J. Appl. Prob. **48,** 467–489 (2011)
Printed in England

EFFICIENT SIMULATION FOR THE MAXIMUM OF INFINITE HORIZON DISCRETE-TIME GAUSSIAN PROCESSES

JOSE BLANCHET [*] [**] AND
CHENXIN LI,[*] [***] *Columbia University*

Abstract

We consider the problem of estimating the probability that the maximum of a Gaussian process with negative mean and indexed by positive integers reaches a high level, say b. In great generality such a probability converges to 0 exponentially fast in a power of b. Under mild assumptions on the marginal distributions of the process and no assumption on the correlation structure, we develop an importance sampling procedure, called the target bridge sampler (TBS), which takes a polynomial (in b) number of function evaluations to achieve a small relative error. The procedure also yields samples of the underlying process conditioned on hitting b in finite time. In addition, we apply our method to the problem of estimating the tail of the maximum of a superposition of a large number, n, of independent Gaussian sources. In this situation TBS achieves a prescribed relative error with a bounded number of function evaluations as $n \nearrow \infty$. A remarkable feature of TBS is that it is *not* based on exponential changes of measure. Our numerical experiments validate the performance indicated by our theoretical findings.

Keywords: Importance sampling; rare-event simulation; Gaussian process; large deviations, fractional Brownian noise

2010 Mathematics Subject Classification: Primary 65C05
Secondary 60G15; 60F10

1. Introduction

Gaussian processes constitute a popular class of models used in various settings in engineering and science, ranging from econometrics to communication networks (see Brockwell and Davis (1991) and Mandjes (2007, Chapter 2)). One of their most convenient features is their ability to capture complex dependence structures by means of linear associations (measured in terms of covariances and correlations). Nevertheless, owing to such complex structures, the probabilistic analysis of nonlinear functionals of a Gaussian process often becomes considerably difficult. One such nonlinear functional is the corresponding all-time maximum, which arises in communication applications and the analysis of queues with Gaussian input (see, for instance, Addie *et al.* (1999)). This is the main focus of this paper.

1.1. Problem setup

In order to provide a concrete framework for further developments, let us first introduce some notation. Let $\{X_k : k \in \mathbb{N}\}$ be a discrete-time Gaussian process with mean 0 and variance

Received 5 February 2010; revision received 9 December 2010.
[*] Postal address: Department of Industrial Engineering and Operations Research, Columbia University, 321 S. W. Mudd, 500 West 120th Street, New York, NY 10027, USA.
[**] Email address: jose.blanchet@columbia.edu
[***] Email address: cl2856@columbia.edu

$\sigma_k^2 = \mathrm{var}(X_k)$. We consider a nonnegative drift sequence $\{\mu_k : k \geq 1\}$. Our main interest is to apply simulation to efficiently estimate the tail probability

$$\alpha(b) = \mathrm{P}\Big(\max_{1 \leq k < \infty} (X_k - \mu_k) > b\Big)$$

as $b \nearrow \infty$. We refer to this asymptotic environment as 'large-buffer scaling' (the terminology is borrowed from the queueing setting; see Mandjes (2007, p. 65)). We will assume that both $\sigma_\cdot$ and $\mu_\cdot$ are regularly varying. In particular, we assume that $\sigma_k = k^{H_\sigma} L_\sigma(k)$ and $\mu_k = k^{H_\mu} L_\mu(k)$, where $0 < H_\sigma < H_\mu < \infty$ and $L_\sigma(\cdot)$ and $L_\mu(\cdot)$ are slowly varying functions at ∞ (i.e. $L_\sigma(ta)/L_\sigma(t) \to 1$ as $t \to \infty$ for all $a > 0$, and similarly for $L_\mu(\cdot)$).

These assumptions cover most cases of applied interest, including the important special case of fractional Gaussian noise with negative linear drift. Note in particular that no restrictions are imposed on the correlation structure nor are the increments of the underlying Gaussian process assumed to be stationary. We assume that $H_\sigma < H_\mu$ because this implies that $\alpha(b) \searrow 0$ as $b \nearrow \infty$. (In the setting of stationary increments we must implicitly assume that $H_\sigma \leq 1$. Since we do not impose any form of stationarity, we do not require this assumption.) In fact, the convergence to 0 is exponentially fast in a positive power of b as $b \nearrow \infty$ (see, for instance, Dębicki (1999) or Theorem 1 below). On the other hand, if $H_\sigma > H_\mu$ and $\mu_k \geq 0$, we obtain

$$\alpha(b) \geq \max_{1 \leq k < \infty} \mathrm{P}(X_k - \mu_k > b) = \max_{1 \leq k < \infty} \left[1 - \Phi\left(\frac{b + \mu_k}{\sigma_k}\right)\right] = \frac{1}{2}.$$

(Here and throughout the rest of the paper, we use $\Phi(x)$ to denote the cumulative distribution function (CDF) of a standard Gaussian random variable evaluated at x.)

One of our goals is to develop an algorithm for estimating $\alpha(b)$ that takes at most a polynomial number (in b) of function evaluations and that yields an estimate that is guaranteed to be close in relative terms to $\alpha(b)$. In addition, we are interested in developing efficient (polynomial time in b) algorithms for generating samples that are close in total variation to the sample path $(X_k : k \leq T(b))$ given that $T(b) < \infty$, where $T(b) := \inf\{k \geq 1 : X_k - \mu_k > b\}$. Finally, we are also concerned with estimation of the tail of the maximum of the superposition of a large number, n, of independent and identically distributed (i.i.d.) Gaussian sources; this asymptotic environment is commonly referred to as the 'many-sources scaling' as $n \nearrow \infty$. In this scaling we replace σ_k^2, μ_k by $n\sigma_k^2$, $n\mu_k$ and b by bn, and we perform our complexity analysis as a function of n.

As a function evaluation, we consider a single addition, a multiplication, a single evaluation of the Gaussian CDF, and the simulation of a single uniform random number. We recognize that these procedures would typically require different numbers of floating-point operations in a computer. However, similar models of computation are often used when dealing with complexity of continuous problems (see, for instance, the discussion in Section 2 of Traub (2003)). Also, any Monte Carlo method in this setting will require simulating Gaussian random variables—a procedure that requires at least evaluating an exponential function if the polar method is used. On the other hand, we can approximate the tail of the Gaussian CDF with a good relative precision uniformly over the real line (no larger than 6×10^{-19} according to Cody (1969)) using rational approximations (which can be computed by evaluating an exponential function and a ratio of two polynomials with degrees no larger than eight). Consequently, we believe that our model of computation, while somewhat coarse, is accurate enough to compare the efficiency of our procedure against alternative Monte Carlo methods when available. Numerical experiments are given in Section 5.

1.2. Asymptotics

There is a rich literature on sharp asymptotic approximations for $\alpha(b)$, both in discrete- and continuous-time settings, and under large-buffer and many-sources scalings. The correlation structure, however, is typically more restricted than what we impose. Pickands (1969) first explored this problem in the context of stationary processes and finite time horizon. The use of the double sum method allows us to obtain refined results in broader environments; see, for instance, the books by Berman (1992, Chapter 9) and Piterbarg (1996, Chapter 4). The text of Adler and Taylor (2007, Chapter 4) contains a comprehensive discussion of asymptotic approximations for Gaussian random fields. For the important example of a queue fed by a fractional Brownian noise, Duffield and O'Connell (1995) first obtained logarithmic asymptotics; Hüsler and Piterbarg (1999) later provided exact asymptotics. Recently, Dieker (2005) extended the exact asymptotics to a more general class of Gaussian processes under four classes of local correlation structures. In the setting of many-sources scaling, Likhanov and Mazumdar (1999) obtained the sharp asymptotics, assuming stationary increments; see also Dębicki and Mandjes (2003) for the continuous-time counterpart. Many-sources asymptotics is also covered by Mandjes (2007).

Sharp asymptotic approximations of $\alpha(b)$ as $b \nearrow \infty$ must rely on the local correlation structure of the process and (as with any such approximation results) there is always an error that might be either difficult to quantify or simply nonnegligible relative to a given precision requirement. In addition, sharp asymptotics often contain constants that are difficult to evaluate, specially in the large-buffer scaling setting where the discrete nature of our formulation gives rise to continuity corrections that even in the case of Gaussian random walks are not entirely straightforward to compute (see, for instance, Chang and Peres (1997)). While sharp asymptotics are rather explicit in the discrete-time many-sources scaling context (see Likhanov and Mazumdar (1999)), we note in our numerical experiments, shown in Section 5, that the asymptotics might incur a substantial error (of the order of 70%), even in situations concerning probabilities of order 10^{-10}. In contrast, our algorithm requires about a second, or even considerably less time, to obtain an estimate with an error of the order of 2%.

1.3. Simulation

The performance of rare-event simulation estimators for overflow probabilities is often quantified according to efficiency notions such as strong or weak efficiency; see Asmussen and Glynn (2007, Chapter 6) or Juneja and Shahabuddin (2006).

In our current Gaussian setting, there are relatively few simulation estimators that can be rigorously quantified in terms of these types of efficiency notion, especially in the large-buffer scaling setting. Huang *et al.* (1999) and Michna (1999) provided two algorithms for queues with fractional Brownian noise. However, it was later proved in Dieker and Mandjes (2006) that their estimators were not efficient. Related literature on rare-event simulation of multivariate Gaussian random variables includes the work of Sadowsky and Bucklew (1990) and Bucklew and Radeke (2003). The algorithms that are closest in spirit to our approach are those of Adler *et al.* (2008), (2010), but an important difference is that they needed to simulate or approximate the whole process of interest. In contrast, as we will see, our algorithm here involves simulating only a random number of components plus an additional 'trimming' procedure. This distinction is particularly useful in our infinite horizon setting.

In the many-sources scaling setting, Dieker and Mandjes (2006) summarized and analyzed several estimators, including an estimator based on a decomposition studied in Boots and Mandjes (2002), another based on dynamic importance sampling ideas developed in Dupuis

https://doi.org/10.1239/jap/1308662639 Published online by Cambridge University Press

and Wang (2004), and a third estimator based on the work of Sadowsky and Bucklew (1990). These three algorithms are also shown to be weakly efficient. Finally, we mention an alternative approach in Giordano *et al.* (2007) called bridge Monte Carlo which shares a common feature with our approach in that both require the construction of a Gaussian bridge. However, the ideas are fundamentally different. First, in contrast to our method, bridge Monte Carlo is not based on importance sampling. Second, and most importantly, bridge Monte Carlo is typically not even weakly efficient.

