Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

LOGARITHMICALLY EFFICIENT ESTIMATION OF THE TAIL OF THE MULTIVARIATE NORMAL DISTRIBUTION

Zdravko Botev

School of Mathematics and Statistics
The University of New South Wales
Sydney, NSW 2052, AUSTRALIA

Daniel Mackinlay

School of Mathematics and Statistics
The University of New South Wales
Sydney, NSW 2052, AUSTRALIA

Yi-Lung Chen

School of Mathematics and Statistics
The University of New South Wales
Sydney, NSW 2052, AUSTRALIA

ABSTRACT

Simulation from the tail of the multivariate normal density has numerous applications in statistics and operations research. Unfortunately, there is no simple formula for the cumulative distribution function of the multivariate normal law, and simulation from its tail can frequently only be approximate. In this article we present an asymptotically efficient Monte Carlo estimator for quantities related to the tail of the multivariate normal distribution. The estimator leverages upon known asymptotic approximations. In addition, we generalize the notion of asymptotic efficiency of Monte Carlo estimators of rare-event probabilities to the sampling properties of Markov chain Monte Carlo algorithms. Regarding these new notions, we propose a simple and practical Markov chain sampler for the normal tail that is asymptotically optimal. We then give a numerical example from finance that illustrates the benefits of an asymptotically efficient Markov chain Monte Carlo sampler.

1 INTRODUCTION

The importance of the normal distribution in mathematics and the applied sciences is well-known. One of the key properties of the univariate Gaussian distribution is Mill's ratio (Mills 1926):

$$\frac{\overline{\Phi}(\gamma)}{\phi(\gamma)} \simeq \frac{1}{\gamma}, \quad \gamma \uparrow \infty,$$

where ϕ and $\overline{\Phi}$ denote the standard normal density and its right tail, respectively. (Here the notation $f(x) \simeq g(x)$ as $x \to a$ stands for $\lim_{x\to a} f(x)/g(x) = 1$. Similarly, we define $f(x) = \mathcal{O}(g(x)) \Leftrightarrow \lim_{x\to a} |f(x)/g(x)| < \text{const.} < \infty$; $f(x) = o(g(x)) \Leftrightarrow \lim_{x\to a} f(x)/g(x) = 0$; also, $f(x) = \Theta(g(x)) \Leftrightarrow f(x) = \mathcal{O}(g(x))$ and $g(x) = \mathcal{O}(f(x))$.)

Given this univariate result, it is then natural to ask what its multivariate analogue is. The first result in this direction was established by (Savage 1962) and states that in d-dimensions:

$$\frac{\overline{\Phi}_\Sigma(\gamma\Sigma\boldsymbol{c})}{\phi_\Sigma(\gamma\Sigma\boldsymbol{c})} \simeq \frac{1}{\gamma^d \prod_{k=1}^d c_k}, \quad \boldsymbol{c} > \mathbf{0}, \quad \gamma \uparrow \infty, \tag{1}$$

where $\phi_\Sigma(\boldsymbol{y})$ and $\overline{\Phi}_\Sigma(\boldsymbol{y}) = \mathbb{P}(\boldsymbol{Y} \geq \boldsymbol{y})$, $\boldsymbol{Y} \sim \mathsf{N}(\mathbf{0}, \Sigma)$, denote the density and tail, respectively, of the multivariate $\mathsf{N}(\mathbf{0}, \Sigma)$ distribution with covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. This result is the simplest extension

of Mill's result to multiple dimensions. Note, however, that the form of the threshold vector, $\Sigma \boldsymbol{c}$, is a weighted linear combination of the columns of Σ, and that this condition can be quite restrictive. It is for this reason that more recently (Hashorva and Hüsler 2003) have sought a more general asymptotic version of the multivariate Mill's ratio.

This more general version of the Mill's ratio can be stated as follows: we wish to estimate

$$\ell(\gamma) = \mathbb{P}(\boldsymbol{Y} \geq \boldsymbol{l}(\gamma)), \quad \boldsymbol{Y} \sim \mathsf{N}(\boldsymbol{0}, \Sigma), \tag{2}$$

where $\max_i l_i > 0$ (l_i is the i-th component of vector $\boldsymbol{l} \in \mathbb{R}^d$), and at least one component of $\boldsymbol{l}(\gamma)$ diverges to ∞, that is, $\lim_{\gamma \uparrow \infty} \|\boldsymbol{l}(\gamma)\| = \infty$. Although (Hashorva and Hüsler 2003) provide the asymptotic approximation to ℓ (see Section 2), it is desirable to also have an asymptotically efficient Monte Carlo estimator of this quantity.

In this article we construct a novel unbiased and logarithmically efficient estimator for ℓ. The logarithmic efficiency is with respect to $\gamma \uparrow \infty$, and refers to an estimator whose relative error grows subexponentially as γ becomes larger and larger. Note that (Botev 2016) introduced a Monte Carlo estimator of (1) that enjoys the much stronger *vanishing relative error* property in the limit as $\gamma \uparrow \infty$ (see, for example, (Kroese et al. 2011, Chapter 10)). Unfortunately, the results in (Botev 2016) do not apply to the more general Mill's ratio in (2). Thus, this aspect of the present work extends and complements the results in (Botev 2016, Botev and L'Ecuyer 2015, Botev et al. 2015).

In addition to this, we introduce a new notion called *logarithmically efficient Markov chain Monte Carlo* (LE-MCMC). What distinguishes such an MCMC sampler from the usual MCMC is that LE-MCMC provides a theoretical guarantee that the reduction of its mixing speed as $\gamma \uparrow \infty$ can be countered by growing the length of the Markov chain at a sub-exponential, or polynomial, rate.

