Efficient simulation for the maximum of infinite horizon Gaussian processes

Jose Blanchet and Chenxin Li
Columbia University

May 4, 2010

Abstract

We consider the problem of estimating the probability that the maximum of a Gaussian process with negative mean function and indexed by positive integers reaches a high level, say b. In great generality such probability converges to zero 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 Target Bridge Sampler (TBS) which is very easy to implement and takes a polynomial (in b) number of function evaluations achieve a small relative error. The procedure also yields samples that are arbitrarily close in total variation to the process given that it hits b in finite time. We also 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 fixed 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 our theoretical findings.

1 Introduction

Gaussian processes constitute an extremely popular class of models used in various settings in engineering and science, ranging from econometrics to communication networks (Campbell, Lo and MacKinlay (1997) and Mandjes (2007)). 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, due to such complex structures, the probabilistic analysis of non-linear functionals of a Gaussian process often becomes considerably difficult. One such non-linear functional of a Gaussian process that arises in communication applications and the analysis of queues with Gaussian input is its all time maximum (see for instance Addie, Mannersalo and Norros (2002)), which is the main focus of this paper.

Problem setup. In order to provide a concrete framework for our development, let us introduce some notations. Let $\{X_k : k \geq 1\}$ be a Gaussian process with mean zero and variance

$\sigma_k^2 = \mathrm{Var}(X_k)$. We consider a non-negative drift sequence $\{\mu_k : k \geq 1\}$. Our main interest is on efficient estimation via simulation of the tail probability

$$\alpha(b) = \mathbb{P}\left(\sup_{k\geq 1}(X_k - \mu_k) > b\right)$$

as $b \nearrow \infty$. We refer to this asymptotic environment as the "large buffer scaling" (the terminology is borrowed from the queueing setting, see Mandjes (2007), p. 65). We shall 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 infinity (i.e. $L_\sigma(ta)/L_\sigma(t) \longrightarrow 1$ as $t \longrightarrow \infty$ for all $a > 0$, 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. In fact, our assumption on $\sigma_\cdot$ and $\mu_\cdot$ are the most flexible in literature since no restrictions are imposed on the correlation structure nor the increments of the underlying Gaussian process are assumed to be stationary. We assume that $H_\sigma < H_\mu$ only because that implies $\alpha(b) \searrow 0$ as $b \nearrow \infty$. In fact, the convergence to zero is exponentially fast in a positive power of b as $b \nearrow \infty$ (see for instance Dębicki (1999)).

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 outputs an estimate that is guaranteed to be close in relative terms to $\alpha(b)$. As a function evaluation we consider a single addition, multiplication, a single evaluation of the Gaussian distribution and the simulation of a single uniform random number. 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 n independent and identically distributed (i.i.d.) Gaussian sources, commonly referred to as the "many sources scaling" as $n \nearrow \infty$. Hence, in this setting we replace σ_k^2, μ_k by $n\sigma_k^2$, $n\mu_k$ and b by bn. Although our results allow for the possibility of sending both $n, b \nearrow \infty$, when discussing the many sources setting we simply hold b fixed and send $n \nearrow \infty$ – just as it is typically done in the literature on many sources asymptotics.

Asymptotics. There is a rich literature on sharp asymptotic approximations for $\alpha(b)$, both under large buffer and many sources scalings, with more restricted conditions on the law of the process than the ones we imposed here. Pickands (1969) first explored this problem in the context of stationary processes and finite time horizon. The use of the double summation method allows to obtain refined results in broader environments, see for instance, the books by Berman (1992) and Piterbarg (1996). The text of Adler and Taylor (2007) 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; Husler and Piterbarg (1999) later provided exact asymptotics. Lately, Dieker (2006) 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 sta-

tionary increments; see also Dębicki and Mandjes (2003) for the continuous time counterpart. Many sources asymptotics is also the focus of 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 non-negligible relative to a given precision requirement. In addition, sharp asymptotics typically contain constants, such as the so-called generalized Pickands constant, that are difficult to evaluate. In fact, one often has to resort to Monte Carlo simulation to estimate such constants (see Burnecki and Michna (2002)).

Our algorithms here allow to evaluate $\alpha(b)$ to arbitrary precision with virtually no assumptions on the correlation structure of the process at a relatively small computational cost. Furthermore, our algorithms also enable the estimation of conditional expectations of the process given large excursions via efficient conditional sampling procedures.

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) Ch. 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 notions, 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 by 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 Adler, Blanchet and Liu (2008, 2010), but an important difference is that they require to simulate or approximate the whole process of interest. In contrast, as we shall see, our algorithm here involves simulating only a random number of components plus an additional "trimming" procedure. This distinction is particularly useful in the infinite horizon case that we consider.

In the many sources scaling setting, Boots and Mandjes (2002) provided a weakly efficient estimator. Dieker and Mandjes (2006) summarize and analyze several estimators, including one based on results of Dupuis and Wang (2004) and another one based on work of Sadowsky and Bucklew (1990). These two algorithms are also shown to be weakly efficient. Finally, we mention an alternative approach by 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.

As we shall discuss in Section 4, the large buffer scaling setting and the many sources scaling settings are intrinsically different; even in the case of self similar type input, such as the case of fractional Gaussian noise increments. The difference arises because of the nature of the discrete time formulation.

Contributions. Our main contributions are as follows:

i) In the large buffer setting we provide a simulation estimator for $\alpha(b)$ with a relative mean squared error at most of order $o\left(b^{1/H_\mu+\xi}\right)$ for any $\xi > 0$. Moreover, each replication of our estimator takes at most $o\left(b^{3/H_\mu+\xi}\right)$ function evaluations to be produced for any $\xi > 0$ (see Theorem 2 and Proposition 2).

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

$$\mathbb{P}_*\left(X_1, ...X_{T(b)} \in \cdot\right) \triangleq \mathbb{P}\left(X_1, ...X_{T(b)} \in \cdot \middle| T(b) < \infty\right) \tag{1}$$

with an expected number of function evaluations of order

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

for any $\xi > 0$ (see Theorem 2 and Proposition 2).

iii) 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 $O(1)$ number of function evaluations as as $n \nearrow \infty$ (see Theorem 3 and Proposition 2).

iv) Our underlying estimator is based on importance sampling and its nature is highly innovative given that it is *not* based on exponential tilting. This element allows us to circumvent the challenge of approximating and tracking the most likely path to overflow under an arbitrary correlation structure. It is remarkable that when applied to the case of Brownian motion, as we shall see in Proposition 1, our importance sampling estimator achieves zero-variance, therefore coinciding with the standard exponential tilting approach in this particular case.

Outline, notation and organization. In contrast to most of the importance sampling estimators proposed for Gaussian processes, our sampler (also based on importance sampling) does not use exponential tilting (also known as exponential change of measure) nor intends to simulate the process sequentially. Instead, we take advantage of the fact that the marginal distributions of the process are directly accessible. This observation allows to randomly select a marginal component according to a rough approximation (as $b \nearrow \infty$) for the conditional probability that such marginal reaches level b given that the all time maximum of the process exceeds level b. The randomly selected marginal is then sampled conditional on being larger than b. The Gaussian bridge, starting from zero up to the selected marginal value, is then simulated according to the conditional dynamics given the observed marginal value. The previous recipe induces an importance sampling distribution for the process up to the first passage time to level b. This is basically the importance sampling strategy that we analyze in this paper.