It is important to point out that most provably efficient simulation procedures for first passage time probabilities of Gaussian processes involve exponential tilting. A typical application of exponential tilting in rare-event simulation involves two steps (see Asmussen and Glynn (2007, Section 6.6)). First, the identification of the most likely path (or 'optimal path') to the rare event, often done in a fluid scale and using large deviation techniques. The optimal path is simply a law of large numbers (or fluid) description of the conditional distribution of the process given the event of interest. This identification step often involves solving a calculus of variations problem based on the associated large deviations rate function of the process of interest (see, for instance, Norros (1999)). The second step involves tracking the optimal path using exponential tilting. In the Gaussian setting this just means simulating sequentially the *increments* of the process with the original variance and correlation structure but with the mean equal to the gradient of the optimal path. It is well known (see the discussion at the end of Section 6.6 of Asmussen and Glynn (2007)) that this two-step procedure is not even guaranteed to yield provably efficient estimators. In contrast, as we will discuss, our sampler (called the 'target bridge sampler' or TBS for short) avoids the problem of attempting to sequentially track the most likely path to level b.

Since we deal both with the large-buffer scaling setting and the many-sources scaling setting separately, we should mention that these two environments are intrinsically different. The main difference is that the most likely time at which overflow occurs remains uniformly bounded as $n \nearrow \infty$ in the many-sources scaling setting but it is unbounded in b in the large-buffer scaling environment. Given this difference, it is not surprising that our results are stronger in the many-sources scaling setting. Let us now list our contributions in a more precise way.

1.4. Contributions

To state our contributions, and throughout the rest of the paper, we will use Landau's notation for the asymptotic behavior of functions. That is, given the functions $f(\cdot), g(\cdot)\colon \mathbb{R}_+^m \to \mathbb{R}$ for some $m \geq 1$, we write $f(x) = o(g(x))$ as $\|x\|_\infty \nearrow \infty$ if $f(x)/g(x) \to 0$ as $\|x\|_\infty \nearrow \infty$ and $f(x) = O(g(x))$ if $|f(x)| \leq cg(x)$ for some $c < \infty$ and all $x \in \mathbb{R}_+^m$.

Our main contributions are as follows.

(i) We propose a provably efficient importance sampling algorithm, *not* based on exponential tilting, that allows us to circumvent the challenge of directly approximating and tracking the most likely path to overflow under an arbitrary correlation structure. As a sanity check, when applied to the Brownian motion in the continuous-time setting, our importance sampling estimator achieves zero variance (see Proposition 1). This coincides with the standard exponential tilting approach in this particular case.

(ii) In the large-buffer setting we provide a simulation estimator for $\alpha(b)$ with $o(b^{1/H_\mu+\xi})$ relative mean squared error for any $\xi > 0$ (i.e. the mean squared error divided by the square of the probability of interest is $o(b^{1/H_\mu+\xi})$ as $b \nearrow \infty$). Moreover, each replication

https://doi.org/10.1239/jap/1308662639 Published online by Cambridge University Press

of our estimator takes $o(b^{3/H_\mu+\xi})$ function evaluations to be produced for any $\xi > 0$ (see Theorem 2 and Proposition 2).

(iii) Also, in the large-buffer scaling context, we provide a sampler that allows us to generate paths that are ε-close in total variation to

$$\mathrm{P}_*(X_1, \ldots, X_{T(b)} \in \cdot) := \mathrm{P}(X_1, \ldots, X_{T(b)} \in \cdot \mid T(b) < \infty), \tag{1}$$

with an expected number of function evaluations which is

$$o(b^{1/H_\mu}(b^{1/H_\mu+\xi} + \log(\varepsilon^{-1})^{1/(H_\mu-H_\sigma+\xi)})^3)$$

as $\max(b, \varepsilon^{-1}) \nearrow \infty$ for any $\xi > 0$ (see Theorem 2 and Proposition 2).

(iv) In the many-sources scaling setting we provide an estimator that is strongly efficient in the sense that the relative mean squared error remains bounded as $n \nearrow \infty$. Moreover, our estimator can be implemented in $O(1)$ function evaluations as $n \nearrow \infty$ (see Theorem 3 and Proposition 2).

1.5. Organization

The rest of the paper is organized as follows. We describe our main strategies and ideas in Section 2. The description and technical analysis of the algorithm for the large-buffer scaling setting is given in Section 3. The case of many-sources scaling is discussed in Section 4. Additional computational issues are discussed in Section 5, together with examples showing the numerical performance of our estimator against other procedures.

2. Basic strategy of target bridge sampling

As we mentioned in the introduction, our method is based on importance sampling. For a review of the importance sampling methodology, the reader may consult the text of Asmussen and Glynn (2007, Section 5.1). It is known that importance sampling dictates that the conditional distribution of the process given the event of interest provides a zero-variance importance sampling distribution for the probability of such an event (see Asmussen and Glynn (2007, p. 128)). In our context, this means that in order to achieve zero variance, we must sample $(X_k : k \geq 1)$ given the event $T(b) < \infty$ (see (1)). Since such a conditional distribution is not directly accessible, the objective is to construct an alternative probability measure, say Q, that suitably mimics the behavior of the zero-variance importance sampling distribution, thereby inducing an estimator with reduced variance.

This intuition is typically exploited in the design of importance sampling algorithms. A standard approach in light-tailed settings (such as our current Gaussian environment) involves constructing Q by sampling the process $X_\cdot$ sequentially using exponential tiltings. In the Gaussian setting this is equivalent to changing the mean of the process at each time conditional on past observations. When the dependence is complex, attempting to track the optimal path to overflow by a sequential mean-shifting procedure could become substantially complicated. We now explain the ideas behind our procedure, which circumvents tracking such an optimal path sequentially in time.

2.1. Target sampling

Gaussian distributions have special features that allow us to deal with complex dependence in a convenient way. Among them the most useful to us is that a family of random variables

https://doi.org/10.1239/jap/1308662639 Published online by Cambridge University Press

that is jointly Gaussian remains jointly Gaussian even after conditioning on specific values of an arbitrary subset of the family. Our construction exploits this particular feature combined with the standard intuition about the zero-variance importance sampling distribution described before.

We start our sampling procedure by placing a point on the 'target set' $\mathcal{T}$, defined as

$$\mathcal{T} = \{(k, X_k) : k \in \{1, 2, \ldots, \}, \ X_k - \mu_k > b\}.$$

This strategy is quite natural given that

$$T(b) < \infty \quad \Longleftrightarrow \quad \mathcal{T} \neq \varnothing.$$

Therefore, if we sample a random point $(\tau, X_\tau) \in \{0, 1, \ldots\} \times [b, \infty)$ according to some procedure and let it be an element of $\mathcal{T}$, this automatically implies that $T(b) < \infty$. We call the sampling of (τ, X_τ) the 'target sampling step'. In some sense, this step can be viewed as an alternative to sampling $T(b)$ directly.

The exact law that we choose to sample (τ, X_τ) is as follows. First, we sample τ according to the probability mass function

$$p_\tau(k) = \frac{\mathrm{P}(X_k - \mu_k > b)}{\sum_{j=1}^{\infty} \mathrm{P}(X_j - \mu_j > b)}. \tag{2}$$

Using Chernoff's bound, we obtain the following bound for the denominator:

$$\sum_{j=1}^{\infty} \mathrm{P}(X_j - \mu_j > b) \leq \sum_{j=1}^{\infty} \exp\left(-\frac{(b + \mu_j)^2}{2\sigma_j^2}\right).$$

Obviously, under our assumption that $H_\mu > H_\sigma$, the last summation is finite. Thus, τ is well defined. An important issue involves the sampling of τ, which might appear to require knowledge of the infinite series in the numerator of (2). However, typically sampling τ can be done efficiently via an acceptance–rejection procedure.

Second, given τ, we sample X_τ according to $\mathrm{P}(X_\tau \in \cdot \mid X_\tau - \mu_\tau > b)$. Simulating X_τ given that $X_\tau > b + \mu_\tau$ can be done via acceptance/rejection with proposal distribution given by $(b + \mu_\tau) + \mathrm{Exp}(1)\sigma_\tau^2/(b + \mu_\tau)$, where $\mathrm{Exp}(1)$ represents an exponential random variable with mean 1. This acceptance–rejection procedure turns out to have an acceptance probability which converges to 1 as $b \nearrow \infty$, so it is quite efficient.

2.2. Bridge sampling

Assume that the target sampling step has been carried out with the output (τ, X_τ). The 'bridge sampling step' proceeds simply by first simulating $X_0, \ldots, X_{\tau-1}$ given (τ, X_τ) under the nominal (original) law, namely $\mathrm{P}(\cdot \mid \tau, X_\tau)$. This is easily done because of the Gaussian property of the process. Since $\tau \geq T(b)$, once we have the path $X_0, \ldots, X_\tau$, we can compute $T(b) = \min(k : 1 \leq k \leq \tau, \ X_k - \mu_k > b)$. The output of the 'bridge sampling step' is simply $(X_0, \ldots, X_{T(b)})$. It is important to note that the path segment $X_{T(b)+1}, \ldots, X_\tau$ is discarded.

The combination of both steps corresponds to our target bridge sampling (TBS) method. Figure 1 illustrates a generic path generated under TBS, with $b = 16$, $T(b) = 13$, and $\tau = 20$. In summary, TBS suitably selects an element from the target set $\mathcal{T}$ which serves as an anchor point for constructing a bridge. Next, in the bridge sampling step, $T(b)$ is computed and we actually discard samples beyond $T(b)$ up to the anchor point.

https://doi.org/10.1239/jap/1308662639 Published online by Cambridge University Press

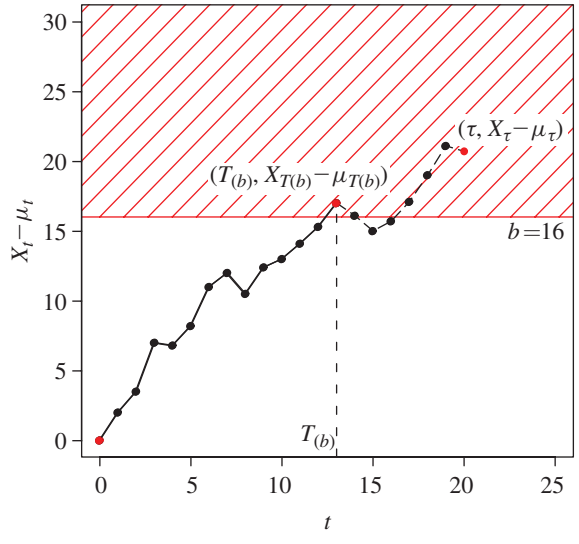

FIGURE 1: Illustration of a path generation under TBS.