In this article, we thus present a logarithmically efficient MCMC for simulating from the conditional distribution:

$$\pi(\boldsymbol{y}) = \frac{\phi_\Sigma(\boldsymbol{y}) \mathbb{I}\{\boldsymbol{y} \geq \boldsymbol{l}(\gamma)\}}{\ell(\gamma)}. \tag{3}$$

There are a number of applications of Mill's ratio in survival analysis and censored data, see, for example, (Klein and Moeschberger 2005). Here, we provide an application of estimating (2) and simulating from (3) from financial engineering (Kroese et al. 2011, Chapter 15). In particular, we show how we can price with high accuracy a digital hindsight put option when the probability of being knocked out is very high.

The rest of the paper is organized as follows. In Section 2 we provide the notation and background needed to state the asymptotic result of (Hashorva and Hüsler 2003). In Section 3 we present our novel estimator of (2), and show that it is logarithmically efficient. Next, in Section 4 we introduce the concept of LE-MCMC and describe a logarithmically efficient sampler. This is followed by a numerical example with pricing a down-and-out option contract. Finally, we explain how the results can be generalized to the setting of a multivariate normal density constrained to more general, but smooth, regions, and draw conclusions.

2 ASYMPTOTIC BEHAVIOR OF THE MULTIVARIATE NORMAL TAIL

Let P be a permutation matrix which maps the vector $(1, \ldots, d)^\top$ into the permutation $\boldsymbol{p} = (p_1, \ldots, p_d)^\top$, that is, $\mathrm{P}(1, \ldots, d)^\top = \boldsymbol{p}$. Note that $\ell(\gamma) = \mathbb{P}(\mathrm{P}\boldsymbol{Y} \geq \mathrm{P}\boldsymbol{l}(\gamma))$ for any permutation $\boldsymbol{p}$, and $\mathrm{P}\boldsymbol{Y} \sim \mathsf{N}(\boldsymbol{0}, \mathrm{P}\Sigma\mathrm{P}^\top)$. We will specify $\boldsymbol{p}$ shortly.

Define the convex quadratic programming:

$$\begin{aligned} \min_{\boldsymbol{x}} \ & \frac{1}{2} \boldsymbol{x}^\top (\mathrm{P}\Sigma\mathrm{P}^\top)^{-1} \boldsymbol{x} \\ \text{subject to: } & \boldsymbol{x} \geq \mathrm{P}\boldsymbol{l}(\gamma). \end{aligned} \tag{4}$$

The Karush-Kuhn-Tucker (KKT) conditions are a necessary and sufficient condition to find the unique solution:

$$\begin{aligned} (\mathrm{P}\Sigma\mathrm{P}^\top)^{-1}\boldsymbol{x} - \boldsymbol{\lambda} &= \mathbf{0} \\ \boldsymbol{\lambda} \geq \mathbf{0},\ \ \mathrm{P}\boldsymbol{l} - \boldsymbol{x} &\leq \mathbf{0} \\ \boldsymbol{\lambda}^\top(\mathrm{P}\boldsymbol{l} - \boldsymbol{x}) &= 0\ , \end{aligned} \tag{5}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^d$ is the Lagrange multiplier. Denote the number of active constraints in (4) by d_1 and the number of inactive constraints as d_2, so that $d_1 + d_2 = d$. Note that the number of active constraints $d_1 \geq 1$, because otherwise the solution to (4) is $\boldsymbol{x} = \mathbf{0}$, which implies $\mathrm{P}\boldsymbol{l} \leq \mathbf{0}$, thus reaching a contradiction.

Given the partition $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^\top, \boldsymbol{\lambda}_2^\top)^\top$ with $\dim(\boldsymbol{\lambda}_1) = d_1$ and $\dim(\boldsymbol{\lambda}_2) = d_2$, one can select the permutation vector $\boldsymbol{p}$ and the corresponding matrix P in such a way that all the active constraints in (5) correspond to $\boldsymbol{\lambda}_1 > \mathbf{0}$ and all the inactive ones to $\boldsymbol{\lambda}_2 = \mathbf{0}$.

Assumption 1 (Active Constraint Set.) Henceforth, we assume that this reordering of the variables via the permutation operator P has been applied as a preprocessing step to both $\boldsymbol{l}$ and Σ so that $\mathrm{P}\boldsymbol{l} = \boldsymbol{l}$ and $\mathrm{P}\Sigma\mathrm{P}^\top = \Sigma$.

If we partition $\boldsymbol{x}, \boldsymbol{l}$, and

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

then the KKT equations tell us that the optimal solution $\boldsymbol{x}^*$ is:

$$\begin{aligned} \boldsymbol{x}_1^* &= \Sigma_{11}\boldsymbol{\lambda}_1 = \boldsymbol{l}_1(\gamma) \\ \boldsymbol{x}_2^* &= \Sigma_{21}\boldsymbol{\lambda}_1 = \Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{l}_1(\gamma) > \boldsymbol{l}_2(\gamma) \end{aligned} \tag{6}$$

with the global minimum $\frac{1}{2}(\boldsymbol{x}^*)^\top\Sigma^{-1}\boldsymbol{x}^* = \frac{1}{2}(\boldsymbol{x}_1^*)^\top\boldsymbol{\lambda}_1 = \frac{1}{2}\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1$.

Note that $\boldsymbol{x}^*(\gamma)$ is implicitly a function of γ and that in general the active constraint set of (4) and its size, d_1, also depends on the value of γ through $\boldsymbol{l}(\gamma)$.

Assumption 2 (Independence of active set from γ) We henceforth assume that $\|\boldsymbol{l}(\gamma)\|$ diverges to infinity as $\gamma \uparrow \infty$ in such a way that, for large enough γ, the active constraint set of (4) becomes independent of γ.