Throughout the paper we shall use Landau's notation for asymptotic behavior of functions. That is, given functions $f(\cdot)$ and $g(\cdot)$ we write $f(b) = o(g(b))$ as $b \nearrow \infty$ if $f(b)/g(b) \longrightarrow 0$

as $b \nearrow \infty$ and $f(b) = O(g(b))$ if $|f(b)| \leq cg(b)$ for some $c < \infty$ and all $b \geq 0$. We also use $\Phi(\cdot)$ to denote the standard Gaussian cumulative distribution function (CDF).

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 many sources scaling development 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 (in the settings in which they are applicable).

2 Basic Strategy of Target Bridge Sampling

As we mentioned in the Introduction, our method is based on importance sampling. For a review on importance sampling methodology the reader may consult the text of Asmussen and Glynn (2007), Section 5.1. The general theory of 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 event (Asmussen and Glynn (2007), p. 128). In our context, this means that in order to estimate $\alpha(b)$ with zero variance we must sample $(X_k : k \geq 1)$ given the event $T(b) < \infty$ (see equation (1) in the Introduction). Since such conditional distribution is not directly accessible, the objective is to construct an alternative probability measure, say $\mathbb{Q}$, 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 $\mathbb{Q}$ 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 conditional distribution of the process by a sequential mean-shifting procedure becomes extremely complicated.

Target sampling. Gaussian distributions, however, have special features that allow one to deal with complex dependence in a convenient way. One such feature is that a family of random variables 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.

In order to exploit this feature we start our sampling procedure by placing a point on the "target set" $\mathcal{T}$, defined as

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

This strategy is quite natural given that

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

Therefore, if we sample a random point $(\tau, X_\tau) \in \{0, 1, ...\} \times [b, \infty)$ according to some procedure and let it be an element of $\mathcal{T}$, this automatically implies $T(b) < \infty$. 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.

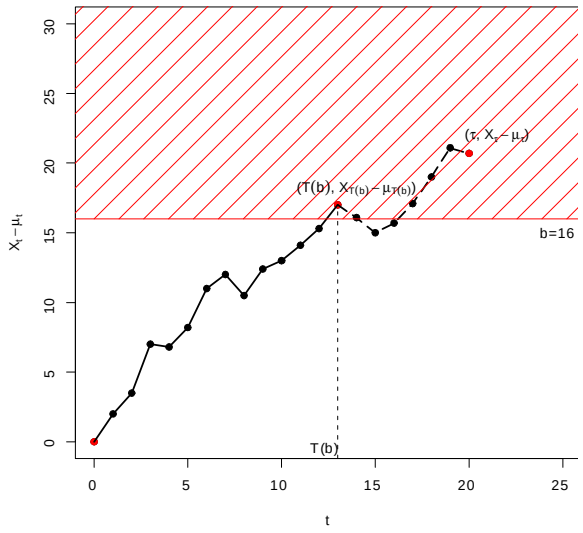

Figure 1: Figure illustrating a path generation under TBS.

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

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

As we shall see, under our assumptions on $\mu_\cdot$ and $\sigma_\cdot$ the denominator in the previous ratio 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). Sampling τ can be done efficiently via acceptance rejection. We will deal with this and related computational complexity issues involved in the implementation in future sections.

Second, given τ, we sample X_τ according to $\mathbb{P}(X_\tau \in \cdot | X_\tau - \mu_\tau > b)$. Simulating X_τ given that $X_\tau > b + \mu_\tau$ can be done via acceptance rejection with proposal distribution given by $(b + \mu_\tau) + \text{Exp}(1)\,\sigma_\tau^2/(b + \mu_\tau)$, where $\text{Exp}(1)$ 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.

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 $X_0, ..., X_{\tau-1}$ given (τ, X_τ) under the nominal (original) law, namely $\mathbb{P}(\cdot | \tau, X_\tau)$. This is easily done because of the Gaussian property of the process (again the computational complexity will be studied later in the paper). Since $\tau \geq T(b)$, once we have the path $X_0, ..., X_\tau$, 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 $(X_0, ..., X_{T(b)})$. *It is important to note that the path segment $X_{T(b)+1}, ..., X_\tau$ is discarded, this is the trimming procedure we alluded to in the Introduction.*

The combination of both steps corresponds to our Target Bridge Sampling (TBS) method. Figure 1 illustrates a generic path generated under TBS, in the picture $b = 16$, $T(b) = 13$ and $\tau = 20$. In summary, TBS does not attempt to directly approximate $T(b)$, as commonly done by conventional exponential tilting approaches. Instead, we suitably select an element from the target set $\mathcal{T}$. Next, in the bridge sampling step, we actually discard samples beyond $T(b)$.

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

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

$$\begin{aligned}
& \mathbb{Q}(X_1 \in B_1, ..., X_k \in B_k, T(b) = k) \\
= & \sum_{j=1}^{\infty} \mathbb{P}\left(X_1 \in B_1, ..., X_k \in B_k, T(b) = k \middle| X_j - \mu_j > b\right) p_\tau(j) \\
= & \sum_{j=k}^{\infty} \mathbb{P}\left(X_1 \in B_1, ..., X_k \in B_k, T(b) = k \middle| X_j - \mu_j > b\right) p_\tau(j) \\
= & \sum_{j=k}^{\infty} \frac{\mathbb{P}\left(X_1 \in B_1, ..., X_k \in B_k, T(b) = k, X_j - \mu_j > b\right)}{\sum_{n=1}^{\infty} \mathbb{P}(X_n - \mu_n > b)} \\
= & \mathbb{P}(X_1 \in B_1, ..., X_k \in B_k) I(T(b) = k) \\
& \times \frac{\sum_{j=k}^{\infty} \mathbb{P}\left(X_j - \mu_j > b \middle| X_1 \in B_1, ..., X_k \in B_k\right)}{\sum_{n=1}^{\infty} \mathbb{P}(X_n - \mu_n > b)}.
\end{aligned}$$

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

$$\begin{aligned}
& I(T(b) = k) \frac{d\mathbb{P}}{d\mathbb{Q}}(X_1, ..., X_k) \\
= & I(T(b) = k) \frac{\sum_{j=1}^{\infty} \mathbb{P}\left(X_j - \mu_j > b\right)}{\sum_{j=k}^{\infty} \mathbb{P}\left(X_j - \mu_j > b \middle| X_1, ..., X_k\right)}.
\end{aligned}$$

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

$$L = \frac{d\mathbb{P}}{d\mathbb{Q}}\left(X_1, ..., X_{T(b)}\right) = \frac{\sum_{j=1}^{\infty} \mathbb{P}\left(X_j - \mu_j > b\right)}{\sum_{j=T(b)}^{\infty} \mathbb{P}\left(X_j - \mu_j > b \middle| X_1, ..., X_{T(b)}\right)}. \tag{3}$$

