View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by CWI's Institut

Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

Fast simulation of overflow probabilities in a queue with Gaussian input

A. B. Dieker; M.R.H. Mandjes

REPORT PNA-E0404 APRIL 20, 2004

CWI is the National Research Institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organization for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

Copyright © 2004, Stichting Centrum voor Wiskunde en Informatica
P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

Fast simulation of overflow probabilities in a queue with Gaussian input

ABSTRACT

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 three 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), and the sequential twist method (simulating source by source).

The asymptotic efficiency of these three methods is investigated as $n \to \infty$: for instance, a necessary and sufficient condition is derived for the efficiency of the method based on a single exponential twist. It turns out that this method is asymptotically inefficient in practice, but the other two methods are asymptotically efficient. We evaluate the three methods by performing a simulation study.

2000 Mathematics Subject Classification: primary 65C05; secondary 60F10, 60G15, 60K10.
Keywords and Phrases: Asymptotic efficiency, Gaussian processes, importance sampling, large deviations, overflow probability
Note: The research was supported by the Netherlands Organization for Scientific Research (NWO) under grant 631.000.002.

Fast simulation of overflow probabilities in a queue with Gaussian input

A. B. Dieker and M. Mandjes
CWI
P.O. Box 94079
1090 GB Amsterdam, the Netherlands
and
University of Twente
Faculty of Mathematical Sciences
P.O. Box 217
7500 AE Enschede, the Netherlands

ABSTRACT

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 three 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), and the sequential twist method (simulating source by source).

The asymptotic efficiency of these three methods is investigated as $n \to \infty$: for instance, a necessary and sufficient condition is derived for the efficiency of the method based on a single exponential twist. It turns out that this method is asymptotically inefficient in practice, but the other two methods are asymptotically efficient. We evaluate the three methods by performing a simulation study.

2000 Mathematics Subject Classification: primary 65C05; secondary 60F10, 60G15, 60K10.
Keywords and Phrases: Asymptotic efficiency, Gaussian processes, importance sampling, large deviations, overflow probability
Note: The research was supported by the Netherlands Organization for Scientific Research (NWO) under grant 631.000.002.

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 in the buffer is lost. Hence, queues are an appropriate tool for describing congestion phenomena.

Gaussian traffic. In this paper, we consider a queue fed by *Gaussian* sources. We focus on stationary sources, 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' closely related to long-range dependence; an example is fractional Brownian motion (fBm). Models based on these processes 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., [12].

Asymptotics. It is extremely hard to calculate the full workload distribution of a queue with Gaussian input; it is only known 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 $B \equiv nb$. Buffer overflow (over level nb) becomes rare when n grows large. For fixed but large n, we study the probability p_n of buffer overflow in a discrete time model. Likhanov and Mazumdar [24] find the asymptotics of p_n, i.e., they identify a function g such that $p_n g(n) \to 1$ as $n \to \infty$. 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. Hence, 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, $g(n)$ tends to underestimate p_n, since $g(n)$ asymptotically only involves the probability of overflow at a single time epoch, cf. Equations (2.1) and (3.4) of [24]. Of course, this is highly undesirable if one is interested in reliable estimates.

Simulation. In absence of analytical results, one could resort to simulation. When simulating loss probabilities in queues with Gaussian input, two problems arise. The first is that it is not straightforward to quickly simulate Gaussian processes. 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 [10]. In the important case of fBm, this leads to an extremely fast algorithm (order of $T \log T$ for a trace of length T) for generating fBm traces. A difficulty with this algorithm is that the trace length should be known before the simulation is started. We cope with this by estimating an approximating probability, while controlling the approximation error.

Another problem of simulation is that it is typically hard to estimate small probabilities; we mainly focus on this difficulty in the remainder. This plays a role in our setting, since the overflow probability decreases to zero as $n \to \infty$. The general rule is that, for an estimate with a fixed 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, the simulation is done under a new measure under which overflow occurs more frequently, where we obtain an unbiased estimator by weighing the simulation output by likelihood ratios. Inherently, there is considerable freedom in choosing the importance sampling measure. 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 results are natural candidates for efficient simulation. In fact, they are asymptotically efficient in many settings; see Asmussen and Rubinstein [3] and Heidelberger [19] and references therein. However, Glasserman and Wang [18] 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 overflow probability. It turns out that this estimator is predominantly asymptotically *inefficient* for a wide range of Gaussian inputs, including fBm and (perhaps surprisingly) standard Brownian motion.

As the large deviation estimator is in practice inefficient, a different approach has to be taken. We present two other methods that can be proven to be asymptotically efficient. The first uses ideas of Boots and Mandjes [7], and simulates timeslot by timeslot. The second is based on a paper by Dupuis and Wang [17], 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 performance evaluation of the (inefficient) large deviation estimator, and the two asymptotically efficient approaches.

Some related results on fast simulation of queues with Gaussian input have been reported by Michna [25] and by Huang *et al.* [20]. Michna focuses on fBm input under the so-called *large-buffer scaling*, but does not consider asymptotic efficiency of his simulation scheme (in fact, one may check that his estimator is asymptotically inefficient). Huang *et al.* also work in the large buffer asymptotic regime, and suggest that the large deviation estimator is asymptotically efficient in the many sources regime; Theorem 1 below entails that this need not be the case. We would like to stress that we focus here *only* on the simulation of overflow in the many-sources regime (not necessarily fBm).

Organization. This paper is organized as follows. Section 2 formalizes the framework of the paper, and discusses how the simulation horizon can be truncated in order to still obtain reasonable estimates. Some preliminaries are given in Section 3, and Section 4 studies the properties of the three simulation methods mentioned above. Section 5 contains a numerical evaluation of these methods, and we conclude the paper with a discussion in Section 6.

2. The overflow probability

The present section contains the description of our queueing model. In particular, we show that the buffer overflow probability relates to an infinite time horizon, see Section 2.1. To simulate the overflow probability, this horizon needs to be truncated, where the neglected probability mass is below a tolerable level. This 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).

Define $A_n(\cdot)$ as the aggregate cumulative traffic process. More precisely, let $A_n(t)$ denote the traffic generated by the n sources in $\{1, \ldots, t\}$; for notational convenience, we set $A_n(0) :=$

0 and we suppose that time is indexed by $\mathbb{N}$. 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, so that $\mathbb{E}A_n(t) = n\mu t$. 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.

