J. Appl. Prob. **51,** 359–376 (2014)
Printed in England

MARKOV CHAIN MONTE CARLO FOR COMPUTING RARE-EVENT PROBABILITIES FOR A HEAVY-TAILED RANDOM WALK

THORBJÖRN GUDMUNDSSON * ** AND
HENRIK HULT,* *** *KTH Royal Institute of Technology*

Abstract

In this paper a method based on a Markov chain Monte Carlo (MCMC) algorithm is proposed to compute the probability of a rare event. The conditional distribution of the underlying process given that the rare event occurs has the probability of the rare event as its normalizing constant. Using the MCMC methodology, a Markov chain is simulated, with the aforementioned conditional distribution as its invariant distribution, and information about the normalizing constant is extracted from its trajectory. The algorithm is described in full generality and applied to the problem of computing the probability that a heavy-tailed random walk exceeds a high threshold. An unbiased estimator of the reciprocal probability is constructed whose normalized variance vanishes asymptotically. The algorithm is extended to random sums and its performance is illustrated numerically and compared to existing importance sampling algorithms.

Keywords: Markov chain Monte Carlo; heavy tail; rare-event simulation; random walk

2010 Mathematics Subject Classification: Primary 65C05; 60J22
Secondary 60G50

1. Introduction

In this paper a Markov chain Monte Carlo (MCMC) methodology is proposed for computing the probability of a rare event. The basic idea is to use an MCMC algorithm to sample from the conditional distribution given that the event of interest occurs, and then extract the probability of the event as the normalizing constant. The methodology will be outlined in full generality and exemplified in the setting of computing hitting probabilities for a heavy-tailed random walk.

A rare-event simulation problem can often be formulated as follows. Consider a sequence of random elements $X^{(1)}, X^{(2)}, \ldots$, e.g. random variables, vectors, or processes, each of which can be sampled repeatedly by a simulation algorithm. The objective is to estimate $p^{(n)} = \mathbb{P}(X^{(n)} \in A_n)$ for some large n, based on a sample $X_0^{(n)}, \ldots, X_{T-1}^{(n)}$. It is assumed that the probability $\mathbb{P}(X^{(n)} \in A_n) \to 0$ as $n \to \infty$, so that the event $\{X^{(n)} \in A_n\}$ can be thought of as rare. The solution to the problem consists of finding a family of simulation algorithms and corresponding estimators whose performance is satisfactory for all n. For unbiased estimators $\hat{p}_T^{(n)}$ of $p^{(n)}$, a useful performance measure is the relative error

$$\mathrm{RE}^{(n)} = \frac{\mathrm{var}(\hat{p}_T^{(n)})}{(p^{(n)})^2}.$$

Received 16 November 2012; revision received 26 June 2013.
* Postal address: Department of Mathematics, KTH Royal Institute of Technology, SE-100 44, Stockholm, Sweden.
** Email address: tgud@kth.se
*** Email address: hult@kth.se

An algorithm is said to have *vanishing relative error* if the relative error tends to 0 as $n \to \infty$ and *bounded relative error* if the relative error is bounded in n.

It is well known that the standard Monte Carlo (MC) algorithm is inefficient for computing rare-event probabilities. As an illustration, consider the standard MC estimate

$$\hat{p}_T^{(n)} = \frac{1}{T} \sum_{t=0}^{T-1} 1\{X_t^{(n)} \in A_n\}$$

of $p^{(n)} = \mathbb{P}(X^{(n)} \in A_n)$ based on the independent replicas $X_0^{(n)}, \ldots, X_{T-1}^{(n)}$. The relative error of the MC estimator is

$$\frac{\operatorname{var}(\hat{p}_T^{(n)})}{(p^{(n)})^2} = \frac{p^{(n)}(1-p^{(n)})}{T(p^{(n)})^2} = \frac{1}{Tp^{(n)}} - \frac{1}{T} \to \infty$$

as $n \to \infty$, indicating that the performance deteriorates when the event is rare.

A popular method to reduce the computational cost is importance sampling; see, e.g. [3]. In importance sampling the random variables $X_0^{(n)}, \ldots, X_{T-1}^{(n)}$ are sampled independently from a different distribution, say $G^{(n)}$, instead of the original distribution $F^{(n)}$. The importance sampling estimator is defined as the weighted empirical estimator

$$\hat{p}_T^{(n)} = \frac{1}{T} \sum_{t=0}^{T-1} L^{(n)}(X_t^{(n)}) 1\{X_t^{(n)} \in A_n\},$$

where $L^{(n)} = \mathrm{d}F^{(n)}/\mathrm{d}G^{(n)}$ is the likelihood ratio, which is assumed to exist on A. The importance sampling estimator $\hat{p}_T^{(n)}$ is unbiased and its performance depends on the choice of the sampling distribution $G^{(n)}$. The optimal sampling distribution is called the zero-variance distribution and is simply the conditional distribution

$$F_A^{(n)}(\cdot) = \mathbb{P}(X^{(n)} \in \cdot \mid X^{(n)} \in A_n) = \frac{\mathbb{P}(X^{(n)} \in \cdot \cap A_n)}{p^{(n)}}.$$

In this case the likelihood ratio weights $L^{(n)}$ are equal to $p^{(n)}$, which implies that $\hat{p}_T^{(n)}$ has zero variance. Clearly, the zero-variance distribution cannot be implemented in practice because $p^{(n)}$ is unknown, but it serves as a starting point for selecting the sampling distribution. A good idea is to choose a sampling distribution $G^{(n)}$ that approximates the zero-variance distribution and such that the random variable $X^{(n)}$ can easily be sampled from $G^{(n)}$; the event $\{X^{(n)} \in A_n\}$ is more likely under the sampling distribution $G^{(n)}$ than under the original $F^{(n)}$, and the likelihood ratio $L^{(n)}$ is unlikely to become too large. Proving efficiency (e.g. bounded relative error) of an importance sampling algorithm can be technically cumbersome and often requires extensive analysis.

The methodology proposed in this paper is also based on the conditional distribution $F_{A_n}^{(n)}$. Because $F_{A_n}^{(n)}$ is known up to the normalizing constant $p^{(n)}$, it is possible to sample from $F_{A_n}^{(n)}$ using an MCMC algorithm, such as a Gibbs sampler or Metropolis–Hastings algorithm. The idea is to generate samples $X_0^{(n)}, \ldots, X_{T-1}^{(n)}$ from a Markov chain with invariant distribution $F_{A_n}^{(n)}$ and construct an estimator of the normalizing constant $p^{(n)}$. An unbiased estimator of $q^{(n)} = (p^{(n)})^{-1}$ is constructed from a known probability distribution $V^{(n)}$ on A_n and the

https://doi.org/10.1239/jap/1402578630 Published online by Cambridge University Press

original distribution $F^{(n)}$ of $X^{(n)}$ by

$$\hat{q}_T^{(n)} = \frac{1}{T}\sum_{t=0}^{T-1} \frac{\mathrm{d}V^{(n)}}{\mathrm{d}F^{(n)}}(X_t^{(n)})1\{X_t^{(n)} \in A_n\}. \tag{1.1}$$

The performance of the estimator depends both on the choice of the distribution $V^{(n)}$ and on the ergodic properties of the MCMC sampler used in the implementation. Roughly speaking, the rare-event properties, as $n \to \infty$, are controlled by the choice of $V^{(n)}$ and the large sample properties, as $T \to \infty$, are controlled by the ergodic properties of the MCMC sampler.