Observe that

$$L \leq \sum_{j=1}^{\infty} \mathbb{P}\left(X_j - \mu_j > b\right).$$

Therefore,

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

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

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 it in principle can be used to construct an acceptance/rejection procedure. More precisely, using our notation for $\mathbb{P}_*(\cdot) = \mathbb{P}(\cdot \,|\, T(b) < \infty)$ introduced in equation (1) we obtain

$$\frac{L}{\alpha(b)} = \frac{d\mathbb{P}_*}{d\mathbb{Q}}\left(X_1, ..., X_{T(b)}\right) \leq \frac{\sum_{n=1}^{\infty} \mathbb{P}\left(X_j - \mu_j > b\right)}{\alpha(b)}. \tag{4}$$

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

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

A sample path accepted under this procedure follows the law $\mathbb{P}_*$. The procedure has acceptance ratio $\sum_{k=1}^{\infty} \mathbb{P}(X_k - \mu_k > b) / \alpha(b)$, and therefore the acceptance probability per sample is equal to $\alpha(b) / \sum_{k=1}^{\infty} \mathbb{P}(X_k - \mu_k > b)$. As a consequence, as we shall show in Theorem 1 below, the acceptance probability under this acceptance / rejection procedure goes to zero polynomially as $b \nearrow \infty$ in great generality.

An insightful case: Brownian motion. In order to have a sense of how good a performance can be expected out of TBS, it is worth considering a case for which $\alpha(b)$ can be explicitly computed. More specifically, consider the case that $X.$ is the standard Brownian motion under $\mathbb{P}$ and $\mu_t = \mu t$ linear for $t \geq 0$. For this setting, we know that $\alpha(b) = \mathbb{P}(\max_{t \geq 0}(X_t - \mu t) > b) = \exp(-2\mu b)$.

Now we apply the idea of TBS to this continuous time setting. As a simple analogy to the procedures we described above for the discrete Gaussian processes, we proceed by first sampling a time τ with density $f_\tau(\cdot)$ proportional to $\mathbb{P}(X. > b)$. Then, given τ, we sample the Brownian bridge $(X_s - \mu s : 0 \leq s \leq t)$ conditional on the observed value X_τ which in turn had been sampled from the distribution of X_τ given that $X_\tau - \mu\tau > b$. Following a simple

discretization procedure we obtain that TBS gives a likelihood ratio for the generated path $(X_s : 0 \leq s \leq T(b))$ equal to

$$L = \frac{\int_0^\infty \mathbb{P}(X_s - \mu s > b)\, ds}{\int_{T(b)}^\infty \mathbb{P}\left(X_u - \mu t > b | X_{T(b)}\right) dt} = \frac{\int_0^\infty \mathbb{P}(X_s - \mu s > b)\, ds}{\int_0^\infty \mathbb{P}(X_t - \mu t > 0)\, dt},$$

which has zero variance because the right most expression is non-random and depends on neither τ nor $T(b)$. Because the sampler is unbiased we must have that $L = \exp(-2\mu b)$. This expression can also be directly checked by applying the Laplace-Fourier transforms to $\int_0^\infty \mathbb{P}(X_s - \mu s > b)\, ds$ 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 we discussed at the beginning of the section, such distribution is equal to the conditional distribution of $(X_s - \mu s : 0 \leq s \leq T(b))$ given that $T(b) < \infty$, which is known to be described by the process $(B_s + \mu s : 0 \leq s \leq T_1(b))$ where $(B_s : s \geq 0)$ is a standard Brownian motion and $T_1(b) \triangleq \inf\{s \geq 0 : B_s + \mu s > b\}$. We record this observation as a proposition.

Proposition 1. *Under the probability measure $\mathbb{Q}$ generated by the sampling strategy described before, it follows that the law of $(X_s - \mu s : 0 \leq s \leq T(b))$ is just that of $(B_s + \mu s : 0 \leq s \leq T_1(b))$. Therefore, $\mathbb{Q}$ is the zero-variance importance sampling probability measure.*

This result shows the potential behind of TBS. The standard approach of describing the zero-variance change-of-measure for estimating $\mathbb{P}(\max_{t\geq 0}(X_t - \mu t) > b)$, by means of an exponential tilting of the drift of $X_\cdot$, could give rise to very complicated descriptions in situations in which $X_\cdot$ exhibits very complex dependence. Our approach, on the other hand is conceptually applicable to virtually any Gaussian situation. In principle, one only requires that: 1) the marginal distributions; 2) conditional distributions given marginals can be sampled and 3) conditional marginal probabilities can also be computed. These features actually appear in other processes of interest beyond the Gaussian case. A careful study of the application of TBS in such cases will appear elsewhere.

Main result of the section. We conclude this section with a summary of the mean squared error properties behind an estimator based on $\mathbb{Q}(\cdot)$.

Theorem 1. *If $(\sigma_k : k \geq 1)$ and $(\mu_k : k \geq 1)$ are regularly varying with indices $0 < H_\sigma < H_\mu < \infty$ respectively, then we have that $\mathbb{E}^{\mathbb{Q}} L = \alpha(b)$ and*

$$\frac{\mathrm{Var}^{\mathbb{Q}}(L)}{\mathbb{P}(\max_{k\geq 1} X_k - \mu_k > b)^2} = o(b^{2/H_\mu + 2\xi}) \qquad \text{as } b \to \infty. \tag{5}$$

for any $\xi > 0$. Moreover, the acceptance ratio of the procedure described in (4) satisfies

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

for any $\xi > 0$.

We shall prove this theorem at the end of 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 non-exponential change-of-measure $\mathbb{Q}$. We noted, however, that from a simulation standpoint, the likelihood ratio estimator L resulted from $\mathbb{Q}$ cannot be directly computed because it requires the exact computation of certain infinite sums. In this section we shall study a couple of ways to address this issue.

First, one way to remedy this situation is via another randomization step as we now explain. Given the path $\left(X_1, \ldots, X_{T(b)}\right)$ obtained out of TBS, define the function

$$\widetilde{L}\left(n; X_1, \ldots, X_{T(b)}\right) = \frac{\mathbb{P}\left(X_{n-T(b)+1} - \mu_{n-T(b)+1} > b \middle| X_1, ..., X_{T(b)}\right)}{\mathbb{P}\left(X_n - \mu_n > b | X_1, ..., X_{T(b)}\right)}, \quad k \geq T\left(b\right).$$

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

$$p_N\left(k\right) = \frac{\mathbb{P}\left(X_k - \mu_k > b | X_1, ..., X_{T(b)}\right)}{\sum_{k=T(b)}^{\infty} \mathbb{P}\left(X_k - \mu_k > b | X_1, ..., X_{T(b)}\right)}, \quad k \geq T\left(b\right).$$