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), in which $\sigma^2(t)$ is proportional to t^{2H}.

Queueing model. We now turn to the queueing model. In this paper we let the queue's (deterministic) service rate scale 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, that we again scale with the number of sources: $B \equiv nb > 0$. It is well-known that this probability can be expressed in terms of the aggregate cumulative arrival process $A_n(\cdot)$, as follows:

$$p_n = P\left(\sup_{t\in\mathbb{N}}[A_n(t) - nct] > nb\right). \tag{2.1}$$

We remark that the probability p_n of exceeding level 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 overflow probability in continuous time is obtained by replacing $\mathbb{N}$ by $\mathbb{R}_+$ in Eq. (2.1). Notably, the asymptotics of the overflow probability in continuous time differ qualitatively from those in (2.1), see [11]. A further discussion of this issue is relegated to Section 6.

2.2 The simulation horizon

As argued in the Introduction, no error bounds for the asymptotics in (2.1) are available. This motivates the research on methods to quickly simulate p_n. Representation (2.1) however shows that the event of overflow corresponds to a 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. Then we approximate p_n by

$$p_n^T := P\left(\sup_{t\in\{1,\ldots,T\}}[A_n(t) - nct] > nb\right). \tag{2.2}$$

Of course, T should be chosen sufficiently large, in order to make sure that the approximation error is small. To investigate this error, let $\tau_n := \inf\{t \in \mathbb{N} : A_n(t) - nct > nb\}$ denote the epoch of the first buffer overflow, 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 $P(\tau_n > T)$. The resulting estimator is *biased*: it has a mean *smaller* than p_n. As in Boots and Mandjes [7], we choose T such that

$$\frac{P(\tau_n > T)}{p_n} < \epsilon, \tag{2.3}$$

for some predefined $\epsilon > 0$. When ϵ is chosen small, the truncation is clearly of minor impact.

The requirement in (2.3) does not directly translate into an explicit expression for the simulation horizon T. Following [7], this problem is tackled by establishing tractable bounds on $P(\tau_n > T)$ and p_n. We write

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

A lower bound on p_n. Obviously, for any $t \in \mathbb{N}$,

$$\begin{aligned} p_n &\geq P(A_n(t) > nb + nct) \\ &= \int_{\sqrt{n}\frac{b+(c-\mu)t}{\sigma(t)}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) dx \\ &\geq \frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{nI_t} + \sqrt{nI_t + 2}} e^{-nI_t}, \end{aligned} \tag{2.4}$$

where the last inequality is a standard bound for the standard normal cumulative density function (see [26, 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 (2.4) 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' overflow epoch: given that overflow occurs, it is most likely that it happens at epoch t^*; see for instance Wischik [31].

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

$$P(\tau_n > T) = \sum_{t=T+1}^{\infty} P(\tau_n = t) \leq \sum_{t=T+1}^{\infty} P(A_n(t) - nct > nb) \leq \sum_{t=T+1}^{\infty} e^{-nI_t} \tag{2.5}$$

In the present generality, it is difficult to bound this quantity further. We could proceed by focusing on a specific correlation structure, for instance fBm with $\sigma^2(t) = t^{2H}$, for $H \in (0, 1)$. 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 [6] with index α, then then $\sigma^2(t)$ can be bounded from above (for t sufficiently large) by Potter's bound $Ct^{\alpha+\epsilon}$, for any $C > 1$ and $\epsilon > 0$ [6, Thm. 1.5.6]. Obviously, it is desirable to choose the horizon as small as possible under the restriction that (2.3) holds; for this, C and H should be chosen as small as possible.

Under $\sigma^2(t) \leq Ct^{2H}$ we can bound (2.5) as follows:

$$\sum_{t=T+1}^{\infty} e^{-nI_t} \leq \sum_{t=T+1}^{\infty} \exp\left(-n\frac{(c-\mu)^2}{2C}t^{2-2H}\right) \leq \int_T^{\infty} \exp\left(-n\frac{(c-\mu)^2}{2C}t^{2-2H}\right) dt. \tag{2.6}$$

Since the natural way of finding an upper bound critically depends on the value of H, we consider the cases $H \leq 1/2$ and $H > 1/2$ separately. As 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 1 *In case $H \leq 1/2$, we have*

$$\int_T^\infty \exp\left(-nC_0 t^{1/q}\right) dt \leq \frac{q}{C_0 n} \exp\left(-nC_0 T^{1/q}\right). \tag{2.7}$$

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

$$\gamma_q := q - 1 - m, \text{ and } \beta_q := \frac{(q-1)\cdots(q-m)}{\gamma_q^m e^{\gamma_q}}. \tag{2.8}$$

This notation plays a central role in the following lemma, which is proven in Appendix A.1.2.

Lemma 2 *In case $H > 1/2$, we have*

$$\int_T^\infty \exp\left(-nC_0 t^{1/q}\right) dt \leq \frac{q\beta_q}{C_0^q (n-\gamma_q)} \exp\left(-(n-\gamma_q) C_0 T^{1/q}\right).$$

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

Corollary 1 *For $H \leq 1/2$, let $T(n)$ be the largest integer smaller 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 largest integer smaller 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 (2.3) *does not exceed ϵ.*
Moreover, $\overline{T} := \lim_{n\to\infty} T(n) = (I_{t^*}/C_0)^{1/(2-2H)}$.