The computation of normalizing constants and ratios of normalizing constants in the context of MCMC is a reasonably well-studied problem in the statistical literature; see, e.g. [11] and the references therein. However, such methods have, to the best of our knowledge, not been studied in the context of rare-event simulation.

To exemplify MCMC methodology, we consider the problem of computing the probability $p^{(n)} = \mathbb{P}(S_n > a_n)$ that a random walk $S_n = Y_1 + \cdots + Y_n$ (where $Y_1, \ldots, Y_n$ are nonnegative, independent, and heavy-tailed random variables) exceeds a high threshold a_n as the number of summands n increases. This problem has received some attention in the context of conditional MC algorithms (see [2] and [4]) and importance sampling algorithms (see [5], [6], [10], and [15]), most notably in the setting where the number of summands is fixed.

In this paper a Gibbs sampler is presented for sampling from the conditional distribution $\mathbb{P}((Y_1, \ldots, Y_n) \in \cdot \mid S_n > a_n)$. The resulting Markov chain is proved to be uniformly ergodic. An estimator for $(p^{(n)})^{-1}$ of the form (1.1) is suggested with $V^{(n)}$ as the conditional distribution of $(Y_1, \ldots, Y_n)$ given $\max\{Y_1, \ldots, Y_n\} > a_n$. The estimator is proved to have vanishing normalized variance when the distribution of Y_1 belongs to the class of subexponential distributions. The proof is elementary and is completed in a few lines. This is in sharp contrast to efficiency proofs for importance sampling algorithms for the same problem, which require more restrictive assumptions on the tail of Y_1 and tend to be long and technical; see [5], [6], and [10]. An extension of the algorithm to a sum with a random number of steps is also presented.

The paper is organized as follows. The basic methodology for computing rare-event probabilities is described in Section 2. Section 3 contains the design and efficiency results for the estimator for computing hitting probabilities in the case of a heavy-tailed random walk with, firstly, a deterministic number of steps, and, secondly, a random number of steps. In Section 4 we present numerical experiments and compare the efficiency of the MCMC estimator against an existing importance sampling algorithm and standard MC.

2. Rare-event simulation by MCMC

In this section an algorithm for rare-event simulation using MCMC is presented and conditions that ensure good convergence are discussed.

2.1. Formulation of the algorithm

Let X be a random element taking values in a measurable space. Denote by F the distribution of X, and let A be a measurable set. The problem is to compute the probability

$$p = \mathbb{P}(X \in A) = \int_A \mathrm{d}F.$$

https://doi.org/10.1239/jap/1402578630 Published online by Cambridge University Press

The event $\{X \in A\}$ is thought of as rare in the sense that p is small. Let F_A be the conditional distribution of X given $X \in A$. Then

$$\frac{\mathrm{d}F_A}{\mathrm{d}F}(x) = \frac{1}{p} 1\{x \in A\}.$$

Consider a Markov chain $(X_t)_{t \geq 0}$ having F_A as its invariant distribution. Such a Markov chain can be constructed by implementing an MCMC algorithm, such as a Gibbs sampler or a Metropolis–Hastings algorithm; see, e.g. [3] and [12].

To construct an estimator of $q = p^{-1}$, consider a probability distribution V with $V \ll F_A$. It follows that $V \ll F$ and it is assumed that the density $\mathrm{d}V/\mathrm{d}F$ is known. An estimator of q is given by

$$\hat{q}_T = \frac{1}{T} \sum_{t=0}^{T-1} u(X_t), \quad \text{where} \quad u(x) = \frac{1}{p} \frac{\mathrm{d}V}{\mathrm{d}F_A}(x) = \frac{\mathrm{d}V}{\mathrm{d}F}(x). \tag{2.1}$$

Note that $\hat{q}_T$ is unbiased since $V \ll F_A$ implies that $V(A) = 1$ and

$$\mathbb{E}_{F_A}[u(X)] = \int_A \frac{1}{p} \frac{\mathrm{d}V}{\mathrm{d}F_A}(x) F_A(\mathrm{d}x) = \frac{1}{p} \int_A V(\mathrm{d}x) = \frac{1}{p}.$$

The expected value above is computed under the invariant distribution F_A of the Markov chain. It is implicitly assumed that the sample size T is sufficiently large that the burn-in period, the time until the Markov chain reaches stationarity, is negligible or alternatively that the burn-in period is discarded.

It is possible to invert the estimator and think of $\hat{p}_T = (\hat{q}_T)^{-1}$ as an estimator for p, but one has to be cautious. Indeed, if the support of V is strictly contained in A then there is a nonzero probability that all the terms in the sum in (2.1) are 0, leading to the estimate $\hat{q}_T = 0$ and $\hat{p}_T = \infty$. To avoid this, we can simply take $\hat{p}_T$ as the minimum of $(\hat{q}_T)^{-1}$ and 1.

There are two essential design choices that determine the performance of the algorithm: the choice of the distribution V and the design of the MCMC sampler. The distribution V influences the variance of $u(X_t)$ in (2.1) and is therefore of main concern for controlling the rare-event properties of the algorithm. It is desirable to choose V such that the normalized variance of the estimator, given by $p^2 \operatorname{var}(\hat{q}_T)$, is not too large. On the other hand, the design of the MCMC sampler is crucial to control the dependence of the Markov chain and thereby the convergence rate of the algorithm as a function of the sample size. To speed up the simulation, it is desirable that the Markov chain mixes fast so that the dependence dies out quickly.

2.2. Controlling the normalized variance

In this subsection we provide an informal discussion of how to control the performance of the estimator $\hat{q}_T$ by controlling its normalized variance.

For the estimator $\hat{q}_T$ to be useful, it is of course important that its variance is not too large. When the probability p is small, it is desirable that $\operatorname{var}(\hat{q}_T)$ is of size comparable to $q^2 = p^{-2}$. To this end, the normalized variance $p^2 \operatorname{var}(\hat{q}_T)$ is studied.

The normalized variance can be decomposed as

$$\begin{aligned} p^2 \operatorname{var}_{F_A}(\hat{q}_T) &= p^2 \operatorname{var}_{F_A}\left(\frac{1}{T} \sum_{t=0}^{T-1} u(X_t)\right) \\ &= p^2 \left(\frac{1}{T} \operatorname{var}_{F_A}(u(X_0)) + \frac{2}{T^2} \sum_{t=0}^{T-1} \sum_{s=t+1}^{T-1} \operatorname{cov}_{F_A}(u(X_s), u(X_t))\right). \end{aligned} \tag{2.2}$$

https://doi.org/10.1239/jap/1402578630 Published online by Cambridge University Press

Let us for the moment focus our attention on the first term. It can be written as

$$\begin{aligned}\frac{p^2}{T}\operatorname{var}_{F_A}(u(X_0)) &= \frac{p^2}{T}(\mathbb{E}_{F_A}[u(X_0)^2] - \mathbb{E}_{F_A}[u(X_0)]^2)\\ &= \frac{p^2}{T}\left(\int_A \left(\frac{1}{p}\frac{\mathrm{d}V}{\mathrm{d}F_A}(x)\right)^2 F_A(\mathrm{d}x) - \frac{1}{p^2}\right)\\ &= \frac{1}{T}\left(\int_A \frac{\mathrm{d}V}{\mathrm{d}F_A}(x)V(\mathrm{d}x) - 1\right).\end{aligned}$$

Therefore, in order to control the normalized variance, the distribution V must be chosen so that $\mathbb{E}_V[\mathrm{d}V/\mathrm{d}F_A]$ is close to 1. It is clear that $V = F_A$ is the optimal choice. This motivates taking V as an approximation of F_A. The method is similar to that of choosing an efficient sampling distribution in importance sampling. In that case F_A is called the zero-variance distribution. An important difference is that here $V \ll F_A \ll F$ is required, whereas in importance sampling F needs to be absolutely continuous with respect to the sampling distribution.