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

$$\begin{aligned}
& \mathbb{E}(\ \widetilde{L}\left(N; X_1, \ldots, X_{T(b)}\right) \middle| X_1, \ldots, X_{T(b)}) \\
= & \sum_{k=T(b)}^{\infty} \frac{\mathbb{P}\left(X_{k-T(b)+1} - \mu_{k-T(b)+1} > b \middle| X_1, ..., X_{T(b)}\right)}{\mathbb{P}\left(X_k - \mu_k > b | X_1, ..., X_{T(b)}\right)} \frac{\mathbb{P}\left(X_k - \mu_k > b | X_1, ..., X_{T(b)}\right)}{\sum_{k=T(b)}^{\infty} \mathbb{P}\left(X_k - \mu_k > b | X_1, ..., X_{T(b)}\right)} \\
= & \frac{\sum_{j=1}^{\infty} \mathbb{P}\left(X_j - \mu_j > b\right)}{\sum_{j=T(b)}^{\infty} \mathbb{P}\left(X_j - \mu_j > b \middle| X_1, ..., X_{T(b)}\right)} = L.
\end{aligned}$$

Although the previous approach provides an unbiased estimator, it will introduce some variance due to the additional randomization step. Another way to deal with the computational issue is to truncate the sums, although this inevitably induces a bias in the estimator. This truncation technique, used by Dieker and Mandjes (2006), is the one that we shall consider in the rest of the section. In particular, we analyze

$$\alpha_{t^+(b)}(b) = \mathbb{P}\left(\max_{1 \leq k \leq t^+(b)} \left(X_k - \mu_k\right) \geq b\right), \tag{6}$$

by choosing $t^+(b)$ sufficiently large so that $\alpha_{t^+(b)}\left(b\right) \leq \alpha(b) \leq \alpha_{t^+(b)}\left(b\right)\left(1+\varepsilon\right)$ for $\varepsilon > 0$ fixed.

Our analysis naturally takes advantage of a time, k, that minimizes the rate of decay to zero of $\mathbb{P}(X_k - \mu_k > b)$. Therefore, we let $k^*(b)$ be any optimizer of such rate of decay, in particular

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

where $g(k;b) \triangleq (b+\mu_k)/\sigma_k$. Clearly we have that $\alpha_{t^+(b)}(b) \leq \alpha(b)$ and also that

$$\mathbb{P}(X_{k^*(b)} - \mu_{k^*(b)} > b) \leq \alpha(b) \leq \alpha_{t^+(b)}(b) + \sum_{k=t^+(b)+1}^{\infty} \mathbb{P}(X_k - \mu_k > b). \tag{7}$$

In order to guarantee that the infinite sum in the previous display is small relative to the lower bound on $\alpha(b)$, the definition of $t^+(b)$ requires a bound on $g(k^*(b);b)$ which is introduced in the following lemma, whose proof is given at the end of the section.

Lemma 1. *Suppose that $(\sigma_k : k \geq 1)$ and $(\mu_k : k \geq 1)$ are regularly varying with indices $0 < H_\sigma < H_\mu < \infty$ respectively. Then for any $0 < \delta < \min\{(H_\mu - H_\sigma)/2, H_\sigma\}$, there exists $M(\delta) \in \mathbb{N}$ such that for $\forall k \geq M(\delta)$,*

$$\begin{aligned} k^{H_\mu-\delta} &\leq \mu_k \leq k^{H_\mu+\delta}, \\ k^{H_\sigma-\delta} &\leq \sigma_k \leq k^{H_\sigma+\delta} \end{aligned}$$

and

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

where $c(\delta;H_\sigma,H_\mu) = \left(\frac{H_\sigma-\delta}{H_\mu-H_\sigma+2\delta}\right)^{-\frac{H\sigma-\delta}{H_\mu+\delta}} \left(1 + \frac{(H_\sigma-\delta)2^{H_\mu-H_\sigma+2\delta}}{H_\mu-H_\sigma+2\delta}\right)$ is a constant independent of b.

Using $h(b)$ it turns out, as we shall explain in the proof of Theorem 2 below, that $t^+(b)$ in (6) can be specified as

$$t^+(b) = \left\lceil \lambda \left(h(b) + \frac{1}{h(b)} \left| \log\left(\frac{\left(1 - h(b)^{-2}\right) p\varepsilon}{2^{\eta-1}\lambda} \right) \right| \right)^{1/q} \right\rceil, \tag{8}$$

where $q = (H_\mu - H_\sigma - 2\delta)$, $\eta = \max\{1, 1/2q\}$, $\lambda = \Gamma(1+\eta)$ and $\varepsilon > 0$ was introduced right after equation (6). This is a proper definition because $h(b) > 1$ for any b. Note that by choosing δ sufficiently small we ensure

$$t^+(b) = o\left(b^{1/H_\mu+\xi} + \log\left(\varepsilon^{-1}\right)^{1/(H_\mu-H_\sigma+2\xi)} \right) \tag{9}$$

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

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 $\mathbb{Q}_0(\cdot)$, $\mathbb{E}^{\mathbb{Q}_0}(\cdot)$ and $\mathrm{Var}^{\mathbb{Q}_0}(\cdot)$ to denote the

probability measure, expectation operator and variance induced by the importance sampling distribution in Algorithm 1.1.

Algorithm 1.1

1. Set $t^+(b)$ according to (8)
2. Targeting:
   - Sample τ according to probability mass function
   $$p_\tau(k) = \frac{\mathbb{P}(X_k > b)}{\sum_{j=1}^{t^+(b)} \mathbb{P}(X_j > b)}.$$
   - Given τ, sample X_τ according to the law $\mathbb{P}(X_\tau \leq \cdot | X_\tau > b + \mu_\tau)$;
3. Bridging: Given X_τ, sample the Gaussian bridge $X_1, X_2, \ldots, X_{\tau-1} | X_\tau$ from the nominal (original) distribution.
4. Find $T(b) = \min\{j \geq 1 : X_j - \mu_j > b\}$
5. Compute and output the likelihood estimator
$$L_{t^+(b)} = \frac{\sum_{j=1}^{t^+(b)} \mathbb{P}\left(X_j - \mu_j > b\right)}{\sum_{j=T(b)}^{t^+(b)} \mathbb{P}\left(X_j - \mu_j > b \middle| X_1, ..., X_{T(b)}\right)}.$$

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

$$\begin{aligned}
\mathbb{P}(A | T(b) < \infty) &\leq \frac{\mathbb{P}(A, T(b) \leq t^+(b))}{\alpha(b)} + \frac{\alpha(b) - \alpha_{t^+(b)}(b)}{\alpha(b)} \\
&\leq \frac{\mathbb{P}(A, T(b) \leq t^+(b))}{\alpha_{t^+(b)}(b)} + 1 - \frac{\alpha_{t^+(b)}(b)}{\alpha(b)} \\
&\leq \mathbb{P}\left(A | T(b) \leq t^+(b)\right) + \varepsilon.
\end{aligned}$$

Also,

$$\begin{aligned}
\mathbb{P}\left(A | T(b) \leq t^+(b)\right) &= \frac{\mathbb{P}(A, T(b) \leq t^+(b))}{\alpha_{t^+(b)}(b)} \\
&\leq \frac{\mathbb{P}(A, T(b) < \infty)}{\alpha_{t^+(b)}(b)} \leq \mathbb{P}(A | T(b) < \infty) + \varepsilon.
\end{aligned}$$

Therefore

$$\sup_{A \in \mathcal{F}_{T(b)}} \left| \mathbb{P}\left(A | T(b) \leq t^+(b)\right) - \mathbb{P}(A | T(b) < \infty) \right| \leq \varepsilon.$$

We now state explicitly our algorithm for generating exact samples from $\mathbb{P}\left(\left(X_1,...,X_{T(b)}\right)\in\cdot\,\middle|\,T(b)\leq t^+(b)\right)$.