With these definitions and assumptions we can now formulate the result of (Hashorva and Hüsler 2003).

Theorem 1 (Mill's Ratio For Multivariate Normal) Under the assumptions above, if $\boldsymbol{Y} \sim \mathsf{N}(\mathbf{0}, \Sigma)$ with $\boldsymbol{Y}^\top = (\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top)$, $\dim(\boldsymbol{Y}_1) = d_1$, then as $\gamma \uparrow \infty$, we have:

$$\mathbb{P}(\boldsymbol{Y} \geq \boldsymbol{l}(\gamma)) = \frac{\mathbb{P}(\boldsymbol{Y}_2 \geq \boldsymbol{l}_\infty \,|\, \boldsymbol{Y}_1 = \mathbf{0})}{(2\pi)^{d_1/2}|\Sigma_{11}|^{1/2}\prod_{k=1}^{d_1}\boldsymbol{e}_k^\top\Sigma_{11}^{-1}\boldsymbol{l}_1} \exp\left(-\frac{\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}{2}\right)(1 + o(1)),$$

where $\boldsymbol{l}_\infty \stackrel{\text{def}}{=} \lim_{\gamma\uparrow\infty}(\boldsymbol{l}_2(\gamma) - \boldsymbol{x}_2^*(\gamma))$ with $\boldsymbol{l}_\infty \leq \mathbf{0}$, and $\boldsymbol{e}_k$ is the unit vector with a 1 in the k-th position.

Given this asymptotic result, we would like to construct a simple and unbiased Monte Carlo estimator of $\ell(\gamma)$, paying close attention to its error behavior as the tail probability becomes smaller and smaller with increasing γ. In the next section we present one such estimator.

3 LOGARITHMICALLY EFFICIENT ESTIMATION OF GAUSSIAN TAIL

The crude Monte Carlo estimator of ℓ is, of course:

$$\mathbb{I}\{\boldsymbol{Y} \geq \boldsymbol{l}\},$$

where $\boldsymbol{Y} \sim \mathsf{N}(\mathbf{0}, \Sigma)$ is simulated under the original measure $\mathbb{P}$. It is clear that this estimator is not efficient, because $\mathbb{V}\mathrm{ar}(\mathbb{I}\{\boldsymbol{Y} \geq \boldsymbol{l}\})/\ell(\gamma)^2 \simeq 1/\ell(\gamma)$ grows at an exponential rate as $\gamma \uparrow \infty$.

Instead, we consider the change of measure $\mathbb{P}_{\boldsymbol{\mu}}$, under which $\boldsymbol{Y} \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$. This gives the importance sampling estimator:

$$\hat{\ell}(\gamma) = \exp\left(\tfrac{1}{2}\boldsymbol{\mu}^\top \Sigma^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^\top \Sigma^{-1}\boldsymbol{Y}\right) \mathbb{I}\{\boldsymbol{Z} \geq \boldsymbol{l}(\gamma)\}, \qquad \boldsymbol{Y} \sim \mathbb{P}_{\boldsymbol{\mu}}. \tag{7}$$

We then have the following efficiency result.

Theorem 2 (Logarithmic Efficiency) Suppose assumptions 1. and 2. in Section 2 hold. If we select $\boldsymbol{\mu} = \boldsymbol{x}^*$, where $\boldsymbol{x}^*$ is the solution to (4), then the estimator $\hat{\ell}$ is asymptotically efficient:

$$\liminf_{\gamma} \frac{\ln \mathbb{E}\hat{\ell}^2(\gamma)}{\ln \ell(\gamma)} = 2.$$

Proof. Since the estimator is unbiased, we need only consider its second moment. The second moment under $\mathbb{P}_{\boldsymbol{\mu}}$ is:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\mu}}\hat{\ell}^2(\gamma) &= \mathbb{E}_{\mathbf{0}} \exp\left(\frac{1}{2}\boldsymbol{\mu}^\top \Sigma^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^\top \Sigma^{-1}\boldsymbol{Y}\right) \mathbb{I}\{\boldsymbol{Y} \geq \boldsymbol{l}(\gamma)\} \\
&= \exp\left(\boldsymbol{\mu}^\top \Sigma^{-1}\boldsymbol{\mu}\right) \mathbb{P}_{-\boldsymbol{\mu}}(\boldsymbol{Y} \geq \boldsymbol{l}(\gamma)) \\
&= \exp\left(\boldsymbol{\mu}^\top \Sigma^{-1}\boldsymbol{\mu}\right) \mathbb{P}_{\mathbf{0}}(\boldsymbol{Y} \geq \boldsymbol{l} + \boldsymbol{\mu}), \text{ where } \boldsymbol{\mu} = (\boldsymbol{l}_1^\top, \boldsymbol{l}_1^\top \Sigma_{11}^{-1}\Sigma_{21})^\top.
\end{aligned}$$

Next, we apply Theorem 1 on the term $\mathbb{P}(\boldsymbol{Y} \geq \tilde{\boldsymbol{l}})$, where

$$\tilde{\boldsymbol{l}} = \begin{pmatrix} 2\boldsymbol{l}_1 \\ \boldsymbol{l}_2 + \Sigma_{12}\Sigma_{11}^{-1}\boldsymbol{l}_1 \end{pmatrix}.$$