We recall that t^* could be interpreted as the most likely epoch of overflow. Given that overflow occurs, most of the probability mass will be around t^*. Hence it is not surprising that $\overline{T} > t^*$:

$$\frac{I_{t^*}}{C_0} = \frac{(b+(c-\mu)t^*)^2}{2\sigma^2(t^*)} \Big/ \frac{(c-\mu)^2}{2C} > (t^*)^{2-2H} = (t^*)^{1/q}. \tag{2.9}$$

3. Preliminaries on rare-event simulation

This section provides some background on the simulation of (small) overflow probabilities. Section 3.1 introduces 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 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

$$\lambda(B) := \int_B \frac{d\lambda}{d\nu} d\nu. \tag{3.1}$$

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* (or: likelihood ratio). 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:

$$\widehat{\nu_\lambda(A)} := \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}_{\{X^{(k)} \in A\}} \frac{d\nu}{d\lambda}(X^{(k)}). \tag{3.2}$$

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., [19]. 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 a measurable space $(\mathcal{X}, \mathcal{B})$. 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 (3.2):

$$\widehat{\nu_{\lambda_n}(B)}_N := \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}_{\left\{X_{\lambda_n}^{(k)} \in B\right\}} \frac{d\nu_n}{d\lambda_n}\left(X_{\lambda_n}^{(k)}\right). \tag{3.3}$$

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

$$\eta_N(\lambda_n, B) := \frac{\mathbb{V}\mathrm{ar}_{\lambda_n}\left(\widehat{\nu_{\lambda_n}(B)}_N\right)}{\nu_n(B)^2} = \frac{\mathbb{E}_{\lambda_n}\left(\widehat{\nu_{\lambda_n}(B)}_N\right)^2}{\nu_n(B)^2} - 1; \tag{3.4}$$

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 (3.4), 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. Set $N_{\lambda_n}^* := \inf\{N \in \mathbb{N} : \eta_N(\lambda_n, B) \leq \eta_{\max}\}$. Asymptotic efficiency is sometimes referred to as *asymptotic optimality*, *logarithmic efficiency*, or *weak efficiency*.

Definition 1 *An importance sampling family* $\{\lambda_n\}$ *is called* asymptotically efficient *if*

$$\limsup_{n\to\infty} \frac{1}{n} \log N_{\lambda_n}^* = 0, \tag{3.5}$$

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

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

$$E_n := \frac{\log \int_B \left(\frac{d\nu_n}{d\lambda_n}\right)^2 d\lambda_n}{\log \nu_n(B)}; \tag{3.6}$$

see [15] 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' overflow probability $p_n^{T(n)}$ defined in (2.2).

As argued in Section 3.2, asymptotic efficiency corresponds to the performance of simulation methods for large n. Notice that, by virtue of Corollary 1, we can safely set $T(n) = \lceil \overline{T} \rceil$ for n large enough; for ease set $T := \lceil \overline{T} \rceil$. Conclude that we can restrict ourselves to assess asymptotic efficiency of methods for estimating p_n^T.

In this paper we concentrate on three methods for simulating p_n^T. The first, which we refer to as *single exponential twist*, is the simplest of the three. 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 two asymptotically efficient alternatives: the first solves the theoretical difficulties by simulating timeslot by timeslot (which we therefore call *cut-and-twist*), the second by simulating source by source (*sequential twist*). The former method uses the ideas of Boots and Mandjes [7], whereas latter has recently been proposed by Dupuis and Wang [17].

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 [2, 8, 9, 21, 22, 29]. However, one has to be careful, as a successful application of such an exponential twist critically depends on the specific problem at hand [15, 17, 18]. Before deriving conditions for asymptotic optimality of the exponential twist in the setup of the present paper, we first provide more background.

We denote by $\mathcal{O}_T \subset \mathbb{R}^T$ the set of *paths* that cause overflow up to time $T \in \mathbb{N}$, i.e.,

$$\begin{aligned} \mathcal{O}_T &:= \{x \in \mathbb{R}^T : \exists t \in \{1,\ldots,T\} : x_t + \mu t \geq b + ct\} \\ &= \bigcup_{\{t:t\in\{1,\ldots,T\}\}} \bigcup_{\{y:y+\mu t\geq b+ct\}} \{x \in \mathbb{R}^T : x(t) = y\}. \end{aligned} \tag{4.1}$$

Note that, with $\nu_n^{(T)}$ denoting the distribution of $\{A_n(t)/n - \mu t : t \in \{1,\ldots,T\}\}$,

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

The following lemma, which is proven the appendix, 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., $\Gamma^{(T)} := (\Gamma(s,t))_{s,t=1}^T$.

Lemma 3 *We have*

$$\lim_{n\to\infty} \frac{1}{n} \log \nu_n^{(T)}(\mathcal{O}_T) = -\frac{1}{2} r^{*\prime} \left(\Gamma^{(T)}\right)^{-1} r^* = -I_{t^*}, \tag{4.2}$$

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

$$r_t^* = \frac{b + (c-\mu)t^*}{\sigma^2(t^*)} \Gamma(t^*, t). \tag{4.3}$$

Time epoch t^* can be thought of as the *most likely epoch of overflow*: as n grows the probability of overflow decays exponentially, but given that it occurs, with overwhelming probability the busy period preceding overflow has duration t^*. Likewise, r^* can informally be interpreted as the *most likely path to overflow*; note that indeed $r_{t^*}^* = b + (c-\mu)t^*$. It is important to realize that r^* is piecewise linear only in the case of (scaled) Brownian input (i.e., $\sigma^2(t) = Ct$ for some $C > 0$); in general r^* is a 'curved' path.

We can now to 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

$$\lambda_n^{(T)}(A) = \int_A \exp\left(n \frac{b + (c-\mu)t^*}{\sigma^2(t^*)} x_{t^*} - n I_{t^*}\right) \nu_n^{(T)}(dx). \tag{4.4}$$

By observing that

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

it can be verified that the new measure $\lambda_n^{(T)}$ corresponds to the distribution of a Gaussian process with mean vector $(r_t^*)_{t=1,\ldots,T}$, 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 r^* for large n, we say that this exponential twist is in accordance with the large deviation behavior of Lemma 3.

The following theorem is the main result of this subsection. It presents sufficient and necessary conditions for asymptotic optimality of the estimator determined by (3.3), where λ_n is given by (4.4). Its proof is given in Appendix A.2.

We recently came across a related theorem by [5]. An important difference is that these authors study the continuous-time overflow 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 [14]. However, we believe that discrete time is more natural in a simulation framework; see also Section 6. Another difference is the proof technique; [5] use recent insights into certain Gaussian martingales, while we take a direct approach.

Theorem 1 *Importance sampling under a 'single exponential twist' is asymptotically efficient for simulating p_n^T if and only if*

$$\inf_{t\in\{1,\ldots,T\}} \frac{b + (c-\mu)t + r_t^*}{\sigma(t)} = 2\,\frac{b + (c-\mu)t^*}{\sigma(t^*)}. \tag{4.5}$$

Clearly

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

hence Theorem 1 states that the change of measure is asymptotically efficient if and only if $h(t) \geq h(t^*)$ for all $t \in \{0,\ldots,T\}$.