Algorithm 1.2

1. Set $t^+(b)$ according to (8).
2. Repeat until acceptance: Sample $L_{t^+(b)}$ according to Algorithm 1.1 and accept if $L_{t^+(b)}\leq U\sum_{j=1}^{t^+(b)}\mathbb{P}\left(X_j-\mu_j>b\right)$, where U is uniformly distributed over $[0,1]$ and independent of $L_{t^+(b)}$.
3. Output the accepted path $X_1,...,X_{T(b)}$ obtained by Algorithm

The next result summarizes the statistical properties in terms of relative mean squared error of the estimator $L_{t^+(b)}$ and the expected number proposals required to terminate Algorithm 1.2 (recall the definition of $\mathbb{Q}_0(\cdot)$ given right before Algorithm 1.1).

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

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

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

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

as $b\nearrow\infty$ and the number of proposals required to terminate Algorithm 1.2 is geometrically distributed with mean $o\left(b^{1/H_\mu+\xi}\right)$ as $b\nearrow\infty$.

Before we provide the proof of Lemma 1 and Theorem 2, we need the following result, which is an adaptation of the bounds obtained by Alzer (1997). The proof is somewhat similar to that of Lemma 2.1 and Lemma 2.2 in Dieker and Mandjes (2006) and elementary. Nevertheless, we provide a full argument here for completeness.

Lemma 2. *For any time $T\geq 1$ and parameters $q,C>0$, we have that*

$$\int_T^\infty\exp(-Ct^q)dt<\frac{\lambda}{qC^\eta}e^{-C(T/\lambda)^q},$$

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

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

$$\begin{aligned}&\int_T^\infty\exp(-Ct^q)dt\\ \leq\;&\frac{1}{qC^{1/q}}\int_{CT^q}^\infty s^{1/q-1}\exp(-s)ds\leq\frac{T^{1-q}}{qC}\int_{CT^q}^\infty\exp(-s)ds\leq\frac{1}{qC}e^{-CT^q}.\end{aligned}$$

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

$$\frac{1}{\Gamma(1+1/q)} \int_x^\infty e^{-t^q} dt < 1 - [1 - e^{-[\Gamma(1+1/q)]^{-q} x^q}]^{1/q},$$

Therefore, we have

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

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

Now we provide the proof 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 Potts bound, i.e. $\forall \delta < (H_\mu - H_\sigma)/2$, $\exists M(\delta) \in \mathbb{N}$ sufficiently large such that $\forall k \geq M(\delta)$

$$\begin{aligned} k^{H_\mu - \delta} &\leq \mu_k \leq k^{H_\mu + \delta}, \\ k^{H_\sigma - \delta} &\leq \sigma_k \leq k^{H_\sigma + \delta}, \\ g_L(k;b) \triangleq k^{H_\mu - H_\sigma - 2\delta} &\leq g(k;b) \leq \frac{b + k^{H_\mu + \delta}}{k^{H_\sigma - \delta}} \triangleq g_U(k;b)\,. \end{aligned}$$

In order to find the minimum of $g_U(k;b)$, we treat $g_U(\cdot;b)$ as a function defined on $\mathbb{R}$. Define

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

Then

$$\begin{aligned} g_U^{'}(t;b) &= \frac{(H_\mu + \delta) t^{H_\mu + H_\sigma - 1} - (H_\sigma - \delta) t^{H_\sigma - \delta - 1} \left(b + t^{H_\mu + \delta}\right)}{t^{2H_\sigma - 2\delta}} \\ &= \frac{(H_\mu - H_\sigma + 2\delta) t^{H_\mu + \delta} - (H_\sigma - \delta) b}{t^{H_\sigma - \delta + 1}} \\ t_U^*(b) &= \left(\frac{H_\sigma - \delta}{H_\mu - H_\sigma + 2\delta} b\right)^{1/(H_\mu + \delta)}. \end{aligned}$$

Therefore, for any $\delta < \min\left((H_\mu - H_\sigma)/2, H_\sigma\right)$, when $b < \frac{H_\mu - H_\sigma + 2\delta}{H_\sigma - \delta} M(\delta)^{H_\mu + \delta}$, we have $t_U^*(b) < M(\delta)$. We can only say that

$$g(k^*;b) \leq g(M(\delta);b) \leq g_U(M(\delta);b) = \frac{b + M(\delta)^{H_\mu + \delta}}{M(\delta)^{H_\sigma - \delta}}.$$

On the other hand, if $b \geq \frac{H_\mu - H_\sigma + 2\delta}{H_\sigma - \delta} M(\delta)^{H_\mu + \delta}$, we have

$$t_U^*(b) \geq M(\delta) \geq 1,$$

therefore

$$g(k^*; b) \leq \min_{k \geq t_U^*(b)} g(k; b) \leq \min_{k \geq t_U^*(b)} g_U(k; b).$$

However, because of discretization, we don't necessarily have $\min_{k \geq t_U^*(b)} g_U(k; b) \leq g_U(t_U^*)$. Rather, considering that $g_U(t)$ is monotonically increasing on $[t_U^*, +\infty)$, we know

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

□

Now we are ready to prove the two theorems. Actually the proofs of them are 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,

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

For the numerator, applying a well known bound on the Gaussian cumulative distribution function, $\Phi(\cdot)$ (see Durrett (2004), p. 6), namely

$$\frac{1}{\sqrt{2\pi}}\left(x^{-1} - x^{-3}\right)\exp\left(-\frac{1}{2}x^2\right) \leq 1 - \Phi(x) \leq \frac{1}{\sqrt{2\pi}} x^{-1} \exp\left(-\frac{1}{2}x^2\right), \tag{10}$$

we get

$$\begin{aligned}\sum_{k=t^+(b)+1}^{\infty}\mathbb{P}(X_k-\mu_k\geq b) &\leq \frac{1}{\sqrt{2\pi}}\sum_{k=t^+(b)+1}^{\infty}\exp\left(-\frac{1}{2}g\left(k;b\right)^2\right)\Big/\, g\left(k;b\right)\\ &\leq \frac{1}{\sqrt{2\pi}h(b)}\sum_{k=t^+(b)+1}^{\infty}\exp\left(-\frac{1}{2}g_L(k;b)^2\right)\\ &\leq \frac{1}{\sqrt{2\pi}h(b)}\int_{t^+(b)}^{\infty}\exp\left(-\frac{1}{2}g_L(t;b)^2\right)dt \qquad (11)\end{aligned}$$