The procedure explained above induces a likelihood associated to the sample path $(X_0, \ldots, X_{T(b)})$, and a corresponding likelihood ratio between the probability measure induced by TBS and $\mathrm{P}(\cdot \mid T(b) < \infty)$. We now provide a precise mathematical description of the probability measure induced by TBS, which we will denote by $\mathrm{Q}(\cdot)$. We will use $\mathrm{E}^{\mathrm{Q}}(\cdot)$ for the associated expectation operator and $\mathrm{var}^{\mathrm{Q}}(\cdot)$ for corresponding variances.

Clearly, $T(b) = \min\{k \geq 1 \colon X_k - \mu_k > b\}$ is a stopping time with respect to the filtration $\mathcal{F}_k = \sigma(\{X_1, \ldots, X_k\})$, $k \geq 1$, generated by the process $(X_n \colon n \geq 1)$. Set $\mathcal{F} = \sigma(\bigcup_{k \geq 1} \mathcal{F}_k)$. The stopped σ-field associated to $T(b)$ is defined as $\mathcal{F}_{T(b)} = \sigma\{A \in \mathcal{F} \colon A \cap \{T(b) = k\} \in \mathcal{F}_k\}$. Because $\tau < \infty$, clearly the importance sampling probability measure $\mathrm{Q}(\cdot)$ is defined on $\mathcal{F}_{T(b)}$ in such a way that $\mathrm{Q}(T(b) < \infty) = 1$. Moreover, translating in mathematical terms the description of TBS given earlier we see that, for any Borel sets $B_1, \ldots, B_k$, we have

$$
\begin{aligned}
&\mathrm{Q}(X_1 \in B_1, \ldots, X_k \in B_k,\ T(b) = k) \\
&\quad = \sum_{j=1}^{\infty} \mathrm{P}(X_1 \in B_1, \ldots, X_k \in B_k,\ T(b) = k \mid X_j - \mu_j > b) p_\tau(j) \\
&\quad = \sum_{j=k}^{\infty} \mathrm{P}(X_1 \in B_1, \ldots, X_k \in B_k,\ T(b) = k \mid X_j - \mu_j > b) p_\tau(j) \\
&\quad = \sum_{j=k}^{\infty} \frac{\mathrm{P}(X_1 \in B_1, \ldots, X_k \in B_k,\ T(b) = k,\ X_j - \mu_j > b)}{\sum_{n=1}^{\infty} \mathrm{P}(X_n - \mu_n > b)} \\
&\quad = \mathrm{P}(X_1 \in B_1, \ldots, X_k \in B_k)\, \mathbf{1}(T(b) = k) \\
&\qquad \times \frac{\sum_{j=k}^{\infty} \mathrm{P}(X_j - \mu_j > b \mid X_1 \in B_1, \ldots, X_k \in B_k)}{\sum_{n=1}^{\infty} \mathrm{P}(X_n - \mu_n > b)}.
\end{aligned}
$$

https://doi.org/10.1239/jap/1308662639 Published online by Cambridge University Press

The previous equations define $\mathrm{Q}(\cdot)$ throughout $\mathcal{F}_{T(b)}$ and we have

$$\mathbf{1}(T(b)=k)\frac{\mathrm{dP}}{\mathrm{dQ}}(X_1,\ldots,X_k)=\mathbf{1}(T(b)=k)\frac{\sum_{j=1}^{\infty}\mathrm{P}(X_j-\mu_j>b)}{\sum_{j=k}^{\infty}\mathrm{P}(X_j-\mu_j>b\mid X_1,\ldots,X_k)}.$$

Consequently, the importance sampling estimator for $\alpha(b)$ generated by Q is simply

$$L=\frac{\mathrm{dP}}{\mathrm{dQ}}(X_1,\ldots,X_{T(b)})=\frac{\sum_{j=1}^{\infty}\mathrm{P}(X_j-\mu_j>b)}{\sum_{j=T(b)}^{\infty}\mathrm{P}(.X_j-\mu_j>b\mid X_1,\ldots,X_{T(b)})}. \tag{3}$$

Observe that

$$\sum_{j=T(b)}^{\infty}\mathrm{P}(X_j-\mu_j>b\mid X_1,\ldots,X_{T(b)})\geq\mathrm{P}(X_{T(b)}-\mu_{T(b)}>b)=1.$$

Therefore,

$$L\leq\sum_{j=1}^{\infty}\mathrm{P}(X_j-\mu_j>b),\qquad \mathrm{E}^{\mathrm{Q}}L^2\leq\left(\sum_{j=1}^{\infty}\mathrm{P}(X_j-\mu_j>b)\right)^2.$$

As a consequence, the behavior of the relative mean squared error of the estimator is upper bounded by the ratio $\sum_{j=1}^{\infty}\mathrm{P}(X_j-\mu_j>b)/\alpha(b)$, which grows graciously (polynomially bounded in b as Theorem 1 at the end of this section indicates) and sometimes it even stays bounded.

2.3. Exact sampling

An additional observation that is useful for the design of exact sampling procedures is that in our case the likelihood ratio, L, is bounded and, therefore, in principle it can be used to construct an acceptance–rejection procedure. More precisely, using our notation for $\mathrm{P}_*(\cdot)=\mathrm{P}(\cdot\mid T(b)<\infty)$ introduced in (1), we obtain

$$\frac{L}{\alpha(b)}=\frac{\mathrm{dP}_*}{\mathrm{dQ}}(X_1,\ldots,X_{T(b)})\leq\frac{\sum_{n=1}^{\infty}\mathrm{P}(X_j-\mu_j>b)}{\alpha(b)}. \tag{4}$$

Therefore, we can use $\mathrm{Q}(\cdot)$ as our proposal distribution, then generate a random variable U uniformly distributed over the interval $[0,1]$ (independent of L) and accept if

$$L\leq U\sum_{k=1}^{\infty}\mathrm{P}(X_k-\mu_k>b).$$

A sample path accepted under this procedure follows the law P_*. The procedure has acceptance ratio $\sum_{k=1}^{\infty}\mathrm{P}(X_k-\mu_k>b)/\alpha(b)$, and, therefore, the acceptance probability per sample is equal to $\alpha(b)/\sum_{k=1}^{\infty}\mathrm{P}(X_k-\mu_k>b)$. As a consequence, as we will show in Theorem 1 below, the acceptance probability under this acceptance–rejection procedure goes to 0 polynomially as $b\nearrow\infty$. Consequently, the expected number of proposed sample paths required to obtain a successful realization from $\mathrm{P}_*(\cdot)$ grows at most polynomially in b.

https://doi.org/10.1239/jap/1308662639 Published online by Cambridge University Press

2.4. An insightful case: Brownian motion

Here we consider the special case in which the underlying Gaussian process is Brownian motion. We do this exercise to verify that our method is not worse than exponential tilting, which in particular describes the optimal importance sampling distribution (in terms of variance minimization) for computing $\alpha(b)$. Let $X_\cdot$ be a standard Brownian motion under P, and set $\mu_t = \mu t$ for $t \geq 0$. It is well known that $\alpha(b) = \mathrm{P}(\max_{t\geq 0}(X_t - \mu t) > b) = \exp(-2\mu b)$.

Now we apply the ideas behind TBS to this continuous-time setting. As a simple analogy to the procedure we described above for the discrete Gaussian processes, we proceed by first sampling a time τ with density $f_\tau(\cdot)$ proportional to $\mathrm{P}(X_\cdot > b)$. Then, given τ, we sample the Brownian bridge $(X_s - \mu s\colon 0 \leq s \leq t)$ conditional on the observed value X_τ, which in turn has been sampled from the distribution of X_τ given that $X_\tau - \mu\tau > b$. Following a simple discretization procedure we find that TBS gives a likelihood ratio for the generated path $(X_s\colon 0 \leq s \leq T(b))$ equal to

$$L = \frac{\int_0^\infty \mathrm{P}(X_s - \mu s > b)\,\mathrm{d}s}{\int_{T(b)}^\infty \mathrm{P}(X_u - \mu t > b \mid X_{T(b)})\,\mathrm{d}t} = \frac{\int_0^\infty \mathrm{P}(X_s - \mu s > b)\,\mathrm{d}s}{\int_0^\infty \mathrm{P}(X_t - \mu t > 0)\,\mathrm{d}t},$$

which has zero variance because the right-most expression is nonrandom and depends on neither τ nor $T(b)$. Because the sampler is unbiased, we must have $L = \exp(-2\mu b)$. This expression can also be directly checked by obtaining the Laplace transform of $\int_0^\infty \mathrm{P}(X_s - \mu s > b)\,\mathrm{d}s$ as a function of b.

The overall outcome is that in the Brownian motion setting TBS yields the zero-variance importance sampling distribution. In turn, as discussed at the beginning of the section, such a distribution is equal to the conditional distribution of $(X_s - \mu s\colon 0 \leq s \leq T(b))$ given that $T(b) < \infty$, which is known to be described by the process $(B_s + \mu s\colon 0 \leq s \leq T_1(b))$, where $(B_s\colon s \geq 0)$ is a standard Brownian motion and $T_1(b) := \inf\{s \geq 0\colon B_s + \mu s > b\}$. We record this observation as a proposition.

Proposition 1. *Under the probability measure* Q *generated by the sampling strategy described in the current section for the Brownian motion case, it follows that the law of* $(X_s - \mu s\colon 0 \leq s \leq T(b))$ *is just that of* $(B_s + \mu s\colon 0 \leq s \leq T_1(b))$. *Therefore,* Q *is the zero-variance importance sampling probability measure.*

It is obviously not surprising that we can obtain a good simulation estimator for $\alpha(b)$ in the context of Brownian motion. Nevertheless, the fact that TBS recovers the optimal sampler in this case is, we believe, remarkable. This is specially so given that we did not even attempt to describe the process generated by TBS sequentially in time; yet, the typical description of Brownian motion with negative drift conditioned on reaching level b based on exponential tilting exploits the Markovian nature of the conditional process. It is because we steer the system to the rare event of interest directly that our approach is applicable to any discrete-time Gaussian process that meets our (mild) regularity conditions. In principle, our approach only requires (i) explicit knowledge of the marginal distributions, (ii) that conditional distributions given marginals can be sampled, and (iii) that conditional marginal probabilities can also be computed. These features actually appear in other processes of interest beyond the Gaussian case. For instance, Blanchet *et al.* (2009) adapted the ideas here to estimate loss probabilities in a Markov-modulated M/G/s queue.

https://doi.org/10.1239/jap/1308662639 Published online by Cambridge University Press

2.5. Main result of the section

We conclude this section with a summary of the mean squared error properties behind an estimator based on $\mathrm{Q}(\cdot)$ based on the likelihood ratio (3).