In the above we represented time by the natural numbers $\mathbb{N}$, i.e., we used a grid with mesh 1. Obviously, the same techniques can be used to prove a similar statement for any 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 appendix to see that the infimum in (4.5) should then be taken over $m\mathbb{N} \cap [0,T]$. Thus, by refining the grid, the left hand side of (4.5) 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 $g(t) := [b + (c-\mu)t + \bar{r}^*(t)]/\sigma(t)$, where $\bar{r}^*$ denotes the continuous-time analogue of (4.3):

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

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

$$\lim_{t\uparrow t^*} \ddot{g}(t) = \frac{1}{4}\frac{b+(c-\mu)t^*}{\sigma^3(t^*)}\left[\frac{[\dot{\sigma}^2(t^*)]^2}{\sigma^2(t^*)} - \ddot{\sigma}^2(t^*) - \ddot{\sigma}^2(0)\right], \tag{4.6}$$

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 the whether the single exponential twist becomes inefficient as the mesh tends to zero in specific cases. In particular, 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. Moreover, if $H > 1/2$, it follows from (4.6) and $\ddot{\sigma}^2(0) = \infty$ that $\lim_{t\uparrow t^*} \ddot{g}(t) < 0$, so that we arrive at the same conclusion as in the case $H \leq 1/2$.

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 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 $d\nu_n/d\lambda_n$ following from (4.4):

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

where the x_{t^*} corresponds to the value of $A_n(t^*)/n - \mu t^*$. For asymptotic optimality, this likelihood should be 'small' for realizations in the overflow set $\mathcal{O}_T$. If there is overflow at time t^*, then clearly

$$\frac{d\nu_n^{(T)}}{d\lambda_n^{(T)}} \leq e^{-nI_{t^*}} \tag{4.7}$$

(use $x_{t^*} \geq b + (c-\mu)t^*$). However, if overflow occurs at any other time epoch, clearly the likelihood 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. Condition (4.5) 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 sub-events, and simulating these individually. To this end, write

$$p_n^T = \nu_n^{(T)}\left(\bigcup_{t\in\{1,\ldots,T\}} \mathcal{O}_T(t)\right) = \sum_{t\in\{1,\ldots,T\}} \nu_n^{(T)}(\mathcal{O}_T(t)), \tag{4.8}$$

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

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

notice that the $\mathcal{O}_T(t)$ are *disjoint* events. Hence, the problem reduces to the simulation of T probabilities of the type $\nu_n^{(T)}(\mathcal{O}_T(t))$. This partitioning approach is based on Boots and Mandjes [7], where this idea is exploited for a queue fed by (discrete-time) on-off sources.

The resulting simulation algorithm, to be called 'cut-and-twist', works as follows. Define the exponential twisted measure ${}^t\lambda_n^{(T)}$ as in (4.4), 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 considering the efficiency of this method, we summarize the approach by noting that the estimator equals

$$\frac{1}{N}\sum_{k=1}^{N}\sum_{t\in\{1,\ldots,T\}} \mathbf{1}_{\left\{X_t^{(k)}\in\mathcal{O}_T(t)\right\}} \frac{d\nu_n^{(T)}}{d\lambda_n^{(T)}}\left(X_t^{(k)}\right), \tag{4.9}$$

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 for any $x \in \mathcal{O}_T(t)$ the corresponding likelihood is uniformly bounded by $\exp(-nI_t)$, cf. (4.7).

Theorem 2 *The 'cut-and-twist' method is asymptotically efficient for estimating p_n^T.*

Despite being asymptotically optimal, the obvious drawback of this method is that it may take a substantial amount of time to simulate the T probabilities individually.

4.3 The sequential twist method

Recently, Dupuis and Wang [17] introduced an intuitively appealing approach for rare-event simulation. We now give a brief description of the method in the setting of overflow in queues with Gaussian input, although the method is known to work in a considerably more general setting. Consider a sequence $\bar{A}_1, \bar{A}_2, \ldots$ of i.i.d. random vectors taking values in $\mathbb{R}^T$, where the $\bar{A}_j$ are distributed as $\{A_1(t) - \mu t : t \in \{1,\ldots,T\}\}$; as a consequence, the $\bar{A}_j$ have distribution $\nu_1^{(T)}$. Note that p_n^T can be written as

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\bar{A}_i(t) \in B\right), \text{ with } B := \left\{x \in \mathbb{R}^T : \sup_{t\in\{1,\ldots,T\}} x_t - (c-\mu)t \geq b\right\},$$

and hence

$$p_n^T = \int_{\{(x^{(1)},\ldots,x^{(n)}):\frac{1}{n}\sum_{i=1}^n x^{(i)}\in B\}} \nu_1^{(T)}(dx^{(1)})\cdots\nu_1^{(T)}(dx^{(n)}). \tag{4.10}$$

Instead of twisting ν_n^T as in the previous methods, the sequential twist method twists *every copy* of $\nu_1^{(T)}$ (i.e., every source) in Equation (4.10) 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 $j\in\{0,\ldots,n-1\}$). We aim to find the 'cheapest' way to reach the overflow set B *given* $\bar{A}_1,\ldots,\bar{A}_j$. Hence, we do not change the measure if already $\frac{1}{n}\sum_{i=1}^j \bar{A}_i \in B$ (under this condition it is not rare anymore to reach B, due to $\mathbb{E}\bar{A}_j(t)=0$), and otherwise we change the mean of the distribution of $\bar{A}_{j+1}$ to μ_{j+1}, where

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

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 *Define for* $j\in\{0,\ldots,n-1\}$

$$t_{j+1}^* := \underset{t\in\{1,\ldots,T\}}{\arg\inf}\ \frac{nb+n(c-\mu)t-\sum_{i=1}^j \bar{A}_i(t)}{(n-j)\sigma(t)}, \tag{4.11}$$

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

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

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

$$\prod_{j=1}^n \exp\left(-\frac{J_j}{\sigma(t_j^*)}\bar{A}_j(t_j^*)+\frac{1}{2}J_j^2\right). \tag{4.12}$$

An estimator is obtained by repeating this procedure N times, and computing the estimate using (3.2); of course, the underlying samples must be independent.

The conditions for the following theorem of Dupuis and Wang [17] are checked in the appendix.

Theorem 3 (Dupuis-Wang) *The 'sequential twist' method is asymptotically efficient for estimating* p_n^T.