Now, applying Lemma 2 to (11), we get

$$\sum_{k=t^+(b)+1}^{\infty}\mathbb{P}(X_k-\mu_k\geq b) < \frac{2^{\eta-1}\lambda}{H_\mu-H_\sigma-2\delta}\frac{\exp\left(-g_L(t^+(b)/\lambda)^2/2\right)}{\sqrt{2\pi}g_L(t^+(b);b)}$$

with $\eta=\max\{1,1/2(H_\mu-H_\sigma-2\delta)\}$ and $\lambda=\Gamma(1+\eta)\geq 1$. Recall that $t^+(b)$ is defined as

$$t^+(b)=\left\lceil\lambda\left(h(b)+\frac{1}{h(b)}\left|\log\frac{\left(1-h(b)^{-2}\right)\left(H_\mu-H_\sigma-2\delta\right)\varepsilon}{2^{\eta-1}\lambda}\right|\right)^{1/(H_\mu-H_\sigma-2\delta)}\right\rceil,$$

therefore

$$\begin{aligned}&\frac{2^{\eta-1}\lambda}{H_\mu-H_\sigma-2\delta}\frac{\exp\left(-g_L(t^+(b)/\lambda)^2/2\right)}{\sqrt{2\pi}g_L(t^+(b);b)}\\ <\;& \frac{\exp\left(-h(b)^2/2\right)}{\sqrt{2\pi}}\left(h(b)^{-1}-h(b)^{-3}\right)\varepsilon\\ \leq\;& \left[1-\Phi(h(x))\right]\varepsilon\leq\varepsilon\mathbb{P}(X_{k^*}-\mu_{k^*}\geq b).\end{aligned}$$

Equivalently,

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

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

$$\begin{aligned}&\frac{\mathrm{Var}^{\mathbb{Q}}(L_{t^+(b)})}{\mathbb{P}(\max_{1\leq k\leq t^+(b)}(X_k-\mu_k)>b)^2}\\ \leq\;& \frac{\mathbb{E}^{\mathbb{Q}}(L_{t^+(b)}^2)}{\mathbb{P}(\max_{1\leq k\leq t^+(b)}(X_k-\mu_k)>b)^2}\\ =\;& \left(\frac{\sum_{k=1}^{t^+(b)}\mathbb{P}(X_k-\mu_k\geq b)}{\mathbb{P}(\max_{1\leq k\leq t^+(b)}(X_k-\mu_k)>b)}\right)^2\mathbb{E}\left(\frac{1}{\sum_{k=t^*}^{t^+(b)}\mathbb{P}(X_k\geq b|X_{[0,t^*]})}\right)^2\\ \leq\;& \left(\frac{\sum_{k=1}^{t^+(b)}\mathbb{P}(X_k-\mu_k\geq b)}{\mathbb{P}(\max_{1\leq k\leq t^+(b)}(X_k-\mu_k)>b)}\right)^2\leq t^+(b)^2=O\left(b^{\frac{H_\mu-H_\sigma+2\delta}{H_\mu-H_\sigma-2\delta}\frac{2}{H_\mu+\delta}}\right).\end{aligned}$$

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

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

for any $\xi > 0$. □

Theorem 1 now follows as a simple extension out 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

$$\mathbb{E}^{\mathbb{Q}}L = \alpha(b) = \mathbb{P}\left(\max_{k\ge 1}(X_k-\mu_k)\ge b\right).$$

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

$$\begin{aligned}
& \frac{\mathrm{Var}^{\mathbb{Q}}(L)}{\mathbb{P}(\max_{k\ge1}(X_k-\mu_k)\ge b)^2} \\
\le\ & \frac{\mathbb{E}^{\mathbb{Q}}(L^2)}{\mathbb{P}(\max_{k\ge1}(X_k-\mu_k)\ge b)^2} \\
=\ & \left(\frac{\sum_{k=1}^{\infty}\mathbb{P}(X_k-\mu_k\ge b)}{\mathbb{P}(\max_{k\ge1}(X_k-\mu_k)\ge b)}\right)^2 \mathbb{E}\left(\frac{1}{\sum_{k=k^*}^{\infty}\mathbb{P}(X_k\ge b|X_{[0,k^*]})}\right)^2 \\
\le\ & \left(t^+(b)+\varepsilon\right)^2 \\
=\ & O\left(b^{\frac{H_\mu-H_\sigma+2\delta}{H_\mu-H_\sigma-2\delta}\frac{2}{H_\mu+\delta}}\right) = O\left(b^{2/H_\mu+\xi}\right).
\end{aligned}$$

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 queuing theory and especially in network load problems. It turns out that our method not only is 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 zero. This may seem surprising given that the large buffer and the many sources regimes are known to be basically equivalent for a class of models in continuous time (in particular fractional Brownian motion, the equivalence follows from self-similarity). As we shall discuss later in the section, the key difference is precisely the continuous versus discrete time formulation.

To state the problem in mathematical terms, let us define

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

where the processes $\left(X_k^{(j)} : k \geq 1\right)$, $j \geq 1$ 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

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

It is important to emphasize that in this section we mainly concentrate on $\tilde{\alpha}\left(nb\right)$ 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 by Dieker and Mandjes (2006). They applied four methods to this problem all of them 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 basically corresponds to applying Algorithm 1.1 to the process $\left(\tilde{X}_k^{(n)} : k \geq 1\right)$, gives rise to a strongly efficient estimator (bounded in n number of function evaluations).

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

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

Because of the factor $n^{1/2}$ appearing in $\tilde{g}\left(k; nb\right)$ it turns out, as our next result shows, that our selection $t^+\left(b\right)$ defined in equation (8) can still be used as a truncation threshold. The proof of the next lemma is given at the end of the section.

Lemma 3. *For any $\varepsilon > 0$, let $t^+(b)$ be defined in equation (8). Then for any $n \geq 1$, we have*

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

The previous lemma allows to concentrate our efforts on estimating

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

at the price of introducing ε-relative bias. A detailed algorithm, completely analogous to Algorithm 1.1 is given next for estimating $\widetilde{\alpha}_{t^+(b)}\left(nb\right)$.

Algorithm 2

1. Set $t^+\left(b\right)$ according to (8)

2. Targeting:

   - Sample τ according to probability mass function

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

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

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

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

5. Compute the likelihood estimator

$$L_n = \frac{\sum_{j=1}^{t^+(b)} \mathbb{P}(\tilde{X}_j^{(n)} - n\mu_j > nb)}{\sum_{j=T(b)}^{t^+(b)} \mathbb{P}\left(\tilde{X}_j^{(n)} - n\mu_j > nb \middle| \tilde{X}_1^{(n)}, ..., \tilde{X}_{T(b)}^{(n)}\right)}.$$

We use $\mathbb{Q}_1(\cdot)$ to denote the probability measure corresponding to the importance sampling strategy introduced before and $\mathbb{E}^{\mathbb{Q}_1}(\cdot)$ and $\mathrm{Var}^{\mathbb{Q}_1}(\cdot)$ for 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 $b = O(1)$ as $n \nearrow \infty$). In particular, we have that*

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

Proof. Just note that

$$\begin{aligned} & \frac{\mathbb{E}^{\mathbb{Q}_1}(L_n^2)}{\mathbb{P}(\max_{k \leq t^+(b)} \tilde{X}_k^{(n)} - n\mu_k \geq nb)} \\ \leq & \left[\frac{\sum_{k=0}^{t^+(b)} \mathbb{P}\left(\tilde{X}_k^{(n)} - n\mu(k) \geq nb\right)}{\max_{k \leq t^+(b)} \mathbb{P}\left(\tilde{X}_k^{(n)} - n\mu(k) \geq nb\right)}\right]^2 \\ \leq & t^+(b)^2. \end{aligned}$$