Theorem 1. *We have* $\mathrm{E}^{\mathrm{Q}} L = \alpha(b)$ *and, for any* $\xi > 0$,

$$\frac{\mathrm{var}^{\mathrm{Q}}(L)}{\mathrm{P}(\max_{k\geq 1} X_k - \mu_k > b)^2} = o(b^{2/H_\mu + 2\xi}) \quad \text{as } b \to \infty.$$

Moreover, the acceptance ratio of the procedure described in (4) satisfies

$$\frac{\sum_{k=1}^{\infty} \mathrm{P}(X_k - \mu_k > b)}{\alpha(b)} = o(b^{1/H_\mu + \xi}) \quad \text{as } b \to \infty$$

for any $\xi > 0$.

We will prove this theorem at the end of the next section as it follows directly from technical results that are required to deal with the truncation of the infinite sums appearing in the definition of L.

3. Large-buffer scaling and time truncation

In the previous section we described the main conceptual ideas behind our importance sampling strategy via the nonexponential change of measure Q. We noted, however, that from a simulation standpoint, the likelihood ratio estimator L resulting from Q cannot be directly computed because it requires the exact computation of certain infinite sums. In this section we will study two ways to address this issue.

The first way to remedy this situation is via another randomization step as we now explain. Given the path $(X_1, \ldots, X_{T(b)})$ obtained by TBS, define the function

$$\widetilde{L}(n; X_1, \ldots, X_{T(b)}) = \frac{\mathrm{P}(X_{n-T(b)+1} - \mu_{n-T(b)+1} > b \mid X_1, \ldots, X_{T(b)})}{\mathrm{P}(X_n - \mu_n > b \mid X_1, \ldots, X_{T(b)})}, \qquad k \geq T(b).$$

Also, given $(X_1, \ldots, X_{T(b)})$, define N with probability mass function

$$p_N(k) = \frac{\mathrm{P}(X_k - \mu_k > b \mid X_1, \ldots, X_{T(b)})}{\sum_{k=T(b)}^{\infty} \mathrm{P}(X_k - \mu_k > b \mid X_1, \ldots, X_{T(b)})}, \qquad k \geq T(b).$$

Generating paths according to $p_N(\cdot)$ can be easily done via acceptance/rejection. Then note that $\widetilde{L}(N; X_1, \ldots, X_{T(b)})$ is an unbiased estimator of L given $(X_1, \ldots, X_{T(b)})$. Indeed,

$$\begin{aligned}
&\mathrm{E}(\widetilde{L}(N; X_1, \ldots, X_{T(b)}) \mid X_1, \ldots, X_{T(b)}) \\
&\quad = \sum_{k=T(b)}^{\infty} \frac{\mathrm{P}(X_{k-T(b)+1} - \mu_{k-T(b)+1} > b \mid X_1, \ldots, X_{T(b)})}{\mathrm{P}(X_k - \mu_k > b \mid X_1, \ldots, X_{T(b)})} \\
&\qquad \times \frac{\mathrm{P}(X_k - \mu_k > b \mid X_1, \ldots, X_{T(b)})}{\sum_{k=T(b)}^{\infty} \mathrm{P}(X_k - \mu_k > b \mid X_1, \ldots, X_{T(b)})} \\
&\quad = \frac{\sum_{j=1}^{\infty} \mathrm{P}(X_j - \mu_j > b)}{\sum_{j=T(b)}^{\infty} \mathrm{P}(X_j - \mu_j > b \mid X_1, \ldots, X_{T(b)})} \\
&\quad = L.
\end{aligned}$$

https://doi.org/10.1239/jap/1308662639 Published online by Cambridge University Press

Although this approach provides an unbiased estimator, it will introduce some variance due to the additional randomization step.

The second way to deal with the issue of evaluating the infinite series in L is to introduce a truncation, although this inevitably induces a bias in the estimator. This truncation technique, used in Dieker and Mandjes (2006), is the one that we will consider in the rest of the section. Obviously, for any time horizon $t^+(b)$, the truncated estimator

$$\alpha_{t^+(b)}(b) = \mathrm{P}\Big(\max_{1\le k\le t^+(b)} (X_k - \mu_k) \ge b\Big) \le \alpha(b). \tag{5}$$

Therefore, $t^+(b)$ has to be chosen sufficiently large to guarantee that $\alpha(b) \le \alpha_{t^+(b)}(b)(1+\varepsilon)$ for some small $\varepsilon > 0$ fixed.

A crucial role in our analysis is played by a time that minimizes the decay rate of $\mathrm{P}(X_k - \mu_k > b)$. Let $k^*(b)$ be any optimizer of such a rate of decay, in particular,

$$k^*(b) \in \arg\min_{k\in\mathbb{N}} g(k; b),$$

where $g(k; b) := (b + \mu_k)/\sigma_k$. Clearly, we have

$$\mathrm{P}(X_{k^*(b)} - \mu_{k^*(b)} > b) \le \alpha(b) \le \alpha_{t^+(b)}(b) + \sum_{k=t^+(b)+1}^{\infty} \mathrm{P}(X_k - \mu_k > b).$$

In order to guarantee that the infinite sum in the previous display is small relative to the lower bound on $\alpha(b)$, we need the following lemma that establishes a bound on $g(k^*(b); b)$. The proof is given at the end of the section.

Lemma 1. *For any* $0 < \delta < \min\{(H_\mu - H_\sigma)/2, H_\sigma\}$, *there exists* $M(\delta) \in \mathbb{N}$ *such that, for all* $k \ge M(\delta)$,

$$k^{H_\mu - \delta} \le \mu_k \le k^{H_\mu + \delta}, \qquad k^{H_\sigma - \delta} \le \sigma_k \le k^{H_\sigma + \delta},$$

and

$$g(k^*(b); b) \le h(b) := \begin{cases} \dfrac{b + M(\delta)^{H_\mu+\delta}}{M(\delta)^{H_\sigma-\delta}} & \text{if } b < \dfrac{H_\mu - H_\sigma + 2\delta}{H_\sigma - \delta} M(\delta)^{H_\mu+\delta}, \\ c(\delta; H_\sigma, H_\mu) b^{(H_\mu - H_\sigma + 2\delta)/(H_\mu+\delta)} & \text{otherwise}, \end{cases}$$

where

$$c(\delta; H_\sigma, H_\mu) = \left(\frac{H_\sigma - \delta}{H_\mu - H_\sigma + 2\delta}\right)^{-(H\sigma-\delta)/(H_\mu+\delta)} \left(1 + \frac{(H_\sigma - \delta)2^{H_\mu - H_\sigma + 2\delta}}{H_\mu - H_\sigma + 2\delta}\right)$$

is a constant independent of b.

Using $h(b)$, it turns out, as we will explain in the proof of Theorem 2 below, that $t^+(b)$ in (5) can be specified as

$$t^+(b, \varepsilon) := \left\lceil \lambda \left(h(b) + \frac{1}{h(b)} \left| \log\left(\frac{(1 - h(b)^{-2}) q \varepsilon}{2^{\eta-1}\lambda} \right) \right| \right)^{1/q} \right\rceil, \tag{6}$$

where $q = (H_\mu - H_\sigma - 2\delta)$, $\eta = \max\{1, 1/2q\}$, $\lambda = \Gamma(1+\eta)$, and $\varepsilon > 0$ was introduced

https://doi.org/10.1239/jap/1308662639 Published online by Cambridge University Press

right after (5). To ease the notation, we will simply write $t^+(b)$ rather than $t^+(b,\varepsilon)$. Note that by choosing δ sufficiently small we ensure that

$$t^+(b) = o(b^{1/H_\mu+\xi} + \log(\varepsilon^{-1})^{1/(H_\mu - H_\sigma + 2\xi)}) \tag{7}$$

as $\max(b, \varepsilon^{-1}) \nearrow \infty$ for any fixed $\xi > 0$.

We then have the following algorithm for generating samples of an unbiased estimator for $\alpha_{t^+(b)}(b)$. In the sequel, we use the notation $\mathrm{Q}_0(\cdot)$, $\mathrm{E}^{\mathrm{Q}_0}(\cdot)$, and $\mathrm{var}^{\mathrm{Q}_0}(\cdot)$ to respectively denote the probability measure, expectation operator, and variance induced by the importance sampling distribution in Algorithm 1.

Algorithm 1. 1. Set $t^+(b)$ according to (6).

2. *Targeting*.

- Sample τ according to the probability mass function

$$p_\tau(k) = \frac{\mathrm{P}(X_k > b)}{\sum_{j=1}^{t^+(b)} \mathrm{P}(X_j > b)}.$$

- Given τ, sample X_τ according to the law $\mathrm{P}(X_\tau \le \cdot \mid X_\tau > b + \mu_\tau)$.

3. *Bridging*. Given X_τ, sample the Gaussian bridge $X_1, X_2, \ldots, X_{\tau-1} \mid X_\tau$ from the nominal (original) distribution.