A drawback of this approach is that all sources should be generated individually. Note that only one vector is generated in the single-twist method (as the aggregate of Gaussian sources is again Gaussian), and T vectors for the cut-and-twist method. However, the sequential approach can obviously 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 $n/M \in \mathbb{N}$, and 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 approach*; since

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

Theorem 3 yields the asymptotic efficiency of the batch sequential estimator.

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 a reasonable n. This issue is addressed empirically in Section 5.4.

5. Evaluation

While the preceding sections are applicable to any Gaussian process with stationary increments (satisfying certain conditions), in this section we evaluate the simulation methods for the practically relevant case of fractional Brownian motion.

Simulation of fractional Brownian motion is highly nontrivial. As the simulation grid is equispaced, it is best to simulate the (stationary!) incremental process, also 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 Davies and Harte method [10]. 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 complexity is of order $T \log T$ for a sample size of length T. For more details on this method, we refer to [16] and [32].

We evaluate the three methods of Section 4 as follows. First, we check empirically whether the overflow probability decays exponentially, and whether our simulations support the claims in Theorems 1, 2, and 3. After this preliminary analysis, we study the reliability of the methods by refining the simulation grid. Further empirical insight into the methods is gained by studying the influence of the Hurst parameter on the simulation horizon and of the batch size on the relative efficiency. The evaluation is concluded by a time-complexity analysis.

5.1 Empirical verification of the theory

In Section 4, we noted that the overflow probability decays exponentially in n and we studied whether the three discussed simulation methods are asymptotically efficient. In the present subsection, our aim is to validate these theoretical results by performing a simulation experiment. The need for a variance reduction technique is illustrated by including 'naive' Monte Carlo simulation in our analysis.

The parameters are chosen as follows: $b = 0.3$, $c - \mu = 0.1$, $H = 0.8$, $M = 1$, $\epsilon = 0.05$, and $\eta_{\max} = (0.1/1.96)^2$. Unless stated otherwise, we use these parameters throughout this

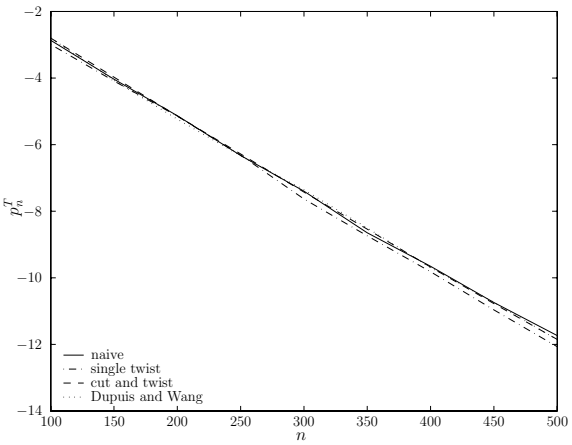

Figure 1: Empirical verification of the exponential decay of the overflow probability.

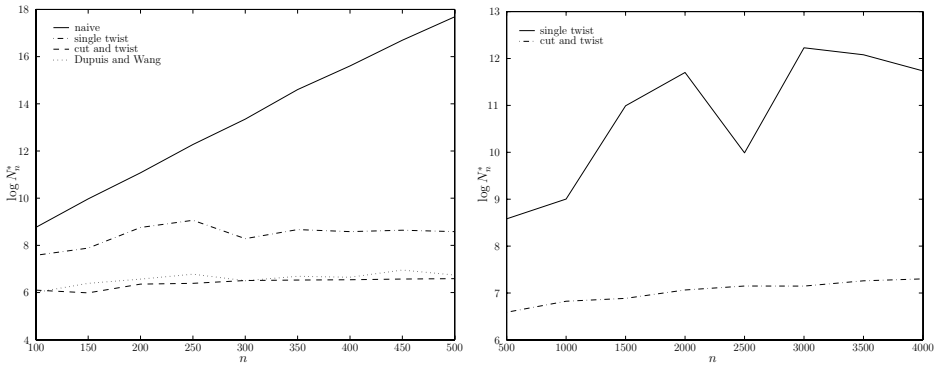

Figure 2: Empirical verification of the asymptotic efficiency of the simulation methods.

section. The choice $H = 0.8$ is supported by measurements on LAN data traffic [23], and $\eta_{\max}$ is chosen such that the width of the confidence interval is 20% of the estimated probability. It is left to the reader to check that condition (4.5) does not hold, i.e., the single-twist estimator is not asymptotically efficient.

We start by checking the exponential decay of p_n; see Figure 1. Notice that the probability is plotted on a logarithmic scale, hence the straight lines. The confidence intervals are not plotted, since the simulation method has almost no influence on their width by construction.

We next 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. Note that under asymptotic efficiency $\log N_n^*$ should be (ultimately) sublinear. Therefore, the plots support Theorem 2, Theorem 3, and the fact that the naive estimator is inefficient.

However, it is not immediate from the left panel of Figure 2 that the single-twist method is asymptotically inefficient. Although the irregular behavior indicates that this might indeed be the case, we find more evidence by increasing n further. This is done in the right panel of

Figure 2. We have also included the results for the (asymptotically efficient) cut-and-twist method in order to show the difference.

The unstable behavior in both plots of the single-twist method is closely related to the interpretation of a possible failure of the exponential twist (see Section 4). As noted there, overflow occurs at time epoch t^* in a 'typical' simulation run, but it might also happen that overflow occurs at some other time $t \neq t^*$. 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 overflow probability is underestimated.

5.2 Simulation grid

While the observations in the previous subsection were predicted by theory, we now shift our attention to experiments that give insights into the performance of the methods in practice. As a first step, investigate the influence of the grid mesh on the estimated probability. We will see that such an analysis provides valuable insights into the reliability of the estimated probabilities.

We evaluate

$$P\left(\sup_{t\in\{\alpha,2\alpha,\ldots\}} \bar{A}_n(t) - n(c-\mu)t > nb\right) \tag{5.1}$$

for a range of $\alpha \geq 0$, in such a way that the simulation grid becomes finer. For instance, one can take $\alpha = 1, 1/2, 1/4, 1/8$; the probability then increases as α is made smaller, as we only add grid points. We get some idea how reliable the simulation methods are by checking whether the estimates indeed increase.