□

Before providing the proof of Lemma 3 and closing the section, let us discuss why one cannot translate strong efficiency from the many sources setting to the large buffer setting in our current discrete setting, although it is well known (Ganesh et al (2004)) that the two regimes are equivalent in continuous time. Let us consider the case when $(X_t : t \geq 0)$ is fBm

and $(\mu_t : t \geq 0)$ is linear. That is, $\mathrm{Cov}(X_t, X_s) = (t^{2H_\sigma} + s^{2H_\sigma} - |t-s|^{2H_\sigma})/2$ and $\mu_t = t$. Then, because of the self-similarity property, $X(a\cdot) =_d a^{H_\sigma} X^{(j)}(\cdot)$ for any $a > 0$ (where $=_d$ denotes equality in distribution). We have

$$\begin{aligned} & \mathbb{P}\left(\max_t \sum_{j=1}^n \left(X_t^{(j)} - t\right) > nb\right) \\ = \; & \mathbb{P}\left(\max_t \left(\sqrt{n}X_t - nt\right) > nb\right) \\ = \; & \mathbb{P}\left(\max_t \left(X\left(n^{1/2(1-H_\sigma)}t\right) - n^{1/2(1-H_\sigma)}t\right) > n^{1/2(1-H_\sigma)}b\right). \end{aligned}$$

Therefore, if we set $s = n^{1/2(1-H_\sigma)}t$, $b^* = n^{1/2(1-H_\sigma)}b$, then the rare event under the many sources regime translates into a large buffer problem via the equality

$$\mathbb{P}\left(\max_t \sum_{j=1}^n \left(X_t^{(j)} - t\right) > nb\right) = \mathbb{P}\left(\max_s \left(X_s - s\right) > b^*\right), \tag{12}$$

where $b^* \to \infty$.

This translation, however, does not carry over at the level of an implementable simulation algorithm directly, which requires a discretization grid. Deep down, the efficiency proof uses the discrete nature of the problem in a crucial way. When we translate from the many sources formulation into a large buffer formulation an implicit side-effect is that the grid width (now in the large buffer formulation) will increase by a factor of b^*/b (or $n^{1/2(1-H_\sigma)}$) as well. So when we send $n \nearrow \infty$, the algorithm implied by (12) does have $b^* \nearrow \infty$, that is, large buffer, but it also has a grid length $n^{1/(2-2H_\sigma)}$. That is different from Algorithm 1.1, where the grid length is 1 regardless of the buffer size b. As a consequence, the asymptotic efficiency of the existing method on the many sources scaling cannot be translated directly into to the type of asymptotic efficiency considered in the large buffer scaling. From this perspective, the discrete large buffer scaling studied in Section 3 is more difficult than the discrete multisource scaling problem studied in this section.

We close the section with the proof of Lemma 3.

Proof of Lemma 3. Once again, the calculation boils down to

$$\begin{aligned} & \frac{\mathbb{P}\left(\max_{k>t^+(b)} \sum_{j=1}^n \left[X_k^{(j)} - \mu_k\right] > nb\right)}{\mathbb{P}\left(\max_{1\leq k<\infty} \sum_{j=1}^n \left[X_k^{(j)} - \mu_k\right] > nb\right)} \\ \leq \; & \frac{\sum_{k>t^+(b)} \mathbb{P}\left(\sum_{j=1}^n \left[X_k^{(j)} - \mu_k\right] > nb\right)}{\max_{1\leq k<\infty} \mathbb{P}\left(\sum_{j=1}^n \left[X_k^{(j)} - \mu_k\right] > nb\right)}. \end{aligned}$$

Since $\mathbb{P}\left(\sum_{j=1}^n \left[X_k^{(j)} - \mu_k\right] > nb\right) = 1 - \Phi\left(\sqrt{n}g(k;b)\right)$, where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard Gaussian random variable, we know (once again appealing

to Durrett (2004) p. 6)

$$\begin{aligned}\max_{1\leq k<\infty}\mathbb{P}\left(\sum_{j=1}^{n}X_k^{(j)}-n\mu_k>nb\right) &\geq 1-\Phi\left(\sqrt{n}h(b)\right)\\ &\geq \left(n^{-1/2}h\left(b\right)^{-1}-n^{-3/2}h(b)^{-3}\right)\exp\left(-\frac{n}{2}h(b)^2\right)\end{aligned}$$

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

$$\begin{aligned}&\sum_{k=t^+(b)+1}^{\infty}\mathbb{P}\left(\sum_{j=1}^{n}X_k^{(j)}-n\mu_k>nb\right)\\ <\ &\frac{1}{\sqrt{2\pi n}h(b)}\int_{t^+(b)}^{\infty}\exp\left(-\frac{n}{2}g_L(t)^2\right)dt\end{aligned}$$

Using Lemmas 1 and 2, we get

$$\begin{aligned}&\int_{t^+(b)}^{\infty}\exp\left(-\frac{n}{2}g_L(t)^2\right)dt\\ <\ &\frac{2^{\eta-1}\lambda}{\left(H_\mu-H_\sigma-2\delta\right)n^\eta}e^{-n(t^+(b)/\lambda)^{2\left(H_\mu-H_\sigma-2\delta\right)}/2}\\ <\ &\frac{2^{\eta-1}\lambda}{\left(H_\mu-H_\sigma-2\delta\right)n}e^{-n(t^+(b)/\lambda)^{2\left(H_\mu-H_\sigma-2\delta\right)}/2},\end{aligned}$$

where $\eta=\max\{1,1/\left(H_\mu-H_\sigma-2\delta\right)\}$ and $\lambda=\Gamma(1+\eta)$.

Once again, recall that

$$t^+(b)=\left\lceil\lambda\left(h(b)+\frac{1}{h(b)}\left|\log\frac{\left(1-h(b)^{-2}\right)\left(H_\mu-H_\sigma-2\delta\right)\varepsilon}{2^{\eta-1}\lambda}\right|\right)^{1/(H_\mu-H_\sigma-2\delta)}\right\rceil$$

then,

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

□

5 Complexity Analysis per Replication and Numerical Example

We have proved the statistical efficiency of our estimator (in terms of 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 such number of function evaluations and compare 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 single addition, multiplication, comparison, generation of a single Gaussian random variable and the evaluation of the Gaussian CDF.

