *Advances in queueing*, J. Dshalalow, Ed. CRC, Boca Raton, FL, 429–461.

Bahadur, R. R. and Zabell, S. L. 1979. Large deviations of the sample mean in general vector spaces. *Ann. Probab. 7*, 587–621.

Baldi, P. and Pacchiarotti, B. 2004. Importance sampling for the ruin problem for general Gaussian processes. Preprint.

Bingham, N. H., Goldie, C. M., and Teugels, J. L. 1989. *Regular variation*. Cambridge University Press, Cambridge.

Boots, N. K. and Mandjes, M. 2002. Fast simulation of a queue fed by a superposition of many (heavy-tailed) sources. *Probab. Engrg. Inform. Sci. 16*, 205–232.

Bucklew, J. A., Ney, P., and Sadowsky, J. S. 1990. Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains. *J. Appl. Probab. 27*, 44–59.

Collamore, J. F. 2002. Importance sampling techniques for the multidimensional ruin problem for general Markov additive sequences of random vectors. *Ann. Appl. Probab. 12*, 382–421.

Davies, R. B. and Harte, D. S. 1987. Tests for Hurst effect. *Biometrika 74*, 95–102.

Dębicki, K. and Mandjes, M. 2003. Exact overflow asymptotics for queues with many Gaussian inputs. *J. Appl. Probab. 40*, 704–720.

Dębicki, K. and Palmowski, Z. 1999. On-off fluid models in heavy traffic environment. *Queueing Syst. 33*, 327–338.

Dembo, A. and Zeitouni, O. 1998. *Large deviations techniques and applications*, Second ed. Springer-Verlag, New York.

Deuschel, J.-D. and Stroock, D. W. 1989. *Large deviations*. Academic Press Inc., Boston, MA.

Dieker, A. B. 2005. Conditional limit theorems for queues with Gaussian input, a weak convergence approach. *Stochastic Process. Appl. 115*, 849–873.

Dieker, A. B. and Mandjes, M. 2003. On spectral simulation of fractional Brownian motion. *Probab. Engrg. Inform. Sci. 17*, 417–434.

Dieker, A. B. and Mandjes, M. 2005. On asymptotically efficient simulation of large deviation probabilities. *Adv. in Appl. Probab. 37*, 539–552.

Dietrich, C. R. and Newsam, G. N. 1997. Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. *SIAM Journal on Scientific Computing 18,* 4, 1088–1107.

Dupuis, P. and Wang, H. 2004. Importance sampling, large deviations, and differential games. *Stoch. Stoch. Rep. 76*, 481–508.

Ganesh, A., O'Connell, N., and Wischik, D. 2004. *Big queues.* Springer-Verlag, Berlin.

Glasserman, P. and Wang, Y. 1997. Counterexamples in importance sampling for large deviations probabilities. *Ann. Appl. Probab. 7*, 731–746.

Heidelberger, P. 1995. Fast simulation of rare events in queueing and reliability models. *ACM Trans. Modeling Comp. Simulation 5*, 43–85.

Huang, C., Devetsikiotis, M., Lambadaris, I., and Kaye, A. R. 1999. Fast simulation of queues with long-range dependent traffic. *Comm. Statist. Stochastic Models 15,* 3, 429–460.

Kelly, F. 1996. Notes on effective bandwidths. In *Stochastic networks: theory and applications*, F. Kelly, S. Zachary, and I. Ziedins, Eds. Oxford University Press, 141–168.

Lehtonen, T. and Nyrhinen, H. 1992a. On asymptotically efficient simulation of ruin probabilities in a Markovian environment. *Scand. Actuar. J.*, 60–75.

Lehtonen, T. and Nyrhinen, H. 1992b. Simulating level-crossing probabilities by importance sampling. *Adv. in Appl. Probab. 24*, 858–874.

Leland, W. E., Taqqu, M. S., Willinger, W., and Wilson, D. V. 1994. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Trans. on Networking 2,* 1, 1–15.

Likhanov, N. and Mazumdar, R. 1999. Cell loss asymptotics for buffers fed with a large number of independent stationary sources. *J. Appl. Probab. 36*, 86–96.

Michna, Z. 1999. On tail probabilities and first passage times for fractional Brownian motion. *Math. Methods Oper. Res. 49*, 335–354.

Mitrinović, D. S. 1970. *Analytic inequalities.* Springer-Verlag, New York.

Norros, I. 1994. A storage model with self-similar input. *Queueing Syst. 16*, 387–396.

Rockafellar, R. T. 1970. *Convex analysis.* Princeton University Press, Princeton, N.J.

Sadowsky, J. 1991. Large deviations and efficient simulation of excessive backlogs in a GI/G/m queue. *IEEE Trans. Autom. Control 36*, 579–588.

Sadowsky, J. S. and Bucklew, J. A. 1990. On large deviations theory and asymptotically efficient Monte Carlo estimation. *IEEE Trans. Inform. Theory 36*, 579–588.

Siegmund, D. 1976. Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist. 4*, 673–684.

Wischik, D. 2001a. Moderate deviations in queueing theory. Preprint.

Wischik, D. 2001b. Sample path large deviations for queues with many inputs. *Ann. Appl. Probab. 11*, 379–404.

Wood, A. T. A. and Chan, G. 1994. Simulation of stationary Gaussian processes in $[0,1]^d$. *Journal of Computational and Graphical Statistics 3,* 4, 409–432.

Received April 2004; revised February 2006; accepted March 2006