4. Find $T(b) = \min\{j \ge 1 \colon X_j - \mu_j > b\}$.

5. Compute and output the likelihood estimator

$$L_{t^+(b)} = \frac{\sum_{j=1}^{t^+(b)} \mathrm{P}(X_j - \mu_j > b)}{\sum_{j=T(b)}^{t^+(b)} \mathrm{P}(X_j - \mu_j > b \mid X_1, \ldots, X_{T(b)})}.$$

As explained in Section 2, a companion algorithm based on acceptance/rejection can be obtained to generate exact samples according to $\mathrm{P}(\cdot \mid T(b) \le t^+(b))$, which in turn is ε-close in total variation to $\mathrm{P}(\cdot \mid T(b) < \infty)$. Indeed, note that, for any measurable set $A \in \mathcal{F}_{T(b)}$, we obtain

$$\begin{aligned}
\mathrm{P}(A \mid T(b) < \infty) &\le \frac{\mathrm{P}(A, T(b) \le t^+(b))}{\alpha(b)} + \frac{\alpha(b) - \alpha_{t^+(b)}(b)}{\alpha(b)} \\
&\le \frac{\mathrm{P}(A, T(b) \le t^+(b))}{\alpha_{t^+(b)}(b)} + 1 - \frac{\alpha_{t^+(b)}(b)}{\alpha(b)} \\
&\le \mathrm{P}(A \mid T(b) \le t^+(b)) + \varepsilon.
\end{aligned}$$

Also,

$$\begin{aligned}
\mathrm{P}(A \mid T(b) \le t^+(b)) &= \frac{\mathrm{P}(A, T(b) \le t^+(b))}{\alpha_{t^+(b)}(b)} \\
&\le \frac{\mathrm{P}(A, T(b) < \infty)}{\alpha_{t^+(b)}(b)} \\
&\le \mathrm{P}(A \mid T(b) < \infty) + \varepsilon.
\end{aligned}$$

https://doi.org/10.1239/jap/1308662639 Published online by Cambridge University Press

Therefore,

$$\sup_{A\in\mathcal{F}_{T(b)}} |\mathrm{P}(A \mid T(b) \le t^+(b)) - \mathrm{P}(A \mid T(b) < \infty)| \le \varepsilon.$$

We now state explicitly our algorithm for generating exact samples from

$$\mathrm{P}((X_1, \ldots, X_{T(b)}) \in \cdot \mid T(b) \le t^+(b)).$$

Algorithm 2. 1. Set $t^+(b)$ according to (6).

2. Sample $X_1, \ldots, X_{T(b)}$ and calculate its corresponding likelihood ratio $L_{t^+(b)}$ according to Algorithm 1.

3. Simulate $U \sim \mathrm{Unif}[0, 1]$. If $L_{t^+(b)} \le U \sum_{j=1}^{t^+(b)} \mathrm{P}(X_j - \mu_j > b)$, accept and output path $X_1, \ldots, X_{T(b)}$. Otherwise, go back to step 2.

The next result summarizes the statistical properties in terms of the relative mean squared error of the estimator $L_{t^+(b)}$ and the expected number of proposals required to terminate Algorithm 2 (recall the definition of $\mathrm{Q}_0(\cdot)$ given immediately before Algorithm 1).

Theorem 2. *The relative bias of $L_{t^+(b)}$ is less than ε, that is,*

$$0 \le 1 - \frac{\mathrm{E}^{\mathrm{Q}_0}(L_{t^+(b)})}{\alpha(b)} = 1 - \frac{\alpha_{t^+(b)}(b)}{\alpha(b)} < \varepsilon.$$

Moreover, for each $\xi > 0$, we have

$$\frac{\mathrm{var}^{\mathrm{Q}_0}(L_{t^+(b)})}{\alpha_{t^+(b)}(b)^2} = o(b^{2/H_\mu + 2\xi})$$

as $b \nearrow \infty$ and the number of proposals required to terminate Algorithm 2 is geometrically distributed with mean $o(b^{1/H_\mu+\xi})$ as $b \nearrow \infty$.

Before we provide the proof of Lemma 1 and Theorem 2, we need the following two results. The first result is a well-known bound on the Gaussian CDF (see Durrett (2004, p. 6)).

Lemma 2. *For any positive x and the standard normal CDF $\Phi(\cdot)$, we have*

$$\frac{1}{\sqrt{2\pi}}(x^{-1} - x^{-3}) \exp\left(-\frac{1}{2}x^2\right) \le 1 - \Phi(x) \le \frac{1}{\sqrt{2\pi}} x^{-1} \exp\left(-\frac{1}{2}x^2\right).$$

The next elementary bound is similar to Lemma 2.1 and Lemma 2.2 of Dieker and Mandjes (2006). However, because our primary target is the large-buffer scaling setting and the estimates in Dieker and Mandjes (2006) are derived with the case of a large number of sources in mind, we provide a slightly different result which is an adaptation of the bounds obtained in Alzer (1997).

Lemma 3. *For any time $T \ge 1$ and parameters $q, C > 0$, we have*

$$\int_T^\infty \exp(-Ct^q)\,\mathrm{d}t < \frac{\lambda}{qC^\eta} \exp\left(-C\left(\frac{T}{\lambda}\right)^q\right),$$

where $\eta = \max\{1, 1/q\}$ and $\lambda = \Gamma(1 + \eta) \ge 1$.

https://doi.org/10.1239/jap/1308662639 Published online by Cambridge University Press

Proof. (i) When $q \geq 1$ and $T \geq 1$, using a simple variable substitution,

$$\begin{aligned}\int_T^\infty \exp(-Ct^q)\,\mathrm{d}t &\leq \frac{1}{qC^{1/q}} \int_{CT^q}^\infty s^{1/q-1}\exp(-s)\,\mathrm{d}s \\ &\leq \frac{T^{1-q}}{qC}\int_{CT^q}^\infty \exp(-s)\,\mathrm{d}s \\ &\leq \frac{1}{qC}\exp(-CT^q).\end{aligned}$$

(ii) When $0 < q < 1$, the main corollary in Alzer (1997) states that

$$\frac{1}{\Gamma(1+1/q)}\int_x^\infty \exp(-t^q)\,\mathrm{d}t < 1 - \left[1 - \exp\left(-\left[\Gamma\left(1+\frac{1}{q}\right)\right]^{-q} x^q\right)\right]^{1/q}.$$

Therefore, we have

$$\begin{aligned}\int_T^\infty \exp(-Ct^q)\,\mathrm{d}t &= C^{-1/q}\int_{C^{1/q}T}^\infty \exp(-s^q)\,\mathrm{d}s \\ &< C^{-1/q}\lambda\left[1 - \left(1 - \exp\left(-C\left(\frac{T}{\lambda}\right)^q\right)\right)^{1/q}\right] \\ &< C^{-1/q}\lambda\left[1 - \left(1 - \frac{\exp(-C(T/\lambda)^q)}{q}\right)\right] \\ &= \frac{\lambda}{qC^{1/q}}\exp\left(-C\left(\frac{T}{\lambda}\right)^q\right),\end{aligned}$$

where $\lambda = \Gamma(1+1/q) > 1$. Combining (i) and (ii), we obtain the result.

Now we provide the proofs of Theorem 1, Lemma 1, and Theorem 2. We start with Lemma 1.

Proof of Lemma 1. Since $\sigma_\cdot$ and $\mu_\cdot$ are regularly varying, the first set of inequalities are from the well-known Potter's bound, i.e. for all $\delta < (H_\mu - H_\sigma)/2$, the exists $M(\delta) \in \mathbb{N}$ sufficiently large such that, for all $k \geq M(\delta)$,

$$k^{H_\mu - \delta} \leq \mu_k \leq k^{H_\mu+\delta}, \qquad k^{H_\sigma-\delta} \leq \sigma_k \leq k^{H_\sigma+\delta}, \qquad g(k;b) \leq \frac{b + k^{H_\mu+\delta}}{k^{H_\sigma-\delta}} =: g_U(k;b).$$

Therefore, $g(k^*(b);b) = \min_{1\leq k\leq\infty} g(k;b) \leq \min_{M(\delta)\leq k\leq\infty} g_U(k;b)$.

Since

$$g_U'(t;b) = \frac{(H_\mu - H_\sigma + 2\delta)t^{H_\mu+\delta} - (H_\sigma - \delta)b}{t^{H_\sigma-\delta+1}} \tag{8}$$

has only one single root in $\mathbb{R}^+$, $g_U(\cdot;b)$ is unimodal on $\mathbb{R}^+$. Define

$$t_U^*(b) := \arg\min_{t\in\mathbb{R}} g_U(t;b);$$

then, clearly,

$$\min_{M(\delta)\leq k\leq\infty} g_U(k;b) \leq \begin{cases} g_U(M(\delta);b) & \text{if } t_U^*(b) < M(\delta), \\ g_U(t_U^*(b)+1;b) & \text{otherwise.}\end{cases} \tag{9}$$

https://doi.org/10.1239/jap/1308662639 Published online by Cambridge University Press

Using (8), we have

$$t_U^*(b) = \left(\frac{H_\sigma - \delta}{H_\mu - H_\sigma + 2\delta} b\right)^{1/(H_\mu+\delta)}.$$

Therefore, when $b \geq ((H_\mu - H_\sigma + 2\delta)/(H_\sigma - \delta))M(\delta)^{H_\mu+\delta}$, we have $t_U^*(b) \geq M(\delta) \geq 1$ and

$$\begin{aligned} g_U(t_U^*(b)+1; b) &< \frac{b}{(t_U^*)^{H\sigma-\delta}} + (t_U^*+1)^{H_\mu-H_\sigma+2\delta} \\ &< \frac{b}{(t_U^*)^{H\sigma-\delta}} + (2t_U^*)^{H_\mu-H_\sigma+2\delta} \\ &= \left(\frac{H_\sigma - \delta}{H_\mu - H_\sigma + 2\delta}\right)^{-(H\sigma-\delta)/(H_\mu+\delta)} \\ &\quad \times \left(1 + \frac{(H_\sigma - \delta)2^{H_\mu-H_\sigma+2\delta}}{H_\mu - H_\sigma + 2\delta}\right) b^{(H_\mu-H_\sigma+2\delta)/(H_\mu+\delta)} \\ &= c(\delta; H_\sigma, H_\mu) b^{(H_\mu-H_\sigma+2\delta)/(H_\mu+\delta)}. \end{aligned}$$

Substituting this bound into (9) we obtain

$$g(k^*(b); b) \leq \begin{cases} \dfrac{b + M(\delta)^{H_\mu+\delta}}{M(\delta)^{H_\sigma-\delta}} & \text{if } b < \dfrac{H_\mu - H_\sigma + 2\delta}{H_\sigma - \delta} M(\delta)^{H_\mu+\delta}, \\ c(\delta; H_\sigma, H_\mu) b^{(H_\mu-H_\sigma+2\delta)/(H_\mu+\delta)} & \text{otherwise.} \end{cases}$$

Now we are ready to prove the two theorems. The proofs are actually very similar in nature. Here we choose to prove Theorem 2 and structure the proof of Theorem 1 as a simple corollary of the proof of Theorem 2.