Keeping in mind that $\boldsymbol{l}$ is such that $\boldsymbol{l}_2 < \Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{l}_1$, the solution to (4) is

$$\begin{aligned}
\tilde{\boldsymbol{x}}_1^* &= \tilde{\boldsymbol{l}}_1(\gamma) = 2\boldsymbol{l}_1 \\
\tilde{\boldsymbol{x}}_2^* &= \Sigma_{21}\Sigma_{11}^{-1}\tilde{\boldsymbol{l}}_1(\gamma) = 2\Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{l}_1 > \tilde{\boldsymbol{l}}_2 = \boldsymbol{l}_2 + \Sigma_{12}\Sigma_{11}^{-1}\boldsymbol{l}_1.
\end{aligned}$$

Theorem 1 then yields:

$$\mathbb{P}(\boldsymbol{Y} \geq \tilde{\boldsymbol{l}}(\gamma)) \simeq \frac{\mathbb{P}(\boldsymbol{Y}_2 \geq \boldsymbol{l}_\infty \,|\, \boldsymbol{Y}_1 = \mathbf{0})}{(2\pi)^{d_1/2}|\Sigma_{11}|^{1/2} \prod_{k=1}^{d_1} \boldsymbol{e}_k^\top \Sigma_{11}^{-1}(2\boldsymbol{l}_1)} \exp\left(-\frac{(2\boldsymbol{l}_1)^\top \Sigma_{11}^{-1}(2\boldsymbol{l}_1)}{2}\right).$$

Therefore,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\mu}}\hat{\ell}^2(\gamma) &= \frac{\mathbb{P}(\boldsymbol{Y}_2 \geq \boldsymbol{l}_\infty \,|\, \boldsymbol{Y}_1 = \mathbf{0})}{(2\pi)^{d_1/2}|\Sigma_{11}|^{1/2} \prod_{k=1}^{d_1} \boldsymbol{e}_k^\top \Sigma_{11}^{-1}(2\boldsymbol{l}_1)} \exp\left(\boldsymbol{\mu}^\top \Sigma^{-1}\boldsymbol{\mu} - 2\boldsymbol{l}_1^\top \Sigma_{11}^{-1}\boldsymbol{l}_1\right)(1 + o(1)) \\
&= \frac{c_1}{\prod_{k=1}^{d_1} \boldsymbol{e}_k^\top \Sigma_{11}^{-1}(2\boldsymbol{l}_1)} \exp\left(-\boldsymbol{l}_1^\top \Sigma_{11}^{-1}\boldsymbol{l}_1\right)(1 + o(1)),
\end{aligned}$$

where $c_1, c_2, \ldots$ are generic constants that do not depend on γ. It follows that we have the asymptotic inequality

$$\frac{\mathbb{E}_{\boldsymbol{\mu}}\hat{\ell}^2(\gamma)}{\ell^2(\gamma)} \lesssim c_2 \prod_{k=1}^{d_1} \boldsymbol{e}_k^\top \Sigma_{11}^{-1}\boldsymbol{l}_1 = c_2 \prod_{k=1}^{d_1} \lambda_k, \qquad \gamma \uparrow \infty$$

and therefore the relative error grows only at a polynomial rate in $\boldsymbol{l}_1$, implying a logarithmically efficient estimator. □

We will give an example of the practical performance of this estimator in Section 5. For the time being, notice how the magnitude of the relative error depends on the product of the Lagrange multipliers, $\prod_{k=1}^{d_1} \lambda_k$, that correspond to the active constraint set. Thus, we can expect this estimator to work well when there are few active constraints in (4), and to work poorly when the number of active constraints is $d_1 = d$.

4 LOGARITHMICALLY EFFICIENT MARKOV CHAIN MONTE CARLO

Next, suppose we wish to simulate from the conditional density (3), and we use an Markov chain independence sampler (Kroese et al. 2011, Chapter 6) with proposal density $\phi_\Sigma(\boldsymbol{z} - \boldsymbol{\mu})$. Then, given the Markov chain is in the state $\boldsymbol{Y} = \boldsymbol{y}$, we accept the proposal move $\boldsymbol{Z} \sim \phi_\Sigma(\boldsymbol{z} - \boldsymbol{\mu})$ with probability

$$\alpha(\boldsymbol{Z}|\boldsymbol{y}) = 1 \wedge \exp(-\boldsymbol{\mu}^\top \Sigma^{-1}(\boldsymbol{Z} - \boldsymbol{y}))\, \mathbb{I}\{\boldsymbol{Z} \geq \boldsymbol{l}\}.$$

This yields the following MCMC algorithm for approximate simulation from (3).

Algorithm 1 : LE-MCMC for $\pi(\boldsymbol{y})$.

Require: $\boldsymbol{\mu}, \Sigma, \boldsymbol{l}$; an initial state of the chain $\boldsymbol{Y}_0$, and length of chain n.
  for $t = 1, \ldots, n$ **do**
    Simulate $\boldsymbol{Z} \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$ and $U \sim \mathsf{U}(0, 1)$, independently
    if $U < \alpha(\boldsymbol{Z}|\boldsymbol{Y}_{t-1})$ **then**
      $\boldsymbol{Y}_t \leftarrow \boldsymbol{Z}$
    else
      $\boldsymbol{Y}_t \leftarrow \boldsymbol{Y}_{t-1}$
  return $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$

A natural question then arises with this sampler. What is the mixing rate of the sampler as a function of n and how is this mixing rate affected by the tail parameter γ? The answer to the first part of the question has been known for a long time (Mengersen and Tweedie 1996). It is the answer to the second part of the question that concerns us here.