We concentrate on Algorithm 1.1. The highest computational burden corresponds to Steps 3 and 4 in terms of generating the sample $(X_k : k \leq \tau)$, which involves a matrix inversion (necessary to compute the conditional variance given the observed value X_k when $\tau = k$). Such inversion procedure requires at most $O(t^+(b)^3)$ function evaluations. Another contribution corresponds to the calculation of our importance sampling estimator

$$L_{t^+(b)} = \frac{\sum_{k=0}^{t^+(b)} \mathbb{P}(X_t > b)}{\sum_{k=T(b)}^{t^+(b)} \mathbb{P}(X_k \geq b | \mathbf{X}_{T(b)})},$$

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

$$\begin{aligned} \mathbb{E}^{\mathbb{Q}}[X_k] &= \mathbf{V}_k^T \mathbf{\Sigma}_{T(b)}^{-1} \mathbf{X}_{T(b)} = \widetilde{\mu}_k \\ \mathrm{Var}^{\mathbb{Q}}[X_k] &= \mathrm{Var}[X_k] - \mathbf{V}_k^T \mathbf{\Sigma}_{T(b)}^{-1} \mathbf{V}_k = \widetilde{\sigma}_k \end{aligned}$$

where $\mathbf{V}_k = \mathrm{Cov}(X_k, \mathbf{X}_{T(b)})$, $\mathbf{\Sigma}_{T(b)} = \mathrm{Cov}(\mathbf{X}_{T(b)}, \mathbf{X}_{T(b)})$. The complexity of these calculation is also $O(t^+(b)^3)$. 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 1.1 is $O(t^+(b)^3)$. In turn, as indicated in equation (9), we have that*

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

for each $\xi > 0$ as $b, \varepsilon^{-1} \nearrow \infty$. Finally, since each proposal of Algorithm 1.2 has an acceptance ratio of order $O\left(t^+(b)\right)$, we conclude that the expected number of function evaluations required to generate a single replication of Algorithm 1.2 is $O\left(t^+(b)^2\right)$.

We now test the performance of our Target Bridge Sampler and compare it against other existing methods in the multisource setting. Mainly, we consider the four algorithms discussed by Dieker and Mandjes (2006). The first one is called the *single twist*, which is a direct derivation of the large deviation principle; the second one is call *cut-and-twist*, which is based

on Boots and Mandjes (2002); the third one is called *random twist*, which is an adaptation of Sadowsky and Bucklew (1990); the fourth 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 $\mathrm{Cov}(X_k, X_j) = (k^{2H_\sigma} + j^{2H_\sigma} - |k-j|^{2H_\sigma})/2$ 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 $H_\sigma = 0.8$, $b = 3$. It is easy to see from the proofs that we can set $\delta = 0$, $M(0) = 1$. Therefore, for $b \geq (H_\mu - H_\sigma)/H_\sigma = 1/4$,

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

We implemented our Algorithm 2, which we denote by TBS in the tables below, and compared it with the results published by Dieker and Mandjes (2006). We follow Dieker and Mandjes (2006)'s 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 and the number of simulations runs needed for that particular algorithm to achieve the criterion are reported in Table 1 and Table 2. The less simulation runs an algorithm needs, the better performance it has. The simulation runs for the first five rows were inferred from the estimated relative variance reported in Dieker and Mandjes (2006).

$n = 300$	Cost of each replication	Estimator	Simulation Runs
Naive	$O(b^3)$	6.12×10^{-4}	32679
Single twist	$O(b^3)$	4.84×10^{-4}	4038
Cut-and-twist	$O(b^4)$	5.95×10^{-4}	703
Random twist	$O(b^3)$	5.50×10^{-4}	3269
Sequential twist	$O(nb^3)$	6.39×10^{-4}	692
Target Bridge	$O(b^3)$	5.84×10^{-4}	26
Benchmark	-	5.75×10^{-4}	-

Table 1: Simulation result of Example 2 with $n = 300$, $b = 3$, $H_\sigma = 0.8$.

It is easily seen from the simulation results that the performance of the Target Bridge Sampler is indeed outstanding. A single replication via TBS is not more expensive than that of any other method and it actually achieves a similar level of relative error with less 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. Also note that number of simulation runs for TBS has been around 25 for both cases. This

$n = 1000$	Cost of each replication	Estimator	Simulation Runs
Naive	$O(b^3)$	-	-
Single twist	$O(b^3)$	1.03×10^{-10}	7257
Cut-and-twist	$O(b^4)$	1.32×10^{-10}	1120
Random twist	$O(b^3)$	1.38×10^{-10}	4287
Sequential twist	$O(nb^3)$	1.41×10^{-10}	1497
TBS	$O(b^3)$	1.36×10^{-10}	24
Benchmark	-	1.372×10^{-10}	-

Table 2: Simulation result of Example 2 with $n = 1000$, $b = 3$, $H_\sigma = 0.8$.

is consistent with Theorem 3 that the coefficient of variation of TBS estimator is bounded as $n \nearrow \infty$.

References

Addie, J., Mannersalo, P. and Norros, I. (1999) Most probable paths and performance formulae for buffers with Gaussian input traffic. *Eur. Trans. Telecomm.* **13**, 183-196.

Adler, R. and Taylor, J. (2007) *Random Fields and Geometry*. Springer. New York.

Adler, R., Blanchet, J. and Liu, J. C. (2008) Efficient simulation for tail probabilities of Gaussian random fields. *Proc. of the Winter Simulation Conference 2008*: 328-336.

Adler, R., Blanchet, J. and Liu, J. C. (2010) Efficient Monte Carlo for high excursions of Gaussian random fields. *Preprint*.

Alzer, H. (1997) On some inequalities for the incomplete gamma function. *Math. Comp.* **66**, 373-389.

Asmussen, S. and Glynn, P. (2007) *Stochastic Simulation: Algorithms and Analysis*. Springer. New York.

Berman, S. (1992) *Sojourns and Extremes of Stochastic Processes*. Wadsworth, Inc. Belmont, CA.

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

Bucklew, J.A. (2004) *Introduction to Rare Event Simulation*. Springer-Verlag. New York.

Burnecki, K. and Michna, Z. (2002) Simulation of Pickands constants. *Prob. Math. Stat.* **22**, 193-199.

Campbell, J., Lo, A., and Mackinlay, A. (1997) *The Econometrics of Financial Markets*. Princeton University Press. Princeton, NJ.

Dębicki, K. (1999). A note on LDP for supremum of Gaussian processes over infinite horizon. *Stat. Prob. Letters.* **44**, 211-220.

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

Dieker, A.B. (2005) Extremes of Gaussian processes over an infinite horizon. *Stoch. Proc. 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. Comp. Sim.* **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. Cam. Phil. Soc.,* **118**, 363-374.

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

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

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

Giordano, S., Gubinelli, M. and Pagano, M. (2007) Rare Events of Gaussian Processes: A Performance Comparison Between Bridge Monte-Carlo and Importance Sampling. Chapter of: *Next Generation Teletraffic and Wired/Wireless Advanced Networking.* Springer. Berlin.

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

Juneja, S. and Shahabuddin, P. (2006) Rare Event Simulation Techniques: An Introduction and Recent Advances. *Handbook on Simulation.* Chapter 11. Elsevier. Editors: Shane Henderson and Barry Nelson 291–350.

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

Mandjes, M. (2007) *Large Deviations for Gaussian Queues.* Wiley. New York.

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

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

Pickands III, J. (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. Inform. Theory.* **36**, 579-588.

Siegmund, D. (1976) Importance Sampling in the Monte Carlo Study of Sequential Tests, *Ann. Stat.* **4**, 673-684.