Proof of Theorem 2. First of all, it is obvious that $L_{t^+(b)}$ is an unbiased estimator of $\alpha_{t^+(b)}(b)$. Therefore,

$$\begin{aligned} \frac{|\mathrm{P}(\max_{k\geq 1}(X_k - \mu_k) \geq b) - \mathrm{E}\, L_{t^+(b)}|}{\mathrm{P}(\max_{k\geq 1}(X_k - \mu_k) \geq b)} &\leq \frac{\mathrm{P}(\max_{k> t^+(b)}(X_k - \mu_k) \geq b)}{\mathrm{P}(\max_{k\geq 1}(X_k - \mu_k) \geq b)} \\ &\leq \frac{\sum_{k=t^+(b)+1}^{\infty} \mathrm{P}(X_k - \mu_k \geq b)}{\mathrm{P}(X_{k^*} - \mu_{k^*} \geq b)}. \end{aligned}$$

For the numerator, applying Lemma 2, we obtain

$$\sum_{k=t^+(b)+1}^{\infty} \mathrm{P}(X_k - \mu_k \geq b) \leq \frac{1}{\sqrt{2\pi}} \sum_{k=t^+(b)+1}^{\infty} \exp\left(-\frac{1}{2} g(k; b)^2\right) \Big/ g(k; b). \tag{10}$$

Now, similar to the proof of Lemma 1, we have

$$g_L(k; b) := b^{H_\mu - H_\sigma - 2\delta} \leq g(k; b) \quad \text{for any } k \geq M(\delta).$$

Furthermore, recalling the definition of $t^+(b)$ in (1), we have

$$g_L(t^+(b); b) = t^+(b)^{1/q} > h(b) > g_L(M(\delta)), \qquad t^+(b) > M(\delta).$$

https://doi.org/10.1239/jap/1308662639 Published online by Cambridge University Press

Therefore, (10) can be further bounded by

$$\frac{1}{\sqrt{2\pi}h(b)}\sum_{k=t^+(b)+1}^{\infty}\exp\left(-\frac{1}{2}g_L(k;b)^2\right)\le\frac{1}{\sqrt{2\pi}h(b)}\int_{t^+(b)}^{\infty}\exp\left(-\frac{1}{2}g_L(t;b)^2\right)\mathrm{d}t. \quad (11)$$

Now, applying Lemma 3 to (11), we obtain

$$\sum_{k=t^+(b)+1}^{\infty}\mathrm{P}(X_k-\mu_k\ge b)<\frac{2^{\eta-1}\lambda}{H_\mu-H_\sigma-2\delta}\frac{\exp(-g_L(t^+(b)/\lambda)^2/2)}{\sqrt{2\pi}h(b)}.$$

Again, recall the definition of $t^+(b)$ in (1) and the lower bound of $\Phi(\cdot)$ in Lemma 2. We have

$$\begin{aligned}\frac{2^{\eta-1}\lambda}{H_\mu-H_\sigma-2\delta}\frac{\exp(-g_L(t^+(b)/\lambda)^2/2)}{\sqrt{2\pi}h(b)}&<\frac{\exp(-h(b)^2/2)}{\sqrt{2\pi}}(h(b)^{-1}-h(b)^{-3})\varepsilon\\&\le[1-\Phi(h(b))]\varepsilon\\&\le\varepsilon\,\mathrm{P}(X_{k^*}-\mu_{k^*}\ge b).\end{aligned}$$

Equivalently,

$$\frac{\sum_{k=t^+(b)+1}^{\infty}\mathrm{P}(X_k-\mu_k\ge b)}{\mathrm{P}(X_{k^*}-\mu_{k^*}\ge b)}\le\varepsilon.$$

Now let us analyze the squared coefficient of variation of $L_{t^+(b)}$, namely,

$$\begin{aligned}&\frac{\mathrm{var}^{\mathrm{Q}}(L_{t^+(b)})}{\mathrm{P}(\max_{1\le k\le t^+(b)}(X_k-\mu_k)>b)^2}\\&\quad\le\frac{\mathrm{E}^{\mathrm{Q}}(L_{t^+(b)}^2)}{\mathrm{P}(\max_{1\le k\le t^+(b)}(X_k-\mu_k)>b)^2}\\&\quad=\left(\frac{\sum_{k=1}^{t^+(b)}\mathrm{P}(X_k-\mu_k\ge b)}{\mathrm{P}(\max_{1\le k\le t^+(b)}(X_k-\mu_k)>b)}\right)^2\mathrm{E}\left(\frac{1}{\sum_{k=t^*}^{t^+(b)}\mathrm{P}(X_k\ge b\mid X_{[0,t^*]})}\right)^2\\&\quad\le\left(\frac{\sum_{k=1}^{t^+(b)}\mathrm{P}(X_k-\mu_k\ge b)}{\mathrm{P}(\max_{1\le k\le t^+(b)}(X_k-\mu_k)>b)}\right)^2\\&\quad\le t^+(b)^2\\&\quad=O(b^{((H_\mu-H_\sigma+2\delta)/(H_\mu-H_\sigma-2\delta))(2/(H_\mu+\delta))}).\end{aligned}$$

Since we can choose any $\delta<\min\{(H_\mu-H_\sigma)/2,H_\sigma\}$, by sending δ to 0, we obtain

$$\frac{\mathrm{var}^{\mathrm{Q}}(L_{t^+(b)})}{\mathrm{P}(\max_{1\le k\le t^+(b)}(X_k-\mu_k)\ge b)^2}=O(b^{2/H_\mu+\xi})$$

for any $\xi>0$.

Theorem 1 now follows as a simple extension of our proof of Theorem 2.

Proof of Theorem 1. First of all, it is easily seen from the construction of our algorithm that the estimator L is unbiased, so

$$\mathrm{E}^{\mathrm{Q}}L=\alpha(b)=\mathrm{P}\Big(\max_{k\ge1}(X_k-\mu_k)\ge b\Big).$$

https://doi.org/10.1239/jap/1308662639 Published online by Cambridge University Press

On the other hand, in terms of the relative mean squared error, similar to the proof of Theorem 2, we have

$$\begin{aligned}
&\frac{\mathrm{var}^{\mathrm{Q}}(L)}{\mathrm{P}(\max_{k\geq 1}(X_k-\mu_k)\geq b)^2}\\
&\quad\leq \frac{\mathrm{E}^{\mathrm{Q}}(L^2)}{\mathrm{P}(\max_{k\geq 1}(X_k-\mu_k)\geq b)^2}\\
&\quad= \left(\frac{\sum_{k=1}^{\infty}\mathrm{P}(X_k-\mu_k\geq b)}{\mathrm{P}(\max_{k\geq 1}(X_k-\mu_k)\geq b)}\right)^2 \mathrm{E}\left(\frac{1}{\sum_{k=k^*}^{\infty}\mathrm{P}(X_k\geq b|X_{[0,k^*]})}\right)^2\\
&\quad\leq (t^+(b)+\varepsilon)^2\\
&\quad= O(b^{((H_\mu-H_\sigma+2\delta)/(H_\mu-H_\sigma-2\delta))(2/(H_\mu+\delta))})\\
&\quad= O(b^{2/H_\mu+\xi})
\end{aligned}$$

for any $\xi > 0$.

4. Many-sources scaling

In this section we consider the maximum of a Gaussian process that is obtained as a superposition of a large number of independent Gaussian sources. It is commonly encountered in queues with Gaussian input (see Mandjes (2007, Chapter 2)). It turns out that our method is not only applicable to this setting, but in fact it can be shown to be strongly efficient, in the sense that the coefficient of variation of the underlying estimator remains bounded as the probability of interest decreases to 0.

To state the problem in mathematical terms, let us define

$$\tilde{X}_k^{(n)} = \sum_{j=1}^{n} X_k^{(j)},$$

where the processes $(X_k^{(j)}: k \geq 1)$, $j \geq 1$, are i.i.d. copies of the process $(X_k: k \geq 1)$ described in Section 3 (in particular, centered Gaussian processes with variance σ_k^2). We are interested in estimating

$$\tilde{\alpha}(nb) = \mathrm{P}\Big(\max_{1\leq k<\infty}(\tilde{X}_k^{(n)} - n\mu_k) > nb\Big).$$

It is important to emphasize that in this section we mainly concentrate on $\tilde{\alpha}(nb)$ as $n \nearrow \infty$ for fixed b.

If we set $\mu_k = ck$ for some $c > 0$ and $X_k^{(j)}$ has stationary increments, then we recover the setting discussed in Dieker and Mandjes (2006). They applied four methods to this problem. All of them were proved to be weakly efficient (in the sense of subexponential complexity as $n \nearrow \infty$, as mentioned in the introduction). In this particular setting our method, which corresponds to applying Algorithm 1 to the process $(\tilde{X}_k^{(n)}: k \geq 1)$, gives rise to a strongly efficient estimator. That is, the number of function evaluations required to obtain an estimate with a given relative accuracy remains bounded as a function of n.

In order to estimate $\tilde{\alpha}(nb)$, we follow a scheme parallel to that of Section 3. In particular, we define the function $\tilde{g}(k; nb)$ as

$$\tilde{g}(k; nb) := \frac{nb + n\mu_k}{n^{1/2}\sigma_k} = n^{1/2} g(k; b).$$

https://doi.org/10.1239/jap/1308662639 Published online by Cambridge University Press

Because of the factor $n^{1/2}$ appearing in $\tilde{g}k; nb$ it turns out, as our next result shows, that our selection $t^+(b)$ defined in (6) can still be used as a truncation threshold. The proof of the next lemma is given later in this section.

Lemma 4. *For any $\varepsilon > 0$, let $t^+(b)$ be truncation time defined in (6). Then, for any $n \geq 1$, we have*

$$0 \leq 1 - \frac{\mathrm{P}(\max_{1\leq k<t^+(b)}(\tilde{X}_k^{(n)} - n\mu_k) > nb)}{\mathrm{P}(\max_{1\leq k<\infty}(\tilde{X}_k^{(n)} - n\mu_k) > nb)} < \varepsilon.$$

Remark. For the linear drift case, Dieker and Mandjes (2006) obtained a different truncation time which also guarantees at most ε relative bias. Their bound directly exploits the fact that the number of sources is large. As an alternative to our truncation time, we can use their bound, which is readily available in Corollary 2.3 of Dieker and Mandjes (2006). None of these bounds dominate the other one in every problem instance in which they are both applicable, so the reader might simply select the minimum between them.

The previous lemma allows us to concentrate our efforts on estimating

$$\widetilde{\alpha}_{t^+(b)}(nb) = \mathrm{P}\Big(\max_{1\leq k<t^+(b)}(\tilde{X}_k^{(n)} - n\mu_k) > nb\Big)$$

at the price of introducing a relative bias of at most ε. A detailed algorithm, completely analogous to Algorithm 1, is now given for estimating $\widetilde{\alpha}_{t^+(b)}(nb)$.