To this end, let $\kappa_t(A|\boldsymbol{x})$ be the t-step transition kernel of a Markov chain with limiting and stationary density π. The total variation distance between the t-step distribution and the target is thus defined as:

$$D_t(\boldsymbol{x}) = \sup_A |\kappa_t(A|\boldsymbol{x}) - \pi(A)|,$$

where $\pi(A) \stackrel{\text{def}}{=} \int_A \pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ is the measure of a Borel set A on $\mathbb{R}^d$. If we wish to keep the total variation error below a given tolerance ϵ, we then need to take at least

$$T = \min\{t : D_t(\boldsymbol{x}) \leq \epsilon\}$$

number of steps. If π depends on the parameter γ, then so does T. Thus, we define the following efficiency notion. We say that a *Markov chain is strongly efficient* if $\limsup_\gamma T(\gamma) < \infty$ and a *Markov chain is logarithmically efficient* if

$$\limsup_\gamma \frac{\ln T(\gamma)}{\ln \gamma} < \infty.$$

In addition, if the statements in the definition are true for $T = \min\{t : \sup_{\boldsymbol{x}} D_t(\boldsymbol{x}) \leq \epsilon\}$, then, in keeping with standard nomenclature, we say that we have a *uniformly* strongly (or logarithmically) efficient Markov chain.

Thus, to ensure that we sample within a total variation error of ϵ from the target π, the length of a strongly efficient chain need not increase as γ becomes larger. In contrast, the length of a logarithmically

efficient chain has to increase at a polynomial rate in γ to keep the error in check, and the length of an inefficient chain has to grow exponentially.

Proposition 1 (Logarithmically Efficient Markov Chain) The Markov chain in Algorithm 1 is (uniformly) logarithmically efficient.

Proof. First note that $\boldsymbol{y}^\top \Sigma^{-1}\boldsymbol{\mu} = \boldsymbol{y}^\top \boldsymbol{\lambda} = \boldsymbol{y}_1 \boldsymbol{\lambda}_1$ from (5), and hence

$$c(\gamma) = \sup_{\boldsymbol{y}} \frac{\pi(\boldsymbol{y})}{\phi_\Sigma(\boldsymbol{y}-\boldsymbol{\mu})} = \sup_{\boldsymbol{y}\geq \boldsymbol{l}} \frac{\exp\left(\frac{1}{2}\boldsymbol{\mu}^\top \Sigma^{-1}\boldsymbol{\mu} - \boldsymbol{y}^\top \Sigma^{-1}\boldsymbol{\mu}\right)}{\ell_\gamma}$$
$$= \frac{\exp\left(\frac{1}{2}\boldsymbol{\mu}^\top \Sigma^{-1}\boldsymbol{\mu}\right)}{\ell_\gamma} \exp(-\min_{\boldsymbol{y}_1 \geq \boldsymbol{l}_1} \boldsymbol{y}_1^\top \Sigma_{11}^{-1}\boldsymbol{l}_1) = \frac{\exp\left(-\frac{1}{2}\boldsymbol{l}_1^\top \Sigma_{11}^{-1}\boldsymbol{l}_1\right)}{\ell_\gamma}.$$

Then, by (Mengersen and Tweedie 1996, Theorem 2.1) we know that the Markov chain is uniformly ergodic and

$$\sup_{\boldsymbol{x}} D_t(\boldsymbol{x}) \leq 2(1-c^{-1}(\gamma))^t \leq 2\left(1 - \frac{c_1}{\prod_{k=1}^{d_1}\lambda_k}(1+o(1))\right)^t.$$

Thus, to ensure the bound remains below ϵ uniformly as $\gamma \uparrow \infty$, we need to run the independence sampler for

$$T \gtrsim \left\lceil \ln(2/\epsilon)\frac{\prod_{k=1}^{d_1}\lambda_k}{c_1} \right\rceil$$

number of steps. In other words, the length of the chain increases at a polynomial rate, and the chain is uniformly logarithmically efficient. □

5 NUMERICAL EXAMPLE:DOWN-AND-OUT CALL OPTION

We consider the pricing and simulation of a discretely-monitored digital down-and-out call option (Kroese et al. 2011, Chapter 15) with payoff defined by

$$\mathbb{I}\{S_{t_n} > K\}\mathbb{I}\{\min_{1\leq i\leq n-1} S_{t_i} \geq \beta\},$$

where:

- under the risk-neutral measure $\mathbb{Q}$, the stock price process $\{S_t, t \geq 0\}$ is governed by the geometric Brownian motion model
$$S_t = S_0 \exp((r-\sigma^2/2)t + \sigma W_t)$$
with $\{W_t, t \geq t\}$ being a Wiener process; r being the annual interest rate; and σ being the volatility parameter;
- β is a knock-out barrier;
- T is the time of maturity of the option with discrete monitoring over $t_1 < \cdots < t_n = T$;
- $K > \beta$ is the strike price of the exotic option.

The value of the option at time $t = 0$ is given by the expected value of the discounted payoff:

$$C_0 = \exp(-rT)\mathbb{Q}(\min_{i\leq n-1} S_{t_i} \geq \beta, S_{t_n} > K).$$

We are interested in simulating from the law of $\mathbb{Q}$, conditional on the option being worth something, that is, the law of $\mathbb{Q}$, conditional on the event $\{\min_{i\leq n-1} S_{t_i} \geq \beta, S_{t_n} > K\}$. This is thus equivalent to simulating from the conditional pdf:

$$\pi(\boldsymbol{z}) = \frac{\phi_\Sigma(\boldsymbol{z})\mathbb{I}\{\boldsymbol{z} \geq \boldsymbol{l}\}}{\ell(\gamma)},$$

where:

- Σ is a covariance matrix with elements $\Sigma_{ij} = \min\{t_i, t_j\}$, corresponding to the covariance function of the Wiener process;
- the vector $\boldsymbol{l}$ has components

$$l_k = \frac{\ln(\beta/S_0) - (r - \sigma^2/2)t_k}{\sigma}, \quad k \leq n-1, \quad l_n = \frac{\ln(K/S_0) - (r - \sigma^2/2)T}{\sigma}.$$

Then, given a draw $\boldsymbol{Z} \sim \pi$, we can map it to a stock price trajectory for which the option is not worthless:

$$S_{t_k} = S_0 \exp((r - \sigma^2/2)t_k + \sigma Z_k).$$

Figure 5 shows one such stock price trajectory, conditional on the event $\{\min_{i\leq n-1} S_{t_i} \geq \beta, S_{t_n} > K\}$, where $\beta = S_0 = 100, \sigma = 0.1, r = 0.03, T = 4/12, K = 130$, and $d = 264$.