If, for some set $B \subset A$, the probability $\mathbb{P}(X \in B)$ can be computed explicitly then a candidate for V is

$$V(\cdot) = \mathbb{P}(X \in \cdot \mid X \in B).$$

This candidate is likely to perform well if $\mathbb{P}(X \in B)$ is a good approximation of p. Indeed, in this case

$$\int_A \frac{\mathrm{d}V}{\mathrm{d}F_A}(x)V(\mathrm{d}x) = \frac{\mathbb{P}(X \in A)}{\mathbb{P}(X \in B)},$$

which will be close to 1.

Now, let us consider the covariance term in (2.2). Since the samples $(X_t)_{t=0}^{T-1}$ form a Markov chain, the X_t are dependent. Therefore, the covariance term in (2.2) is nonzero and may not be ignored. The crude upper bound

$$\operatorname{cov}_{F_A}(u(X_s), u(X_t)) \le \operatorname{var}_{F_A}(u(X_0))$$

leads to the upper bound

$$\frac{2p^2}{T^2}\sum_{t=0}^{T-1}\sum_{s=t+1}^{T-1}\operatorname{cov}_{F_A}(u(X_s), u(X_t)) \le p^2\left(1 - \frac{1}{T}\right)\operatorname{var}_{F_A}(u(X_0))$$

for the covariance term. This is a very crude upper bound as it does not decay to 0 as $T \to \infty$. However, at the moment, the emphasis is on small p and so we will proceed with this upper bound anyway. As indicated above, the choice of V controls the term $p^2 \operatorname{var}_{F_A}(u(X_0))$ and, therefore, controls the normalized variance.

2.3. Ergodic properties

As we have just seen, the choice of the distribution V controls the normalized variance of the estimator for small p. The design of the MCMC sampler, on the other hand, determines the strength of the dependence in the Markov chain. Strong dependence implies slow convergence, which results in a high computational cost. The convergence rate of MCMC samplers can be analyzed within the theory of φ-irreducible Markov chains. Fundamental results, for φ-irreducible Markov chains, are given in [18] and [19]. We will focus on conditions that imply a geometric convergence rate. The conditions given below are well studied in the context

https://doi.org/10.1239/jap/1402578630 Published online by Cambridge University Press

of MCMC samplers. Conditions for geometric ergodicity in the context of Gibbs samplers are considered in, e.g. [7], [21], and [22], and, for Metropolis–Hastings algorithms, in [17].

A Markov chain $(X_t)_{t\geq 0}$ with transition kernel $p(x,\cdot) = \mathbb{P}(X_{t+1} \in \cdot \mid X_t = x)$ is φ-irreducible if there exists a measure φ such that $\sum_t p^{(t)}(x,\cdot) \ll \varphi(\cdot)$, where $p^{(t)}(x,\cdot) = \mathbb{P}(X_t \in \cdot \mid X_0 = x)$ denotes the t-step transition kernel. A Markov chain with invariant distribution π is called geometrically ergodic if there exists a positive function M and a constant $r \in (0,1)$ such that

$$\|p^{(t)}(x,\cdot) - \pi(\cdot)\|_{\mathrm{TV}} \leq M(x) r^t, \tag{2.3}$$

where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation norm. This condition ensures that the distribution of the Markov chain converges at a geometric rate to the invariant distribution. If the function M is bounded then the Markov chain is said to be uniformly ergodic. Conditions such as (2.3) may be difficult to establish directly and are therefore substituted by suitable minorization or drift conditions. A minorization condition holds on a set C if there exist a probability measure ν, a positive integer t_0, and $\delta > 0$ such that

$$p^{(t_0)}(x, B) \geq \delta \nu(B)$$

for all $x \in C$ and Borel sets B. In this case C is said to be a small set. Minorization conditions have been used to obtain rigorous bounds on the convergence of MCMC samplers; see, e.g. [20].

If the entire state space is small then the Markov chain is uniformly ergodic. Typically, uniform ergodicity does not hold for Metropolis samplers; see [17, Theorem 3.1]. Hence, useful sufficient conditions for geometric ergodicity are often given in the form of drift conditions; see [7] and [17]. Drift conditions, established through the construction of appropriate Lyapunov functions, are also useful for establishing central limit theorems for MCMC algorithms; see [14], [18], and the references therein. When studying simulation algorithms for random walks, in Section 3, we will encounter Gibbs samplers that are uniformly ergodic.

2.4. Asymptotic efficiency

Asymptotic efficiency can be conveniently formulated in terms of a limit as a large deviation parameter tends to ∞. As is usual in problems related to rare-event simulation, the problem at hand is embedded in a sequence of problems, indexed by $n = 1, 2, \ldots$. The general setup is formalized as follows.

Let $(X^{(n)})_{n\geq 1}$ be a sequence of random elements with $X^{(n)}$ having distribution $F^{(n)}$, and let A_n be the sets of interest. Suppose that

$$p^{(n)} = \mathbb{P}(X^{(n)} \in A_n) \to 0$$

as $n \to \infty$. For the nth problem, a Markov chain $(X_t^{(n)})_{t=0}^{T-1}$ with invariant distribution $F_{A_n}^{(n)}(\cdot) = \mathbb{P}(X^{(n)} \in \cdot \mid X^{(n)} \in A_n)$ is generated by an MCMC algorithm. The estimator of $q^{(n)} = (p^{(n)})^{-1}$ (which is based on a probability distribution $V^{(n)}$ that has known density with respect to $F^{(n)}$ and satisfies $V^{(n)} \ll F_{A_n}^{(n)}$) is given by

$$\hat{q}_T^{(n)} = \frac{1}{T}\sum_{t=0}^{T-1} u^{(n)}(X_t^{(n)}), \quad \text{where} \quad u^{(n)}(x) = \frac{\mathrm{d}V^{(n)}}{\mathrm{d}F^{(n)}}(x).$$

In the rest of the paper we will be concerned with the following notions of efficiency.

Rare-event efficiency. Select the probability distributions $V^{(n)}$ such that

$$(p^{(n)})^2 \operatorname{var}_{F_{A_n}^{(n)}}(u^{(n)}(X)) \to 0 \quad \text{as } n \to \infty.$$

https://doi.org/10.1239/jap/1402578630 Published online by Cambridge University Press

Large sample size efficiency. Design the MCMC sampler, by finding an appropriate Gibbs sampler or a proposal density for the Metropolis–Hastings algorithm, such that, for each $n \geq 1$, the Markov chain $(X_t^{(n)})_{t\geq 0}$ is geometrically ergodic.