Before we can compare the estimated probabilities for different α, we have to take into account what happens to the simulation horizon as α decreases. Indeed, if the horizon decreases as $\alpha \to 0$, one cannot conclude that the above monotonicity of carries over to the 'truncated' approximating probabilities. Since $A_n(t)$ is a scaled fractional Brownian motion by assumption, the self-similarity property yields that (5.1) equals

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

Therefore, 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 (2.9),

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

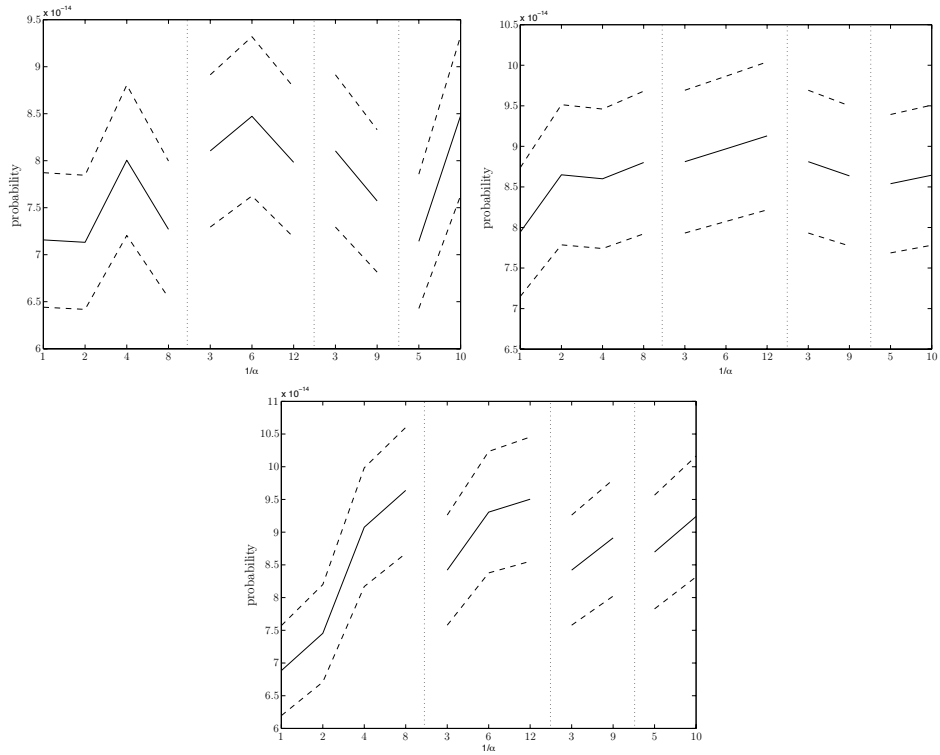

Figure 3: The influence of the grid mesh on the probability for the single-twist method, the cut-and-twist method, and the sequential twist method. The solid lines represent the estimates, while the dashed lines correspond to confidence intervals.

which is monotonic in α and tends to infinity as $\alpha \to 0$. An ideally chosen simulation horizon has a factor b in place of b/α; this factor does not appear here since we used the lower bound $b \geq 0$ in the approximation procedure, see (2.6). Therefore, for small α, it takes a substantial amount of effort to account for non-significant contributions to the probability, but this cannot decrease the probability; the monotonicity is preserved.

In order to investigate whether the estimates indeed decrease in α, we perform some simulations with parameters $n = 150$, $b = 0.9$, $c - \mu = 0.3$, $H = 0.8$, and $M = 1$. ϵ and $\eta_{\max}$ are chosen as before. 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 three different methods. The dotted lines correspond to the boundaries of the confidence intervals. Note that it is possible to draw increasing lines between the boundaries of the confidence intervals for every method. However, we see that the estimates themselves only show the expected (and theoretically correct) behavior for the sequential twist method. The estimates for the single-twist method are disappointing, as expected from its asymptotic inefficiency (see Theorem 1). The estimates for cut-and-twist method are reasonable, with the exception of $\alpha = 1/9$.

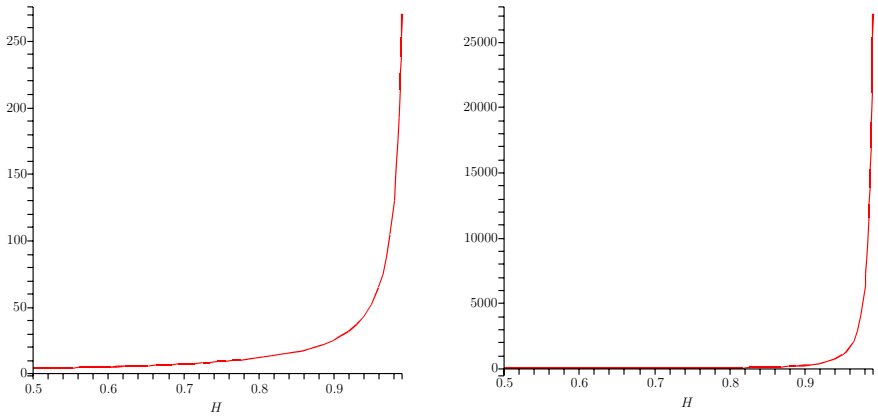

Figure 4: $\tilde{T}(H)$ as a function of H (left panel) and its derivative (right panel).

5.3 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 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 entire halfline, we see that $T(H)$ can be approximated by

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

It becomes clear that $\tilde{T}(H)$ 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 4, we have plotted this function and its derivative.

It is intuitively clear that $\tilde{T}(H)$ increases in H. The higher H, the more long-term correlations are present, and the more time is required 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. (2.3).

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. Then one only twists n/M times in one simulation run, which limits the flexibility of the method and therefore makes it less efficient. Hence, there is a trade-off between the batch size M and the efficiency.

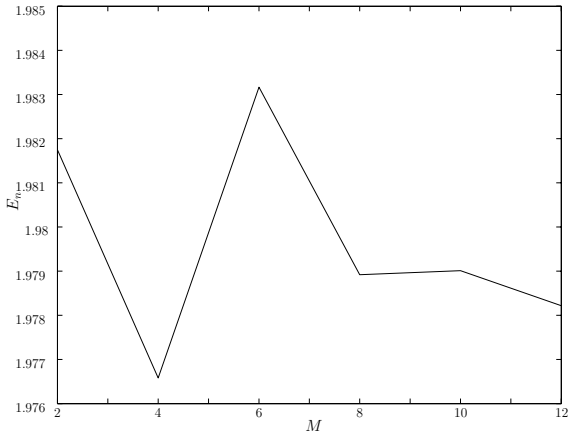

Figure 5: The relative efficiency as a function of M for small M.