The figure was created using the last state of the chain generated by Algorithm 1 with $n = 10^5$. Figure 5 shows the estimated auto-correlation function for Z_3 in $\boldsymbol{Z}$. The figure suggests that the chain becomes stationary after approximately 100 steps.

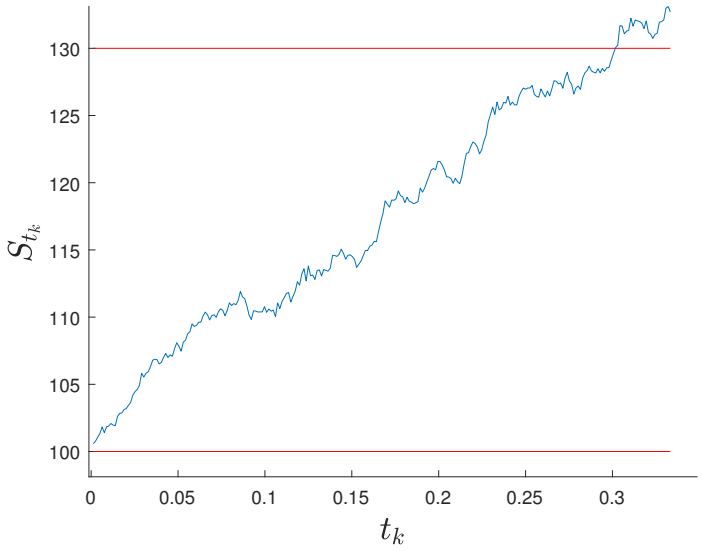

Figure 1: Stock price trajectory conditional on positive payoff. The barrier, $\beta = 100$, and strike price, $K = 130$, are shown in red. Here $S_0 = 100, \sigma = 0.1, r = 0.03, T = 4/12$, and $d = 264$.

We compare our LE-MCMC sampler against the *random walk* Metropolis-Hastings sampler (Kroese et al. 2011, Chapter 6). Given the current state, $\boldsymbol{y}$, of the chain, the proposal move is generated via

$$\boldsymbol{Z} = s \times \mathrm{t}_\nu(\boldsymbol{0}, \Sigma) + \boldsymbol{y}, \tag{8}$$

where:

- $s > 0$ is a scale parameter to be tuned for optimal performance;
- $\mathrm{t}_\nu(\mathbf{0}, \Sigma)$ denotes a random variable drawn from the multivariate student distribution with ν degrees of freedom with location $\mathbf{0}$ and scale matrix Σ.

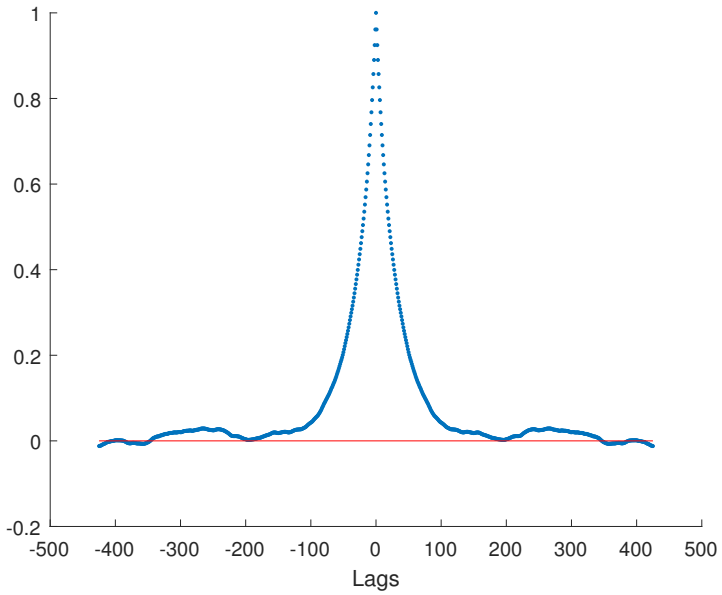

Figure 2: Estimated autocorrelation function of chain in Algorithm 1 with length $n = 10^5$.

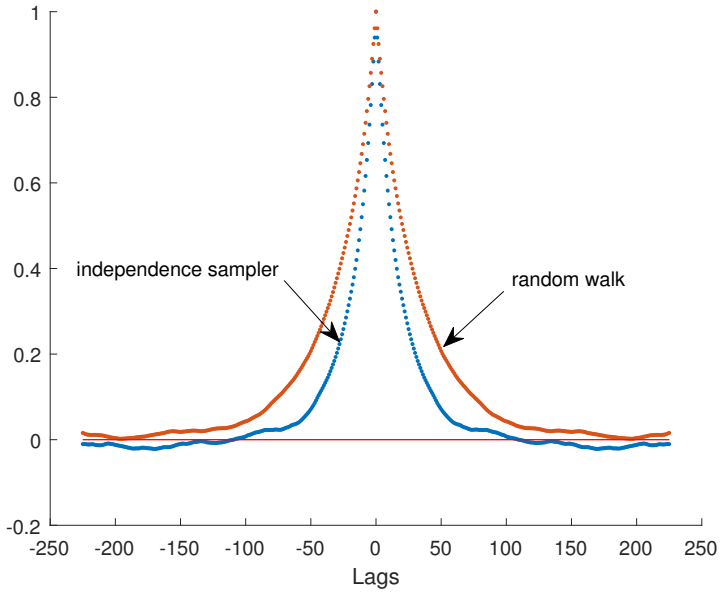

Figure 3: Independence sampler in Algorithm 1 versus the (optimized) random walk sampler with $s = 0.03$.