Remark 2.1. The rare-event efficiency criteria is formulated in terms of the efficiency of estimating $(p^{(n)})^{-1}$ by $\hat{q}_T^{(n)}$. If we insist on studying the mean and variance of $\hat{p}_T^{(n)} = (\hat{q}_T^{(n)})^{-1}$ then the effects of the transformation $x \mapsto x^{-1}$ must be taken into account. For instance, the estimator $\hat{p}_T^{(n)}$ is biased and its variance could be infinite. The bias can be reduced, for instance, via the delta method illustrated in [3, p. 76]. We also remark that even in the estimation of $(p^{(n)})^{-1}$ by $\hat{q}_T^{(n)}$ there is a bias coming from the fact that the Markov chain is not perfectly stationary.

Remark 2.2. The proposed methodology can obviously be generalized to estimate the expectations $\theta^{(n)} = \mathbb{E}[h_n(X^{(n)})]$ for integrable functions h_n, in which case the conditional distribution $F_A^{(n)}$ must be replaced by $F_{h_n}^{(n)}$, given by $\mathrm{d}F_{h_n}^{(n)}/\mathrm{d}F^{(n)} = (\theta^{(n)})^{-1}h_n$.

3. A random walk with heavy-tailed steps

In this section the estimator introduced in Section 2 is applied to compute the probability that a random walk with heavy-tailed steps exceeds a high threshold.

Let $Y_1, \ldots, Y_n$ be nonnegative, independent, and identically distributed random variables with common distribution F_Y. Consider the random walk $S_n = Y_1 + \cdots + Y_n$ and the problem of computing the probability

$$p^{(n)} = \mathbb{P}(S_n > a_n),$$

where $a_n \to \infty$ sufficiently fast that $p^{(n)} \to 0$ as $n \to \infty$.

It is convenient to denote by $\boldsymbol{Y}^{(n)}$ the n-dimensional random vector $(Y_1, \ldots, Y_n)^\top$ and by A_n the set $\{\boldsymbol{y} \in \mathbb{R}^n : \mathbf{1}^\top \boldsymbol{y} > a_n\}$, where $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^n$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$. With this notation,

$$p^{(n)} = \mathbb{P}(S_n > a_n) = \mathbb{P}(\mathbf{1}^\top \boldsymbol{Y}^{(n)} > a_n) = \mathbb{P}(\boldsymbol{Y}^{(n)} \in A_n),$$

and the conditional distribution given the event is $F_{A_n}^{(n)}(\cdot) = \mathbb{P}(\boldsymbol{Y}^{(n)} \in \cdot \mid \boldsymbol{Y}^{(n)} \in A_n)$.

The first step towards defining the estimator of $q^{(n)} = (p^{(n)})^{-1}$ is to construct the Markov chain $(\boldsymbol{Y}_t^{(n)})_{t\geq 0}$ whose invariant distribution is $F_{A_n}^{(n)}$, using a Gibbs sampler. In short, the Gibbs sampler updates one element of $\boldsymbol{Y}_t^{(n)}$ at a time, keeping the other elements constant. Formally, the algorithm proceeds as follows.

Algorithm 3.1. *Start at an initial state* $\boldsymbol{Y}_0^{(n)} = (Y_{0,1}, \ldots, Y_{0,n})^\top$, *where* $Y_{0,1} + \cdots + Y_{0,n} > a_n$. *Given* $\boldsymbol{Y}_t^{(n)} = (Y_{t,1}, \ldots, Y_{t,n})^\top$ *for some* $t = 0, 1, \ldots$, *the next state* $\boldsymbol{Y}_{t+1}^{(n)}$ *is sampled as follows.*

1. *Draw* $j_1, \ldots, j_n$ *from* $\{1, \ldots, n\}$ *without replacement and proceed by updating the components of* $\boldsymbol{Y}_t^{(n)}$ *in the order thus obtained.*

2. *For each* $k = 1, \ldots, n$, *repeat the following.*

   (a) *Let* $j = j_k$ *be the index to be updated, and write*

$$\boldsymbol{Y}_{t,-j} = (Y_{t,1}, \ldots, Y_{t,j-1}, Y_{t,j+1}, \ldots, Y_{t,n})^\top.$$

https://doi.org/10.1239/jap/1402578630 Published online by Cambridge University Press

Sample $Y'_{t,j}$ from the conditional distribution of Y given that the sum exceeds the threshold. That is,

$$\mathbb{P}(Y'_{t,j} \in \cdot \mid \boldsymbol{Y}_{t,-j}) = \mathbb{P}\left(Y \in \cdot \;\middle|\; Y + \sum_{k \neq j} Y_{t,k} > a_n\right).$$

(b) *Put $\boldsymbol{Y}'_t = (Y_{t,1}, \ldots, Y_{t,j-1}, Y'_{t,j}, Y_{t,j+1}, \ldots, Y_{t,n})^\top$ and, for the next k, apply (a) to $\boldsymbol{Y}'_t$.*

3. *Draw π at random (uniformly) from the set of permutations of the numbers $\{1, \ldots, n\}$, and put $\boldsymbol{Y}^{(n)}_{t+1} = (Y'_{t,\pi(1)}, \ldots, Y'_{t,\pi(n)})^\top$.*

Iterate steps 1–3 until the entire Markov chain $(\boldsymbol{Y}^{(n)}_t)_{t=0}^{T-1}$ is constructed.

Remark 3.1. (i) In the heavy-tailed setting the trajectories of the random walk leading to the rare event are likely to consist of one large increment (a big jump) while the other increments are average. The purpose of the permutation step is to force the Markov chain to mix faster by moving the big jump to different locations. However, the permutation step in Algorithm 3.1 is not really needed when considering the probability $\mathbb{P}(S_n > a_n)$. This is due to the fact that the summation is invariant of the ordering of the steps.

(ii) The algorithm requires sampling from the conditional distribution $\mathbb{P}(Y \in \cdot \mid Y > c)$ for arbitrary c. This is easy whenever inversion is feasible, see [3, p. 39], or acceptance/rejection sampling can be employed. There are, however, situations where sampling from the conditional distribution $\mathbb{P}(Y \in \cdot \mid Y > c)$ may be difficult; see [13, Section 2.2].

The following proposition confirms that the Markov chain $(\boldsymbol{Y}^{(n)}_t)_{t\geq 0}$, generated by Algorithm 3.1, has $F^{(n)}_{A_n}$ as its invariant distribution.

Proposition 3.1. *The Markov chain $(\boldsymbol{Y}^{(n)}_t)_{t\geq 0}$, generated by Algorithm 3.1, has the conditional distribution $F^{(n)}_{A_n}$ as its invariant distribution.*

Proof. The goal is to show that each updating step (steps 2 and 3) of the algorithm preserves stationarity. Since the conditional distribution $F^{(n)}_{A_n}$ is permutation invariant, it is clear that step 3 preserves stationarity. Therefore, it is sufficient to consider step 2 of the algorithm.