Theorem 3 states that the method is asymptotically efficient as $n \to \infty$, regardless the value of M. Suppose that for $M = 1$ the method is 'near' efficiency for $n \geq n_0$. One can then expect that, for general M, the method is only 'nearly' efficient for $n \geq Mn_0$, i.e., that the simulation becomes less efficient as M increases. We now check whether this is indeed the case.

As earlier, 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 5. From the plot, it is not so obvious that an increase in M makes the simulation less efficient. Therefore, we also compute the efficiency for $M = 80, 160, 240, 320, 480, 640, 960$; the relative efficiency is then estimated as 1.967898, 1.974371, 1.967983, 1.967482, 1.968573, 1.954502 and 1.953645 respectively. These values indeed suggest that the simulation becomes less efficient as M increases. Although the differences look small, one must keep in mind that this quantity is relates the *exponential* decay rate of the variance of the estimator to the exponential decay rate of the probability to be estimated.

Therefore, small differences blow up exponentially, and we propose to always chose M as small as possible. Still, it is an option to increase M if the simulation takes too long due to the chosen system parameters.

5.5 Time-complexity analysis of efficient methods

As the cut-and-twist method and the sequential twist method are the only efficient methods in practice, it is natural to ask which one is the fastest. An unambiguous answer to this question cannot be given, as it depends on the parameters which method should be preferred.

We have already observed that if the simulation horizon is large (i.e., large $b/(c - \mu)$ or large Hurst parameter H), much time is needed to generate the fractional Brownian motion samples. In the cut-and-twist method, such a sample is needed for each time epoch (of which there are T). In the sequential twist method, such a sample is needed for each source (of which there are n). Moreover, the best twist has to be calculated n times, which amounts to computing the infimum in (4.11). This computation is of order T.

An obvious advantage of the cut-and-twist method is that the required simulation time

large	cut-and-twist	sequential simulation
n	$+$	$-$
H	$-$	$+$
$b/(c-\mu)$	$-$	$+$

Table 1: Comparison of the cut-and-twist method and the sequential simulation method.

mildly depends on n (due to the fact that $T(n) \to T$). Therefore, the method is more attractive if n is large. The computational effort of the sequential twist method is roughly proportional to n.

We summarize our findings in Table 5.5.

6. Discussion

In this section, we stress three issues related to the findings of the present paper. First, we explain why the overflow probability in discrete time cannot be considered as 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. Finally, we discuss another approach that yields asymptotic efficiency, which is not addressed in this paper.

Discrete time vs. continuous time. It is important to realize that the probability (2.1) behaves qualitatively different in continuous time, i.e., when $\mathbb{N}$ is replaced by $\mathbb{R}_+$. We illustrate this by recalling the asymptotics of (2.1) 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 [24]

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

where t^* minimizes $[b+(c-\mu)t]/\sigma(t)$. 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, [11]

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

for some constant $\mathcal{K}'$. We note that an unknown constant, *Pickands' constant*, appears in the asymptotics of $p_n^{\mathbb{R}_+}$; this constant is not present in the discrete time case. 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. In reality, network traffic is clearly non-Gaussian (the traffic is non-negative), which raises the question why one may be interested in an overflow probability with Gaussian input. To answer this question, one should keep in mind that the overflow probability in a Gaussian model may still be a good approximation for the 'real' overflow probability. 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 recent paper by Wischik [30] for a detailed discussion.

Other approaches. Another approach can be taken to obtain asymptotically efficient estimates (for the discrete-time probability!). The method is discussed in Sadowsky and Bucklew [28], and is related to the 'cut-and-twist' approach, where one simulates for all t the probability $\nu_n^{(T)}(\mathcal{O}_T(t))$ with importance sampling distribution ${}^t\lambda_n^{(T)}$. In the method of [28], a *random* t is drawn in each simulation run according to some (arbitrary) distribution P with support $\{1,\ldots,T\}$, and then the single-twist method is performed with the distribution ${}^t\lambda_n^{(T)}$. However, it is unclear what distribution P should be chosen. For instance, if P is mostly concentrated on t^*, one can expect similar problems as for the single-twist method.

A. Appendix: proofs

In this appendix, we provide proofs of the assertions in this paper. We start in Section 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 is given in Appendix A.4.

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

We distinguish the cases $q \le 1$ (Lemma 1) and $q > 1$ (Lemma 2).

A.1.1 Proof of Lemma 1 Since $q \le 1$ and $T \in \mathbb{N}$, we can bound the left hand side of (2.7) 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 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,

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

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

$$y^{q-1} \le \beta e^{\gamma y} \tag{A.2}$$

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 [27] we have for all $y \in \mathbb{R}_+$,

$$y^{q-1} \le 1 + (q-1)(y-1). \tag{A.3}$$

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

$$\beta e^{\gamma y} \geq \beta e^{\gamma} + \beta\gamma e^{\gamma}(y-1). \tag{A.4}$$

By comparing (A.3) to (A.4), we see that $y^{q-1} \leq \beta e^{\gamma y}$ upon choosing $\gamma = q-1$ and $\beta = e^{-\gamma}$.

To find β, γ such that (A.2) 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 (A.2) 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

$$\beta\gamma^m e^{\gamma y} \geq (q-1)\cdots(q-m)y^{q-m-1}. \tag{A.5}$$

Note that the right hand side of (A.5) 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 (2.8) solve this system of equations uniquely. As before, Theorem 25.1 of Rockafellar [27] 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 (A.1):

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

A.2 Proofs for the single-twist method

The key ingredient in the proof of Theorem 1 is a large deviation principle (LDP) known as Cramér's theorem. Therefore, we start by discussing this theorem in more detail.

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 [13] for a rigorous introduction to the theory, or to Deuschel and Stroock [14].

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 [4].

Theorem 4 (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 $G \subset \mathbb{R}^T$*:* $\liminf_{n\to\infty} \frac{1}{n} \log \nu_n^{(T)}(G) \geq -\frac{1}{2} \inf_{x\in G} \|x\|_{\mathcal{H}}^2$.