Algorithm 3. 1. Set $t^+(b)$ according to (6).

2. *Targeting*.

- Sample τ according to the probability mass function

$$p_\tau(k) = \frac{\mathrm{P}(\tilde{X}_k^{(n)} - n\mu_k > nb)}{\sum_{k=1}^{t^+(b)} \mathrm{P}(\tilde{X}_k^{(n)} - n\mu_k > b)} = \frac{\mathrm{P}(n^{1/2}X_k - n\mu_k > nb)}{\sum_{k=1}^{t^+(b)} \mathrm{P}(n^{1/2}X_k - n\mu_k > nb)}.$$

- Given τ, sample $\tilde{X}_\tau^{(n)}$ according to the law $\mathrm{P}(\tilde{X}_\tau^{(n)} \leq \cdot \mid \tilde{X}_\tau^{(n)} > b + n\mu_\tau)$.

3. *Bridging*. Given $\tilde{X}_\tau^{(n)}$, sample the Gaussian bridge $\tilde{X}_1^{(n)}, \tilde{X}_2^{(n)}, \ldots, \tilde{X}_{\tau-1}^{(n)} \mid \tilde{X}_\tau^{(n)}$ from the nominal (original) distribution.

4. Find $T(b) = \min\{k\colon \tilde{X}_k^{(n)} > b + n\mu_k\}$.

5. Compute the likelihood estimator

$$L_n = \frac{\sum_{j=1}^{t^+(b)} \mathrm{P}(\tilde{X}_j^{(n)} - n\mu_j > nb)}{\sum_{j=T(b)}^{t^+(b)} \mathrm{P}(\tilde{X}_j^{(n)} - n\mu_j > nb \mid \tilde{X}_1^{(n)}, \ldots, \tilde{X}_{T(b)}^{(n)})}.$$

We use $\mathrm{Q}_1(\cdot)$ to denote the probability measure corresponding to the importance sampling strategy introduced before, and $\mathrm{E}^{\mathrm{Q}_1}(\cdot)$ and $\mathrm{var}^{\mathrm{Q}_1}(\cdot)$ to denote the corresponding expectation and variance operators, respectively. We have the following result.

Theorem 3. *The estimator L_n for the multisource Gaussian process is strongly efficient for estimating $\widetilde{\alpha}_{t^+(b)}(nb)$ as $n \nearrow \infty$ (assuming that $b = O(1)$ as $n \nearrow \infty$). In particular, we have*

$$\frac{\mathrm{var}^{\mathrm{Q}_1}(L_n)}{\widetilde{\alpha}_{t^+(b)}(nb)} \leq t^+(b)^2.$$

https://doi.org/10.1239/jap/1308662639 Published online by Cambridge University Press

Proof. Just note that

$$\frac{\mathrm{E}^{\mathrm{Q}_1}(L_n^2)}{\mathrm{P}(\max_{k\le t^+(b)} \tilde{X}_k^{(n)} - n\mu_k \ge nb)} \le \left[\frac{\sum_{k=0}^{t^+(b)} \mathrm{P}(\tilde{X}_k^{(n)} - n\mu(k) \ge nb)}{\max_{k\le t^+(b)} \mathrm{P}(\tilde{X}_k^{(n)} - n\mu(k) \ge nb)}\right]^2 \le t^+(b)^2.$$

Before providing the proof of Lemma 4 and closing the section, let us provide some words about an important distinction between the many-sources and the large-buffer environments. A distinction that might explain why the performance of our estimator is different when applied to each of these settings. In the many-sources scaling setting, we can typically compute k^* (independent of n) such that $\tilde{\alpha}(nb) = \mathrm{P}(\tilde{X}_{k^*}^{(n)} - n\mu_{k^*} > nb)(1 + o(1))$ as $n \nearrow \infty$ (see Likhanov and Mazumdar (1999)). So, asymptotically, the infinite horizon problem effectively translates into a finite horizon problem which in fact involves only a single marginal distribution, namely, that of k^*. This situation is not at all applicable to the large-buffer scaling setting.

We close the section with the proof of Lemma 4.

Proof of Lemma 4. Once again, the calculation boils down to

$$\frac{\mathrm{P}(\max_{k>t^+(b)} \sum_{j=1}^n [X_k^{(j)} - \mu_k] > nb)}{\mathrm{P}(\max_{1\le k<\infty} \sum_{j=1}^n [X_k^{(j)} - \mu_k] > nb)} \le \frac{\sum_{k>t^+(b)} \mathrm{P}(\sum_{j=1}^n [X_k^{(j)} - \mu_k] > nb)}{\max_{1\le k<\infty} \mathrm{P}(\sum_{j=1}^n [X_k^{(j)} - \mu_k] > nb)}.$$

Since

$$\mathrm{P}\left(\sum_{j=1}^n [X_k^{(j)} - \mu_k] > nb\right) = 1 - \Phi(\sqrt{n} g(k; b)),$$

we know (once again appealing to Lemma 2) that

$$\max_{1\le k<\infty} \mathrm{P}\left(\sum_{j=1}^n X_k^{(j)} - n\mu_k > nb\right) \ge 1 - \Phi(\sqrt{n} h(b)) \ge (n^{-1/2} h(b)^{-1} - n^{-3/2} h(b)^{-3}) \exp\left(-\frac{n}{2} h(b)^2\right),$$

where $h(\cdot)$ is defined in Lemma 1. Similar to the argument in the proof of Theorem 2, we have

$$\sum_{k=t^+(b)+1}^{\infty} \mathrm{P}\left(\sum_{j=1}^n X_k^{(j)} - n\mu_k > nb\right) < \frac{1}{\sqrt{2\pi n} h(b)} \int_{t^+(b)}^{\infty} \exp\left(-\frac{n}{2} g_L(t)^2\right) \mathrm{d}t.$$

Using Lemmas 1 and 3, we obtain

$$\int_{t^+(b)}^{\infty} \exp\left(-\frac{n}{2} g_L(t)^2\right) \mathrm{d}t < \frac{2^{\eta-1}\lambda}{(H_\mu - H_\sigma - 2\delta) n^\eta} \exp\left(\frac{-n(t^+(b)/\lambda)^{2(H_\mu - H_\sigma - 2\delta)}}{2}\right) < \frac{2^{\eta-1}\lambda}{(H_\mu - H_\sigma - 2\delta) n} \exp\left(\frac{-n(t^+(b)/\lambda)^{2(H_\mu - H_\sigma - 2\delta)}}{2}\right).$$

https://doi.org/10.1239/jap/1308662639 Published online by Cambridge University Press

Consequently, we have

$$\begin{aligned}
&\frac{\mathrm{P}(\max_{k>t^+(b)} \sum_{j=1}^n X_k^{(j)} - n\mu_k > nb)}{\mathrm{P}(\max_{1\le k<\infty} \sum_{j=1}^n X_k^{(j)} - n\mu_k > nb)} \\
&\quad < \frac{2^{\eta-1}\lambda}{(H_\mu - H_\sigma - 2\delta)n} \frac{1}{\sqrt{2\pi n}h(b)} \frac{\exp(-n(t^+(b)/\lambda)^{2(H_\mu - H_\sigma - 2\delta)}/2)}{1 - \Phi(\sqrt{n}h(b))} \\
&\quad < \frac{2^{\eta-1}\lambda}{H_\mu - H_\sigma - 2\delta} \frac{\exp(-n|\log((1-h(b)^{-2})(H_\mu - H_\sigma - 2\delta)\varepsilon/2^{\eta-1}\lambda)|)}{n - h(b)^{-2}} \\
&\quad < \frac{1 - h(b)^{-2}}{n - h(b)^{-2}}\varepsilon \\
&\quad \le \varepsilon.
\end{aligned}$$

5. Complexity analysis per replication and a numerical example

We have proved the statistical efficiency of our estimator (in terms of the relative mean squared error). Nevertheless, the actual computational complexity also involves the number of function evaluations required to implement a single replication of our estimator. In this section we study the number of function evaluations and provide comparisons to other existing algorithms in the setting where there exist alternative methods, namely, in the many-sources scaling setting. As we mentioned in the introduction, in our discussion below a function evaluation corresponds to a single addition, a multiplication, a comparison, the generation of a single Gaussian random variable, and the evaluation of the Gaussian CDF.

We concentrate on Algorithm 1. The highest computational burden corresponds to steps 3 and 4 in terms of generating the sample $(X_k : k \le \tau)$, which involves a matrix inversion (necessary to compute the conditional variance given the observed value X_k when $\tau = k$). Such an inversion procedure requires $O(t^+(b)^3)$ function evaluations. Another contribution corresponds to the calculation of our importance sampling estimator

$$L_{t^+(b)} = \frac{\sum_{k=0}^{t^+(b)} \mathrm{P}(X_t > b)}{\sum_{k=T(b)}^{t^+(b)} \mathrm{P}(X_k \ge b \mid \boldsymbol{X}_{T(b)})},$$

where $\boldsymbol{X}_{T(b)} = (X_1, X_2, \ldots, X_{T(b)})^\top$. A convenient feature is that the numerator is constant across all replications. Therefore, we only consider the evaluation of the denominator, which involves the following conditional Gaussian calculation:

$$\mathrm{E}^{\mathrm{Q}}[X_k] = \boldsymbol{V}_k^T \Sigma_{T(b)}^{-1} \boldsymbol{X}_{T(b)} = \widetilde{\mu}_k, \qquad \mathrm{var}^{\mathrm{Q}}[X_k] = \mathrm{var}[X_k] - \boldsymbol{V}_k^T \Sigma_{T(b)}^{-1} \boldsymbol{V}_k = \widetilde{\sigma}_k.$$

Here

$$\boldsymbol{V}_k = \mathrm{cov}(X_k, \boldsymbol{X}_{T(b)}) \quad \text{and} \quad \Sigma_{T(b)} = \mathrm{cov}(\boldsymbol{X}_{T(b)}, \boldsymbol{X}_{T(b)}).$$

These calculations take $O(t^+(b)^3)$ function evaluations. We summarize our observations in the form of a proposition.