Let $P_j(\boldsymbol{y}, \cdot)$ denote the transition probability of the Markov chain $(\boldsymbol{Y}^{(n)}_t)_{t\geq 0}$ corresponding to the jth component being updated. It is sufficient to show that, for all $j = 1, \ldots, n$ and all Borel sets in the form of a product $B_1 \times \cdots \times B_n \subset A_n$, the following equality holds:

$$F^{(n)}_{A_n}(B_1 \times \cdots \times B_n) = \mathbb{E}_{F^{(n)}_{A_n}}[P_j(\boldsymbol{Y}, B_1 \times \cdots \times B_n)].$$

Observe that, because $B_1 \times \cdots \times B_n \subset A_n$,

$$\begin{aligned}
&F^{(n)}_{A_n}(B_1 \times \cdots \times B_n) \\
&\quad = \mathbb{E}\left[\prod_{k=1}^{n} 1\{Y_k \in B_k\} \;\middle|\; S_n > a_n\right] \\
&\quad = \mathbb{P}(S_n > a_n)^{-1}\mathbb{E}\left[1\{Y_j \in B_j\}1\{S_n > a_n\}\prod_{k\neq j} 1\{Y_k \in B_k\}\right] \\
&\quad = \mathbb{P}(S_n > a_n)^{-1}\mathbb{E}\left[\frac{\mathbb{E}[1\{Y_j \in B_j\} \mid Y_j > a_n - S_{n,-j}, \boldsymbol{Y}^{(n)}_{-j}]\prod_{k\neq j} 1\{Y_k \in B_k\}}{\mathbb{P}(Y_j > a_n - S_{n,-j} \mid \boldsymbol{Y}^{(n)}_{-j})}\right]
\end{aligned}$$

https://doi.org/10.1239/jap/1402578630 Published online by Cambridge University Press

$$= \mathbb{P}(S_n > a_n)^{-1}\mathbb{E}\left[P_j(\boldsymbol{Y}^{(n)}, B_1 \times \cdots \times B_n)\prod_{k\neq j} 1\{Y_k \in B_k\}\right]$$
$$= \mathbb{E}[P_j(\boldsymbol{Y}^{(n)}, B_1 \times \cdots \times B_n) \mid S_n > a_n]$$
$$= \mathbb{E}_{F_{A_n}^{(n)}}[P_j(\boldsymbol{Y}, B_1 \times \cdots \times B_n)],$$

with the notation $\boldsymbol{Y}_{-j}^{(n)} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_n)^\top$ and $S_{n,-j} = S_n - Y_j$. This completes the proof.

As for the ergodic properties, Algorithm 3.1 produces a Markov chain which is uniformly ergodic.

Proposition 3.2. *For each $n \geq 1$, the Markov chain $(\boldsymbol{Y}_t^{(n)})_{t\geq 0}$ is uniformly ergodic. In particular, it satisfies the following minorization condition: there exists $\delta > 0$ such that*

$$\mathbb{P}(\boldsymbol{Y}_1^{(n)} \in B \mid \boldsymbol{Y}_0^{(n)} = \boldsymbol{y}) \geq \delta F_{A_n}^{(n)}(B)$$

for all $\boldsymbol{y} \in A_n$ and all Borel sets $B \subset A_n$.

Proof. Take an arbitrary $n \geq 1$. Uniform ergodicity can be deduced from the minorization condition; see [19]. There exists a probability measure ν, a constant $\delta > 0$, and an integer t_0 such that

$$\mathbb{P}(\boldsymbol{Y}_{t_0}^{(n)} \in B \mid \boldsymbol{Y}_0^{(n)} = \boldsymbol{y}) \geq \delta\nu(B)$$

for every $\boldsymbol{y} \in A_n$ and Borel set $B \subset A_n$. Take $\boldsymbol{y} \in A_n$ and write $g^{(n)}(\cdot \mid \boldsymbol{y})$ for the Radon–Nikodym derivative of $\mathbb{P}(\boldsymbol{Y}_1^{(n)} \in \cdot \mid \boldsymbol{Y}_0^{(n)} = \boldsymbol{y})$ with respect to $F_{A_n}^{(n)}$. The goal is to show that the minorization condition holds with $t_0 = 1$, $\delta = p^{(n)}/n!$, and $\nu = F_{A_n}^{(n)}$.

For any $\boldsymbol{x} \in A_n$, there exists an ordering $j_1, \ldots, j_n$ of the numbers $\{1, \ldots, n\}$ such that

$$y_{j_1} \leq x_{j_1}, \ldots, y_{j_k} \leq x_{j_k}, y_{j_{k+1}} > x_{j_{k+1}}, \ldots y_{j_n} > x_{j_n}$$

for some $k \in \{0, \ldots, n\}$. The probability to draw this particular ordering in step 1 of the algorithm is at least $1/n!$. It follows that

$$g(\boldsymbol{x} \mid \boldsymbol{y}) \geq \frac{p}{n!}\frac{1\{x_{j_1} \geq a_n - \sum_{i\neq j_1} y_i\}}{\bar{F}_Y(a_n - \sum_{i\neq j_1} y_i)}\frac{1\{x_{j_2} \geq a_n - \sum_{i\neq j_1,j_2} y_i - x_{j_1}\}}{\bar{F}_Y(a_n - \sum_{i\neq j_1,j_2} y_i - x_{j_1})}\cdots$$
$$\times \frac{1\{x_{j_n} \geq a_n - x_{j_1} - \cdots - x_{j_{n-1}}\}}{\bar{F}_Y(a_n - x_{j_1} - \cdots - x_{j_{n-1}})}$$
$$\geq \frac{p}{n!},$$

where the last inequality follows since, by construction of the ordering $j_1, \ldots, j_n$, all the indicators are equal to 1. The proof is completed by integrating, with respect to $F_{A_n}^{(n)}$, both sides of the inequality over any Borel set $B \subset A_n$.

Remark 3.2. To keep the proof of Proposition 3.2 simple, we have not used the permutation step of the algorithm in the proof and not tried to optimize δ. By taking advantage of the permutation step, we believe that the constant δ could, with some additional effort, be increased by a factor $n!$.

Note that, so far, the distributional assumption for the steps $Y_1, \ldots, Y_n$ of the random walk have been completely general. For the rare-event properties of the estimator, the design of $V^{(n)}$ is essential and this is where the distributional assumptions become important. In this section

https://doi.org/10.1239/jap/1402578630 Published online by Cambridge University Press

a heavy-tailed random walk is considered. To be precise, assume that the variables $Y_1, \dots, Y_n$ are nonnegative and that the tail of F_Y is heavy in the sense that there is a sequence (a_n) of real numbers such that

$$\lim_{n\to\infty} \frac{\mathbb{P}(S_n > a_n)}{\mathbb{P}(M_n > a_n)} = 1, \tag{3.1}$$

where M_n denotes the maximum of $Y_1, \dots, Y_n$. The class of distributions for which (3.1) holds is large and includes the subexponential distributions. General conditions on the sequence (a_n) for which (3.1) holds are given in [9] (see also [8]). For instance, if $\bar{F}_Y$ is regularly varying at ∞ with index $\beta > 1$ then (3.1) holds with $a_n = an$ for $a > 0$.