T.he proof can be found in Dembo and Zeitouni [13, 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 3 Lemma 3 is an application of Cramér's theorem. We have to prove that

$$\lim_{n\to\infty} \frac{1}{n} \log \nu_n^{(T)}(\mathcal{O}_T) = -\frac{1}{2} \inf_{x\in\mathcal{O}_T} \|x\|_{\mathcal{H}}^2 = -\frac{1}{2}\|r^*\|_{\mathcal{H}}^2. \tag{A.6}$$

The second equality in (A.6) is due to Addie *et al.* [1]. 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 (A.6) now follows upon combining these upper and lower bounds with the following lemma (applied for $r = 0$).

Lemma 5 *For all* $r \in \mathbb{R}^T$*, we have*

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

F.irst note that the interior of the overflow 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,

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

where

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

A similar reasoning that led to the second equality in (A.6) now yields the desired. □

A.2.3 Proof of Theorem 1 As outlined in Section 2.2 of [15], it is a consequence of Lemma 5 (with $r = 0$) that the single exponential twist is asymptotically efficient if and only if

$$\limsup_{n\to\infty} \frac{1}{n} \log \int_{\mathcal{O}_T} \frac{d\lambda_n^{(T)}}{d\nu_n^{(T)}}(x)\lambda_n^{(T)}(dx) \leq -\frac{(b+(c-\mu)t^*)^2}{\sigma^2(t^*)} = -2I_{t^*}, \tag{A.7}$$

cf. (3.6). In principle, the statement can be proven using Theorem 1 of [15]. However, the argument can be given directly in this case. We apply Varadhan's Integral Lemma (Theorem 4.3.1 of Dembo and Zeitouni [13]) to the left hand side of (A.7). 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 (A.7), but, in view of Lemma 5, the resulting bounds coincide. We may therefore conclude that the lim sup is actually a proper limit; the reader is referred to Section 3.1 of [15] for more details on this reasoning. Application of Varadhan's Lemma gives

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

where the last equality is due to Lemma 5. The claim follows by combining this with (A.7).

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

In this subsection, we proof Theorem 2. Observe that for any $j \in \mathbb{N}$, 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 (4.9):

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

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. (3.6).

A.4 Proofs for the sequential twist method

A.4.1 Proof of Lemma 4 We have to prove that

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

From Lemma 3, 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 3 The two assumptions in Condition 2.1 of Dupuis and Wang [17] hold: since we are in a multivariate Gaussian setup we obviously have an everywhere finite moment generating function, and Lemma 5 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 [17].

References

1. R. Addie, P. Mannersalo, and I. Norros. Most probable paths and performance formulae for buffers with Gaussian input traffic. *European Transactions on Telecommunications*, 13:183–196, 2002.
2. S. Asmussen. Risk theory in a Markovian environment. *Scand. Actuar. J.*, pages 69–100, 1989.
3. S. Asmussen and R. Y. Rubinstein. Steady state rare events simulation in queueing models and its complexity properties. In J. Dshalalow, editor, *Advances in queueing*, pages 429–461. CRC, Boca Raton, FL, 1995.

4. R. R. Bahadur and S. L. Zabell. Large deviations of the sample mean in general vector spaces. *Ann. Probab.*, 7:587–621, 1979.
5. P. Baldi and B. Pacchiarotti. Importance sampling for the ruin problem for general Gaussian processes. Preprint, 2004.
6. N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular variation*. Cambridge University Press, Cambridge, 1989.
7. N. K. Boots and M. Mandjes. Fast simulation of a queue fed by a superposition of many (heavy-tailed) sources. *Probab. Engrg. Inform. Sci.*, 16:205–232, 2002.
8. J. A. Bucklew, P. Ney, and J. S. Sadowsky. Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains. *Journal of Applied Probability*, 27:44–59, 1990.
9. J. F. Collamore. Importance sampling techniques for the multidimensional ruin problem for general Markov additive sequences of random vectors. *Ann. Appl. Probab.*, 12:382–421, 2002.
10. R. B. Davies and D. S. Harte. Tests for Hurst effect. *Biometrika*, 74:95–102, 1987.
11. K. Dębicki and M. Mandjes. Exact overflow asymptotics for queues with many Gaussian inputs. *J. Appl. Probab.*, 40:704–720, 2003.
12. K. Dębicki and Z. Palmowski. On-off fluid models in heavy traffic environment. *Queueing Systems Theory Appl.*, 33:327–338, 1999.
13. A. Dembo and O. Zeitouni. *Large deviations techniques and applications*. Springer-Verlag, New York, second edition, 1998.
14. J.-D. Deuschel and D. W. Stroock. *Large deviations*. Academic Press Inc., Boston, MA, 1989.
15. A. B. Dieker and M. Mandjes. On asymptotically efficient simulation of large deviation probabilities. Technical Report PNA-R0311, CWI, the Netherlands, 2003.
16. C. R. Dietrich and G. N. Newsam. Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. *SIAM Journal on Scientific Computing*, 18(4):1088–1107, 1997.
17. P. Dupuis and H. Wang. Importance sampling, large deviations, and differential games. Technical report, Lefschetz Center for Dynamical Systems, Brown University, 2002.
18. P. Glasserman and Y. Wang. Counterexamples in importance sampling for large deviations probabilities. *Ann. Appl. Probab.*, 7:731–746, 1997.
19. Ph. Heidelberger. Fast simulation of rare events in queueing and reliability models. *ACM Trans. Modeling Comp. Simulation*, 5:43–85, 1995.
20. C. Huang, M. Devetsikiotis, I. Lambadaris, and A. R. Kaye. Fast simulation of queues with long-range dependent traffic. *Communications in Statistics. Stochastic Models*, 15(3):429–460, 1999.
21. T. Lehtonen and H. Nyrhinen. On asymptotically efficient simulation of ruin probabilities in a Markovian environment. *Scand. Actuar. J.*, pages 60–75, 1992.

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

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

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

25. Z. Michna. On tail probabilities and first passage times for fractional Brownian motion. *Mathematical Methods of Operations Research*, 49:335–354, 1999.

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

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

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

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

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

31. D. Wischik. Sample path large deviations for queues with many inputs. *Annals of Applied Probability*, 11:379–404, 2001.

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