In implementing the sampler, we attempted to optimize the mixing of the Markov chain by tuning the scale parameter s for $\nu = 100$. The fastest decaying (estimated) autocorrelation function was observed for values of around $s = 0.03$. Figure 5 shows the resulting autocorrelation plot superimposed on the plot from Figure 5. It appears that the proposed independence sampler achieves stationarity faster than the best random walk sampler with proposal of the form (8).

Finally, we note that with $n = 10^6$ the logarithmically efficient estimator (7) yielded the following estimate of $\mathbb{Q}(\min_{i\leq n-1} S_{t_i} \geq \beta, S_{t_n} > K)$ with 95% (asymptotic) confidence interval:

$$[1.88 \pm 0.014] \times 10^{-6}.$$

This small quantity is difficult to estimate efficiently via crude Monte Carlo.

6 EXTENSION TO SMOOTH DOMAINS

A possible extension of the approach has been suggested by (Gollwitzer and Rackwitz 1988), who consider the following more general setting. Suppose one wishes to estimate:

$$p = \mathbb{P}(\boldsymbol{g}(\boldsymbol{Z}) \geq \mathbf{0}),$$

where $\boldsymbol{Z} \sim \mathsf{N}(\mathbf{0}, \mathrm{I}_m)$, and $\boldsymbol{g}(\boldsymbol{z}) = (g_1(\boldsymbol{z}), \ldots, g_d(\boldsymbol{z}))^\top$ is a vector-valued function, $\boldsymbol{g} : \mathbb{R}^m \mapsto \mathbb{R}^d,\ m > d$, such that all g_j's are twice continuously differentiable and $g_j(\mathbf{0}) < 0$ for at least one j. Let $\mathrm{G}(\boldsymbol{z}) = \nabla \boldsymbol{g}(\boldsymbol{z})$ be an $d \times m$ Jacobian matrix containing all gradients, $\nabla g_j,\ j = 1, \ldots, d$, as row vectors.

Define the non-linear optimization program

$$\begin{aligned} \min_{\boldsymbol{z}} \ & \frac{1}{2}\boldsymbol{z}^\top \boldsymbol{z} \\ \text{subject to: } & \boldsymbol{g}(\boldsymbol{z}) \geq \mathbf{0}, \end{aligned} \tag{9}$$

whose solution, $\boldsymbol{z}^* \neq \mathbf{0}$, satisfies the necessary conditions ($\mathrm{G} = \mathrm{G}(\boldsymbol{z}^*)$):

$$\begin{aligned} \boldsymbol{z}^* - \mathrm{G}^\top \boldsymbol{\eta} &= \mathbf{0} \\ \boldsymbol{\eta} \geq \mathbf{0}, \ \ -\boldsymbol{g}(\boldsymbol{z}^*) &\leq \mathbf{0} \\ \boldsymbol{\eta}^\top \boldsymbol{g}(\boldsymbol{z}^*) &= 0\ , \end{aligned}$$

We assume, again, that the first d_1 elements, $\boldsymbol{\eta}_1$, of the Lagrange multiplier $\boldsymbol{\eta}$ are positive (corresponding to the active constraints).

Then, as $\|\boldsymbol{z}^*\| \uparrow \infty$, we have the asymptotic result (Gollwitzer and Rackwitz 1988):

$$\mathbb{P}(\boldsymbol{g}(\boldsymbol{Z}) \geq \mathbf{0}) = \Theta(\mathbb{P}(\boldsymbol{Y} \geq \boldsymbol{l})),$$

where $\boldsymbol{Y} \sim \mathsf{N}(\mathbf{0}, \Sigma)$ with $\Sigma = \mathrm{G}\mathrm{G}^\top$, and

$$\boldsymbol{l} = \mathrm{G}\boldsymbol{z}^* - \boldsymbol{g}(\boldsymbol{z}^*) = \Sigma\boldsymbol{\eta} - \boldsymbol{g}(\boldsymbol{z}^*).$$

Due to our partitioning of active and inactive constraints, we have $\boldsymbol{g}_1(\boldsymbol{z}^*) = \mathbf{0}$ and $\boldsymbol{g}_2(\boldsymbol{z}^*) > \mathbf{0}$. Therefore, the partitioned $\boldsymbol{l} = (\boldsymbol{l}_1^\top, \boldsymbol{l}_2^\top)^\top$ is given by

$$\begin{aligned} \boldsymbol{l}_1 &= \Sigma_{11}\boldsymbol{\eta}_1 \\ \boldsymbol{l}_2 &= \Sigma_{21}\boldsymbol{\eta}_1 - \boldsymbol{g}_2(\boldsymbol{z}^*) < \Sigma_{21}\boldsymbol{\eta}_1. \end{aligned}$$

The last applied to (4) with $\mathrm{P} = \mathrm{I}$ yields the solution (see (6))

$$\begin{aligned} \boldsymbol{x}_1^* &= \boldsymbol{l}_1 = \Sigma_{11}\boldsymbol{\eta}_1 \\ \boldsymbol{x}_2^* &= \Sigma_{21}\Sigma_1 1^{-1}\boldsymbol{l}_1 = \Sigma_{21}\boldsymbol{\eta}_1 > \boldsymbol{l}_2. \end{aligned}$$

This implies that, in fact, $\mathbb{P}(\boldsymbol{Y} \geq \boldsymbol{l}) = \Theta(\mathbb{P}(\boldsymbol{Y}_1 \geq \boldsymbol{l}_1))$, and hence $\mathbb{P}(\boldsymbol{g}(\boldsymbol{Z}) \geq \boldsymbol{0}) = \Theta(\mathbb{P}(\boldsymbol{Y}_1 \geq \boldsymbol{l}_1))$, and only the active constraints matter asymptotically.