Next consider the choice of $V^{(n)}$. As observed in Section 2, a good approximation to the conditional distribution $F^{(n)}_{A_n}$ is a candidate for $V^{(n)}$. For a heavy-tailed random walk, the 'one big jump' heuristics says that the sum is large most likely because one of the steps is large. Based on assumption (3.1), a good candidate for $V^{(n)}$ is the conditional distribution

$$V^{(n)}(\cdot) = \mathbb{P}(\boldsymbol{Y}^{(n)} \in \cdot \mid M_n > a_n).$$

Then $V^{(n)}$ has a known density with respect to $F^{(n)}(\cdot) = \mathbb{P}(\boldsymbol{Y}^{(n)} \in \cdot)$ given by

$$\frac{\mathrm{d}V^{(n)}}{\mathrm{d}F^{(n)}}(\boldsymbol{y}) = \frac{1}{\mathbb{P}(M_n > a_n)} 1\left\{\boldsymbol{y} \colon \bigvee_{j=1}^n y_j > a_n\right\} = \frac{1}{1 - F_Y(a_n)^n} 1\left\{\boldsymbol{y} \colon \bigvee_{j=1}^n y_j > a_n\right\}.$$

The estimator of $q^{(n)} = \mathbb{P}(S_n > a_n)^{-1}$ is given by

$$\hat{q}^{(n)}_T = \frac{1}{T} \sum_{t=0}^{T-1} \frac{\mathrm{d}V^{(n)}}{\mathrm{d}F^{(n)}}(\boldsymbol{Y}^{(n)}_t) = \frac{1}{1 - F_Y(a_n)^n} \frac{1}{T} \sum_{t=0}^{T-1} 1\left\{\bigvee_{j=1}^n Y_{t,j} > a_n\right\}, \tag{3.2}$$

where $(\boldsymbol{Y}^{(n)}_t)_{t\geq 0}$ is generated by Algorithm 3.1. Note that estimator (3.2) can be viewed as the asymptotic approximation $(1 - F_Y(a_n)^n)^{-1}$ of $(p^{(n)})^{-1}$ multiplied by the random correction factor $T^{-1} \sum_{t=0}^{T-1} 1\{\bigvee_{j=1}^n Y_{t,j} > a_n\}$. The efficiency of this estimator relies on the fact that the random correction factor is likely to be close to 1 and has small variance.

Theorem 3.1. *Suppose that (3.1) holds. Then the estimator $\hat{q}^{(n)}_T$ in (3.2) has vanishing normalized variance for estimating $(p^{(n)})^{-1}$. That is,*

$$\lim_{n\to\infty} (p^{(n)})^2 \operatorname{var}_{F^{(n)}_{A_n}}(\hat{q}^{(n)}_T) = 0.$$

Proof. With $u^{(n)}(\boldsymbol{y}) = (1 - F_Y(a_n)^n)^{-1} 1\{\bigvee_{j=1}^n y_j > a_n\}$, it follows from (3.1) that

$$\begin{aligned}
&(p^{(n)})^2 \operatorname{var}_{F^{(n)}_{A_n}}(u^{(n)}(\boldsymbol{Y}^{(n)})) \\
&\quad = \frac{\mathbb{P}(S_n > a_n)^2}{\mathbb{P}(M_n > a_n)^2} \operatorname{var}_{F^{(n)}_{A_n}}\left(1\left\{\boldsymbol{Y} \colon \bigvee_{j=1}^n Y_j > a_n\right\}\right) \\
&\quad = \frac{\mathbb{P}(S_n > a_n)^2}{\mathbb{P}(M_n > a_n)^2} \mathbb{P}(M_n > a_n \mid S_n > a_n)\mathbb{P}(M_n \leq a_n \mid S_n > a_n) \\
&\quad = \frac{\mathbb{P}(S_n > a_n)}{\mathbb{P}(M_n > a_n)}\left(1 - \frac{\mathbb{P}(M_n > a_n)}{\mathbb{P}(S_n > a_n)}\right) \\
&\quad \to 0.
\end{aligned}$$

This completes the proof.

https://doi.org/10.1239/jap/1402578630 Published online by Cambridge University Press

Remark 3.3. Theorem 3.1 covers a wide range of heavy-tailed distributions and even allows the number of steps to increase with n. Its proof is elementary. This is in sharp contrast to the existing proofs of efficiency (bounded relative error, say) for importance sampling algorithms that cover less general models and tend to be long and technical; see, e.g. [5], [6], and [10]. It must be mentioned though that Theorem 3.1 proves efficiency for computing $(p^{(n)})^{-1}$, whereas the authors of [5], [6], and [10] proved efficiency for a direct computation of $p^{(n)}$.

3.1. An extension to random sums

In application to queueing and ruin theory, there is particular interest in sums consisting of a random number of heavy-tailed steps. For instance, the stationary distribution of the waiting time and the workload of an M/G/1 queue can be represented as a random sum; see [1, Theorem 5.7, p. 237]. The classical Cramér–Lundberg model for the total claim amount faced by an insurance company is another standard example of a random sum. In this section Algorithm 3.1 is modified to efficiently estimate hitting probabilities for heavy-tailed random sums.

Let $Y_1, Y_2, \ldots$ be nonnegative independent random variables with common distribution F_Y. Let $(N^{(n)})_{n\geq 1}$ be integer-valued random variables independent of $Y_1, Y_2, \ldots$. Consider the random sum $S_{N^{(n)}} = Y_1 + \cdots + Y_{N^{(n)}}$ and the problem of computing the probability

$$p^{(n)} = \mathbb{P}(S_{N^{(n)}} > a_n),$$

where $a_n \to \infty$ at an appropriate rate.

Denote by $\bar{\boldsymbol{Y}}^{(n)}$ the vector $(N^{(n)}, Y_1, \ldots, Y_{N^{(n)}})^\top$. The conditional distribution of $\bar{\boldsymbol{Y}}^{(n)}$, given $S_{N^{(n)}} > a_n$, is given by

$$\mathbb{P}(N^{(n)} = k,\ (Y_1, \ldots, Y_k) \in \cdot \mid S_{N^{(n)}} > a_n) = \frac{\mathbb{P}((Y_1, \ldots, Y_k) \in \cdot,\ S_k > a_n)\mathbb{P}(N^{(n)} = k)}{p^{(n)}}.$$

A Gibbs sampler for sampling from the above conditional distribution can be constructed essentially as in Algorithm 3.1. The only additional difficulty is to update the random number of steps in an appropriate way. In the following algorithm a particular distribution for updating the number of steps is proposed that gives the correct invariant distribution. To ease the notation, the superscript n is suppressed in the description of the algorithm and in the analysis of its ergodic properties.

Algorithm 3.2. *To initiate, sample N_0 from $\mathbb{P}(N \in \cdot)$ and $Y_{0,1}, \ldots, Y_{0,N_0}$ such that $Y_{0,1} + \cdots + Y_{0,N_0} > a_n$. Each iteration of the algorithm consists of the following steps. Suppose that $\bar{\boldsymbol{Y}}_t = (k_t, y_{t,1}, \ldots, y_{t,k_t})^\top$ with $y_{t,1} + \cdots + y_{t,k_t} > a_n$. Write $k_t^* := \min\{j : y_{t,1} + \cdots + y_{t,j} > a_n\}$.*

1. *Sample the number of steps N_{t+1} from the distribution*

$$p(k_{t+1} \mid k_t^*) = \frac{\mathbb{P}(N = k_{t+1})1\{k_{t+1} \geq k_t^*\}}{\mathbb{P}(N \geq k_t^*)}.$$

   If $N_{t+1} > N_t$, sample $Y_{t+1,k_t+1}, \ldots, Y_{t+1,N_{t+1}}$ independently from F_Y and put $\boldsymbol{Y}_t^{(1)} = (Y_{t,1}, \ldots, Y_{t,k_t}, Y_{t+1,k_t+1}, \ldots, Y_{t+1,N_{t+1}})^\top$.