Proposition 2. *The number of function evaluations required to terminate a single replication of Algorithm 1 is $O(t^+(b)^3)$. In turn, as indicated in (7), we have*

$$t^+(b) = o(b^{1/H_\mu + \xi} + \log(\varepsilon^{-1})^{1/(H_\mu - H_\sigma + 2\xi)})$$

for each $\xi > 0$ as $b, \varepsilon^{-1} \nearrow \infty$. Finally, since the acceptance ratio in step 3 of Algorithm 2 is

https://doi.org/10.1239/jap/1308662639 Published online by Cambridge University Press

$O(t^+(b))$, we conclude that the expected number of function evaluations required to generate a single replication of Algorithm 2 is $O(t^+(b)^2)$.

We now test the performance of our target bridge sampler and compare it against other existing methods in the multisource setting. We consider the four algorithms discussed in Dieker and Mandjes (2006). The first algorithm is called *single twist*, which is a direct derivation of the large deviation principle; the second algorithm is called *cut-and-twist*, which is developed in Dieker and Mandjes (2006) based on a partitioning approach of Boots and Mandjes (2002); the third algorithm is called *random twist*, which is an adaptation of Sadowsky and Bucklew (1990); the fourth algorithm is called *sequential twist*, which is based on the ideas of Dupuis and Wang (2004). Among them, cut-and-twist, random twist, and sequential twist were proved to be weakly efficient under the many-sources scaling. Single-twist is shown in Dieker and Mandjes (2006) not to be efficient generally.

Example 1. (*Fractional Gaussian noise.*) If we set $\operatorname{cov}(X_k, X_j) = (k^{2H_\sigma} + j^{2H_\sigma} - |k - j|^{2H_\sigma})/2$ and $\mu_k = k$, then we obtain fractional Gaussian noise increments with negative linear drift.

In particular, in this numerical example, we choose the parameter as $H_\sigma = 0.8$, $b = 0.3$. It is easy to see from the proofs that we can set $\delta = 0$ and $M(0) = 1$. Therefore, for $b \geq (H_\mu - H_\sigma)/H_\sigma = \frac{1}{4}$,

$$h(b) = (2^{-8/5} + 2^{3/5})b^{1/5},$$
$$t^+(b) = \left\lceil \Gamma(3.5)\left(h(b) + \frac{1}{h(b)}\left|\log\left(\frac{(1 - h(b)^{-2})\varepsilon}{5 \cdot 2^{3/2}\Gamma(3.5)}\right)\right|\right)^5 \right\rceil.$$

We implemented our Algorithm 2, which we denote by TBS in Tables 1 and 2. We follow Dieker and Mandjes' (2006) performance comparison criterion. That is, each algorithm is kept running until the output confidence interval shrinks to 20% relative to the estimated value. The estimated value, the number of simulations runs needed for each particular algorithm to achieve the criterion, and the corresponding CPU times are reported in Table 1 and Table 2. We also report the asymptotic approximation obtained in Likhanov and Mazumdar (1999) as LM asymptotics. This asymptotic value corresponds to $[1 - \Phi((0.3 + 0.1 \times k^*) \times n^{1/2}/(k^*)^{0.8})]$ with $k^* = 12$. The simulation runs for the first five rows were obtained using the codes of Ton Dieker, who kindly shared them with us for these experiments. The row Benchmark was

TABLE 1: Simulation results for Example 2 with $n = 300$, $b = 3$, and $H_\sigma = 0.8$.

Algorithm	Cost of each replication	Estimator	Simulation runs	CPU time (in seconds)
Naive	$O(b^3)$	6.12×10^{-4}	627 101	960
Single twist	$O(b^3)$	4.84×10^{-4}	3964	6.32
Cut-and-twist	$O(b^4)$	5.72×10^{-4}	640	641
Random twist	$O(b^3)$	5.50×10^{-4}	3315	5.13
Sequential twist	$O(nb^3)$	6.39×10^{-4}	668	432
TBS	$O(b^3)$	5.85×10^{-4}	27	1.27
Benchmark	—	5.75×10^{-4}	—	—
LM asymptotics	—	1.86×10^{-4}	—	—

https://doi.org/10.1239/jap/1308662639 Published online by Cambridge University Press

TABLE 2: Simulation results for Example 2 with $n = 1000$, $b = 3$, and $H_\sigma = 0.8$.

Algorithm	Cost of each replication	Estimator	Simulation runs	CPU time (in seconds)
Naive	$O(b^3)$	—	—	—
Single twist	$O(b^3)$	1.03×10^{-10}	8132	3.62
Cut-and-twist	$O(b^4)$	1.37×10^{-10}	959	61.8
Random twist	$O(b^3)$	1.38×10^{-10}	3875	2.27
Sequential twist	$O(nb^3)$	1.41×10^{-10}	1365	758
TBS	$O(b^3)$	1.36×10^{-10}	25	0.16
Benchmark	—	1.37×10^{-10}	—	—
LM asymptotics	—	4.08×10^{-11}	—	—

obtained by running 1000 replications of our algorithm. All the codes were run in the same computing environment, AMD 2218HE at 2.6 GHz.

It is seen from the simulation results that the TBS compares favorably to the rest of the algorithms. A single replication TBS is somewhat more expensive, mostly due to the Cholesky factorization procedure which we perform in our implementation but is avoided in the codes of Dieker and Mandjes (2006) because they take advantage of a so-called 'circulant embedding' of the covariance matrix. This type of embedding exploits the stationary structure of the underlying process. The method that we used for sampling τ in Algorithm 2 is also a naive one, without resorting to any efficient acceptance–rejection procedure. We note however that our method achieves a similar level of relative error with fewer than 5% as many replications as the closest competitor. This may not be surprising given the fact that theoretically our algorithm is strongly efficient, while the others are at most weakly efficient. It should be pointed out that in Section 5.5 of Dieker and Mandjes (2006) it is stated that the random twist and the cut-and-twist methods lead to bounded relative error (strongly efficient) estimators. This claim is incorrect; in fact, the squared coefficient of variation of these estimators would typically grow at rate $cn^{1/2}$ for some $c \in (0, \infty)$. Finally, we note that the asymptotic approximations appear to converge slowly, incurring an error of the order of 70%, so Monte Carlo indeed appears to be a natural approach.

References

ADDIE, R., MANNERSALO, P. AND NORROS, I. (1999). Most probable paths and performance formulae for buffers with Gaussian input traffic. *Europ. Trans. Telecommun.* **13,** 183–196.

ADLER, R. J. AND TAYLOR, J. E. (2007). *Random Fields and Geometry*. Springer, New York.

ADLER, R. J., BLANCHET, J. AND LIU, J. (2008). Efficient simulation for tail probabilities of Gaussian random fields. In *Proc. 40th Conf. Winter Simul.*, pp. 328–336.

ADLER, R. J., BLANCHET, J. AND LIU, J. (2010). Efficient Monte Carlo for large excursions of Gaussian random fields. Preprint. Available at http://arxiv.org/abs/1005.0812v3.

ALZER, H. (1997). On some inequalities for the gamma and psi functions. *Math. Comput.* **66,** 373–389.

ASMUSSEN, S. AND GLYNN, P. W. (2007). *Stochastic Simulation: Algorithms and Analysis*. Springer, New York.

BERMAN, S. M. (1992). *Sojourns and Extremes of Stochastic Processes*. Wadsworth & Brooks/Cole, Pacific Grove, CA.

BLANCHET, J., GLYNN, P. AND LAM, H. (2009). Rare event simulation for a slotted time M/G/s model. *Queueing Systems* **63,** 33–57.

BOOTS, N. K. AND MANDJES, M. (2002). Fast simulation of a queue fed by a superposition of many (heavy-tailed) sources. *Prob. Eng. Inf. Sci*. **16,** 205–232.

BROCKWELL, P. J. AND DAVIS, R. A. (1991). *Time Series: Theory and Methods*, 2nd edn. Springer, New York.

Bucklew, J. A. and Radeke, R. (2003). On the Monte Carlo simulation of digital communication systems with Gaussian noise. *IEEE Trans. Commun.* **51,** 267–274.

Chang, J. T. and Peres, Y. (1997). Ladder heights, Gaussian random walks and the Riemann zeta function. *Ann. Prob.* **25,** 787–802.

Cody, W. J. (1969). Rational Chebyshev approximations for the error function. *Math. Comput.* **23,** 631–637.

Dębicki, K. (1999). A note on LDP for supremum of Gaussian processes over infinite horizon. *Statist. Prob. Lett.* **44,** 211–219.

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

Dieker, A. B. (2005). Extremes of Gaussian processes over an infinite horizon. *Stoch. Process. Appl.* **115,** 207–248.

Dieker, A. B. and Mandjes, M. (2006). Fast simulation of overflow probabilities in a queue with Gaussian input. *ACM Trans. Model. Comput. Simul.* **16,** 1–33.

Duffield, N. G. and O'Connell, N. (1995). Large deviations and overflow probabilities for the general single-server queue, with applications. *Math. Proc. Camb. Phil. Soc.* **118,** 363–374.

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

Durrett, R. (2004). *Probability: Theory and Examples*, 3rd edn. Duxbury Press, Belmont, CA.

Giordano, S., Gubinelli, M. and Pagano, M. (2007). Rare events of Gaussian processes: a performance comparison between bridge Monte-Carlo and importance sampling. In *Next Generation Teletraffic and Wired/Wireless Advanced Networking* (Lecture Notes Comput. Sci. **4712**), Springer, Berlin, pp. 269–280.

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

Hüsler, J. and Piterbarg, V. (1999). Extremes of a certain class of Gaussian processes. *Stoch. Process. Appl.* **83,** 257—271.

Juneja, S. and Shahabuddin, P. (2006). Rare event simulation techniques: an introduction and recent advances. In *Handbook on Simulation*, eds S. Henderson and B. Nelson, North-Holland, Amsterdam, pp. 291–350.

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

Mandjes, M. (2007). *Large Deviations for Gaussian Queues*. John Wiley, Chichester.

Michna, Z. (1999). On tail probabilities and first passage times for fractional Brownian motion. *Math. Meth. Operat. Res.* **49,** 335–354.

Norros, I. (1999). Busy periods of fractional Brownian storage: a large deviations approach. *Adv. Performance Analysis* **2,** 1–19.

Pickands, J., III (1969). Asymptotic properties of the maximum in a stationary Gaussian process. *Trans. Amer. Math. Soc.* **145,** 75–86.

Piterbarg, V. I. (1996). *Asymptotic Methods in the Theory of Gaussian Processes and Fields*. American Mathematical Society, Providence, RI.

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

Traub, J. F. (2003). Information-based complexity. In *Encyclopedia of Computer Science*, 4th edn. John Wiley, Chichester, pp. 850–854.