The key consequence of the above discussion is that the asymptotically optimal estimation of $p = \mathbb{P}(\boldsymbol{g}(\boldsymbol{Z}) \geq \boldsymbol{0})$ is equivalent to the asymptotically optimal estimation of $\mathbb{P}(\boldsymbol{Y} \geq \boldsymbol{l})$, which we already know how to handle. Therefore, without further work, we can conclude that the following estimator of p:

$$\hat{p} = \exp(\|\boldsymbol{\mu}\|_2^2/2 - \boldsymbol{\mu}^\top \boldsymbol{Z})\mathbb{I}\{\boldsymbol{g}(\boldsymbol{Z}) \geq \boldsymbol{0}\}, \qquad \boldsymbol{Z} \sim \mathsf{N}(\boldsymbol{\mu}, \mathrm{I}),$$

where $\boldsymbol{\mu} = \boldsymbol{z}^*$ is the solution to (9), is logarithmically efficient as $\|\boldsymbol{z}^*\| \uparrow \infty$.

7 CONCLUSIONS

In this article we reviewed recent generalizations of the Mill's ratio to the multivariate normal setting. We then used these asymptotic results to motivate the construction of an exponentially tilted estimator that is weakly efficient in a very general rare-event probability estimation setting. A financial application shows that this simple estimator can sometimes yield low variance estimates in cases where the covariance function Σ incorporates strong positive correlation.

In addition to generalizing the proposed estimator to the estimation of general Gaussian integrals as in (Gollwitzer and Rackwitz 1988), another contribution of this article is the formulation of new qualitative properties of Markov chain samplers. In particular, we find that asymptotic efficiency is desirable not only in the robust estimation of rare-event probabilities, but also in the Markov chain Monte Carlo simulation from the corresponding conditional measure (conditional on the rare event).

Given that here we have only analyzed the independence-Metropolis sampler, in future investigations we would like to establish logarithmic or strong efficiency for the more sophisticated Gibbs and/or random-walk Metropolis samplers.

REFERENCES

Botev, Z. I. 2016. "The Normal Law under Linear Restrictions: Simulation and Estimation via Minimax Tilting". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. DOI: 10.1111/rssb.12162.

Botev, Z. I., and P. L'Ecuyer. 2015. "Efficient Probability Estimation and Simulation of the Truncated Multivariate Student-t Distribution". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 380–391. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Botev, Z. I., M. Mandjes, and A. Ridder. 2015. "Tail Distribution of the Maximum of Correlated Gaussian Random Variables". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 633–642. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Gollwitzer, S., and R. Rackwitz. 1988. "An Efficient Numerical Solution to the Multinormal Integral". *Probabilistic Engineering Mechanics* 3 (2): 98–101.

Hashorva, E., and J. Hüsler. 2003. "On Multivariate Gaussian Tails". *Annals of the Institute of Statistical Mathematics* 55 (3): 507–522.

Klein, J. P., and M. L. Moeschberger. 2005. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer Science & Business Media.

Kroese, D. P., T. Taimre, and Z. I. Botev. 2011. *Handbook of Monte Carlo Methods*, Volume 706. John Wiley & Sons.

Mengersen, K. L., and R. L. Tweedie. 1996. "Rates of Convergence of the Hastings and Metropolis Algorithms". *The annals of Statistics* 24 (1): 101–121.

Mills, J. P. 1926. "Table of the Ratio: Area to Bounding Ordinate, for any Portion of Normal Curve". *Biometrika*:395–400.

Savage, R. I. 1962. "Mills' Ratio for Multivariate Normal Distributions". *J. Res. Nat. Bur. Standards Sect. B* 66:93–96.

ACKNOWLEDGMENTS

Zdravko Botev has been supported by the Australian Research Council grant DE140100993. Daniel Mackinlay is supported by the University of New South Wales *University Postgraduate Award*. Yi-Lung Chen is supported by an Australian government *Domestic Research Training Program* Scholarship.

AUTHOR BIOGRAPHIES

ZDRAVKO BOTEV is Senior Lecturer at the School of Mathematics and Statistics at the University of New South Wales in Sydney, Australia. Previously he was an Australian Research Council fellow and a postdoctoral fellow at the University of Montreal. Currently, he is serving as an associate editor of the INFORMS Journal of Computing. His research interests include splitting and adaptive importance sampling methods for rare-event simulation. He is the author the most accurate nonparametric density estimation software, which has been widely used in the applied sciences. For more information, see: http://web.maths.unsw.edu.au/~zdravkobotev/ or email at botev@unsw.edu.au.

DANIEL MACKINLAY is currently pursuing a PhD degree at the University of New South Wales in Sydney, Australia. He holds a Master of Science degree from ETH Zürich. His email address is d.mackinlay@student.unsw.edu.au.

YI-LUNG CHEN is a PhD candidate at the School of Mathematics and Statistics and a tutor at the University of New South Wales in Sydney. He holds a Bachelor of Science degree with first class Honors from the University of New South Wales. His email address is yi-lung.chen@unsw.edu.au.