2. *Proceed by updating all the individual steps as in Algorithm 3.1.*

https://doi.org/10.1239/jap/1402578630 Published online by Cambridge University Press

(a) *Draw* $j_1, \dots, j_{N_{t+1}}$ *from* $\{1, \dots, N_{t+1}\}$ *without replacement and proceed by updating the components of* $\boldsymbol{Y}_t^{(1)}$ *in the order thus obtained.*

(b) *For each* $k = 1, \dots, N_{t+1}$, *repeat the following.*

(i) *Let* $j = j_k$ *be the index to be updated, and write*

$$\boldsymbol{Y}_{t,-j}^{(1)} = (Y_{t,1}^{(1)}, \dots, Y_{t,j-1}^{(1)}, Y_{t,j+1}^{(1)}, \dots, Y_{t,N_{t+1}}^{(1)})^\top.$$

Sample $Y_{t,j}^{(2)}$ *from the conditional distribution of* Y *given that the sum exceeds the threshold. That is,*

$$\mathbb{P}(Y_{t,j}^{(2)} \in \cdot \mid \boldsymbol{Y}_{t,-j}^{(1)}) = \mathbb{P}\bigg(Y \in \cdot \;\bigg|\; Y + \sum_{k \neq j} Y_{t,k}^{(1)} > a_n\bigg).$$

(ii) *Put* $\boldsymbol{Y}_t^{(2)} = (Y_{t,1}^{(1)}, \dots, Y_{t,j-1}^{(1)}, Y_{t,j}^{(2)}, Y_{t,j+1}^{(1)}, \dots, Y_{t,N_{t+1}}^{(1)})^\top$ *and, for the next* k, *apply (i) to* $\boldsymbol{Y}_t^{(2)}$.

(c) *Draw* π *at random (uniformly) from the set of permutations of the numbers* $\{1, \dots, N_{t+1}\}$, *and put* $\bar{\boldsymbol{Y}}_{t+1} = (N_{t+1}, Y_{t,\pi(1)}^{(2)}, \dots, Y_{t,\pi(N_{t+1})}^{(2)})^\top$.

Iterate until the entire Markov chain $(\bar{\boldsymbol{Y}}_t)_{t=0}^{T-1}$ *is constructed.*

Proposition 3.3. *The Markov chain* $(\bar{\boldsymbol{Y}}_t)_{t \geq 0}$, *generated by Algorithm 3.2, has the conditional distribution* $\mathbb{P}((N, Y_1, \dots, Y_N) \in \cdot \mid Y_1 + \cdots + Y_N > a_n)$ *as its invariant distribution.*

Proof. The only essential difference from Algorithm 3.1 is the first step of the algorithm, where the number of steps and possibly the additional steps are updated. Therefore, by Proposition 3.1, it is sufficient to prove that the first step of the algorithm preserves stationarity. The transition probability of the first step, starting from a state $(k_t, y_{t,1}, \dots, y_{t,k_t})$ with $k_t^* = \min\{j : y_{t,1} + \cdots + y_{t,j} > a_n\}$, can be written as

$$\begin{aligned} &P^{(1)}(k_t, y_{t,1}, \dots, y_{t,k_t}; k_{t+1}, A_1 \times \cdots \times A_{k_{t+1}}) \\ &\quad = \mathbb{P}(N_{t+1} = k_{t+1},\ (Y_{t,1}, \dots, Y_{t,k_{t+1}}) \in A_1 \times \cdots \times A_{k_{t+1}} \mid N_t = k_t, \\ &\qquad Y_{t,1} = y_{t,1}, \dots, Y_{t,k_t} = y_{t,k_t}) \\ &\quad = \begin{cases} p(k_{t+1} \mid k_t^*) \displaystyle\prod_{k=1}^{k_{t+1}} 1\{y_{t,k} \in A_k\}, & k_{t+1} \leq k_t, \\ p(k_{t+1} \mid k_t^*) \displaystyle\prod_{k=1}^{k_t} 1\{y_{t,k} \in A_k\} \prod_{k=k_t+1}^{k_{t+1}} F_Y(A_k), & k_{t+1} > k_t. \end{cases} \end{aligned}$$

Consider the stationary probability of a set of the form $\{k_{t+1}\} \times A_1 \times \cdots \times A_{k_{t+1}}$. With π denoting the conditional distribution $\mathbb{P}((N, Y_1, \dots, Y_N) \in \cdot \mid Y_1 + \cdots + Y_N > a_n)$, we have

$$\begin{aligned} &\mathbb{E}_\pi[P^{(1)}(N_t, Y_{t,1}, \dots, Y_{t,N_t}; k_{t+1}, A_1 \times \cdots \times A_{k_{t+1}})] \\ &\quad = \frac{1}{\mathbb{P}(S_N > a_n)} \mathbb{E}[P^{(1)}(N, Y_1, \dots, Y_N; k_{t+1}, A_1 \times \cdots \times A_{k_{t+1}}) 1\{S_N > a_n\}]. \end{aligned} \tag{3.3}$$

https://doi.org/10.1239/jap/1402578630 Published online by Cambridge University Press

By conditioning on N, and using the independence of N and $Y_1, Y_2, \ldots$, the right-hand side of (3.3) equals

$$\frac{1}{\mathbb{P}(S_N > a_n)} \sum_{k_t=1}^{\infty} \mathbb{P}(N = k_t)\mathbb{E}[P^{(1)}(k_t, Y_1, \ldots, Y_{k_t}; k_{t+1}, A_1 \times \cdots \times A_{k_{t+1}})1\{S_{k_t} > a_n\}]. \tag{3.4}$$

With $B_{k^*} = \{(y_1, y_2, \ldots) \in \bigcup_{q=k^*}^{\infty} \mathbb{R}^q : \min\{j : y_1 + \cdots + y_j > a\} = k^*\}$, $A_{k_t}^{\otimes} = A_1 \times \cdots \times A_{k_t}$, and $A_{k_{t+1}}^{\otimes} = A_1 \times \cdots \times A_{k_{t+1}}$, (3.4) can be written as

$$\begin{aligned}
&\frac{1}{\mathbb{P}(S_N > a_n)} \\
&\times \Bigg(\sum_{k_t=1}^{k_{t+1}} \mathbb{P}(N = k_t)\mathbb{E}\Bigg[\sum_{k^*=1}^{k_t} 1\{(Y_1, \ldots, Y_{k_t}) \in B_{k^*}\} P^{(1)}(k_t, Y_1, \ldots, Y_{k_t}; k_{t+1}, A_{k_{t+1}}^{\otimes}) \Bigg] \\
&\quad + \sum_{k_t=k_{t+1}+1}^{\infty} \mathbb{P}(N = k_t)\mathbb{E}\Bigg[\sum_{k^*=1}^{k_{t+1}} 1\{(Y_1, \ldots, Y_{k_{t+1}}) \in B_{k^*}\} P^{(1)}(k_t, Y_1, \ldots, Y_{k_t}; k_{t+1}, A_{k_{t+1}}^{\otimes}) \Bigg] \Bigg).
\end{aligned}$$

Inserting the expression for $P^{(1)}$, this expression becomes

$$\begin{aligned}
&\frac{1}{\mathbb{P}(S_N > a)} \Bigg(\sum_{k_t=1}^{k_{t+1}} \mathbb{P}(N = k_t) \sum_{k^*=1}^{k_t} \mathbb{P}((Y_1, \ldots, Y_{k_t}) \in B_{k^*} \cap A_{k_t}^{\otimes}) p(k_{t+1} \mid k^*) \prod_{j=k_t+1}^{k_{t+1}} F_Y(A_j) \\
&\qquad + \sum_{k_t=k_{t+1}+1}^{\infty} \mathbb{P}(N = k_t) \sum_{k^*=1}^{k_{t+1}} \mathbb{P}((Y_1, \ldots, Y_{k_{t+1}}) \in B_{k^*} \cap A_{k_{t+1}}^{\otimes}) p(k_{t+1} \mid k^*) \Bigg).
\end{aligned}$$

Changing the order of the summation, this expression becomes

$$\begin{aligned}
&\frac{1}{\mathbb{P}(S_N > a_n)} \Bigg(\sum_{k^*=1}^{k_{t+1}} \sum_{k_t=k^*}^{k_{t+1}} \mathbb{P}(N = k_t)\mathbb{P}((Y_1, \ldots, Y_{k_t}) \in B_{k^*} \cap A_{k_t}^{\otimes}) p(k_{t+1} \mid k^*) \prod_{j=k_t+1}^{k_{t+1}} F_Y(A_j) \\
&\qquad + \sum_{k^*=1}^{k_{t+1}} \sum_{k_t=k_{t+1}+1}^{\infty} \mathbb{P}(N = k_t)\mathbb{P}((Y_1, \ldots, Y_{k_{t+1}}) \in B_{k^*} \cap A_{k_{t+1}}^{\otimes}) p(k_{t+1} \mid k^*) \Bigg).
\end{aligned}$$

Since $\mathbb{P}((Y_1, \ldots, Y_{k_t}) \in B_{k^*} \cap A_{k_t}^{\otimes}) \prod_{j=k_t+1}^{k_{t+1}} F_Y(A_j) = \mathbb{P}((Y_1, \ldots, Y_{k_{t+1}}) \in B_{k^*} \cap A_{k_{t+1}}^{\otimes})$, the last expression equals

$$\begin{aligned}
&\frac{1}{\mathbb{P}(S_N > a_n)} \Bigg(\sum_{k^*=1}^{k_{t+1}} \sum_{k_t=k^*}^{k_{t+1}} \mathbb{P}(N = k_t)\mathbb{P}((Y_1, \ldots, Y_{k_{t+1}}) \in B_{k^*} \cap A_{k_{t+1}}^{\otimes}) p(k_{t+1} \mid k^*) \\
&\qquad + \sum_{k^*=1}^{k_{t+1}} \sum_{k_t=k_{t+1}+1}^{\infty} \mathbb{P}(N = k_t)\mathbb{P}((Y_1, \ldots, Y_{k_{t+1}}) \in B_{k^*} \cap A_{k_{t+1}}^{\otimes}) p(k_{t+1} \mid k^*) \Bigg).
\end{aligned}$$

https://doi.org/10.1239/jap/1402578630 Published online by Cambridge University Press

Summing over k_t, this expression becomes

$$\frac{1}{\mathbb{P}(S_N > a_n)}\Bigg(\sum_{k^*=1}^{k_{t+1}} \mathbb{P}((Y_1,\dots,Y_{k_{t+1}}) \in B_{k^*} \cap A_{k_{t+1}}^{\otimes}) p(k_{t+1} \mid k^*)\mathbb{P}(k^* \le N \le k_{t+1}) + \sum_{k^*=1}^{k_{t+1}} \mathbb{P}((Y_1,\dots,Y_{k_{t+1}}) \in B_{k^*} \cap A_{k_{t+1}}^{\otimes}) p(k_{t+1} \mid k^*)\mathbb{P}(N \ge k_{t+1}+1)\Bigg).$$

From the definition of $p(k_{t+1} \mid k^*)$, it follows that this expression is equal to

$$\begin{aligned}
&\frac{1}{\mathbb{P}(S_N > a_n)} \sum_{k^*=1}^{k_{t+1}} \mathbb{P}((Y_1,\dots,Y_{k_{t+1}}) \in B_{k^*} \cap A_{k_{t+1}}^{\otimes}) p(k_{t+1} \mid k^*)\mathbb{P}(N \ge k^*) \\
&\quad = \frac{1}{\mathbb{P}(S_N > a_n)} \sum_{k^*=1}^{k_{t+1}} \mathbb{P}((Y_1,\dots,Y_{k_{t+1}}) \in B_{k^*} \cap A_{k_{t+1}}^{\otimes})\mathbb{P}(N = k_{t+1}) \\
&\quad = \frac{1}{\mathbb{P}(S_N > a_n)} \mathbb{P}((Y_1,\dots,Y_{k_{t+1}}) \in A_{k_{t+1}}^{\otimes})\mathbb{P}(N = k_{t+1}) \\
&\quad = \mathbb{P}(N = k_{t+1},\ (Y_1,\dots,Y_{k_{t+1}}) \in A_{k_{t+1}}^{\otimes} \mid Y_1 + \cdots + Y_N > a_n),
\end{aligned}$$

which is the desired invariant distribution. This completes the proof.

Proposition 3.4. *The Markov chain $(\bar{\boldsymbol{Y}}_t)_{t\ge 0}$, generated by Algorithm 3.2, is uniformly ergodic. In particular, it satisfies the following minorization condition: there exists $\delta > 0$ such that*

$$\mathbb{P}(\bar{\boldsymbol{Y}}_1 \in B \mid \bar{\boldsymbol{Y}}_0 = \bar{\boldsymbol{y}}) \ge \delta \mathbb{P}((N, Y_1, \dots, Y_N) \in B \mid Y_1 + \cdots + Y_N > a_n)$$

for all $\bar{\boldsymbol{y}} \in A = \bigcup_{k\ge 1}\{(k, y_1, \dots, y_k)\colon y_1 + \cdots + y_k > a_n\}$ and all Borel sets $B \subset A$.

The proof requires only a minor modification to the nonrandom case, Proposition 3.2, and is therefore omitted.

Next, consider the distributional assumptions and the design of $V^{(n)}$. Let the distribution of the number of steps $\mathbb{P}(N^{(n)} \in \cdot)$ depend on n. By a similar reasoning to that used in the case of a nonrandom number of steps, the following assumption is imposed: the variables $N^{(n)}$, $Y_1, Y_2, \dots$, and the numbers a_n are such that

$$\lim_{n\to\infty} \frac{\mathbb{P}(Y_1 + \cdots + Y_{N^{(n)}} > a_n)}{\mathbb{P}(M_{N^{(n)}} > a_n)} = 1, \tag{3.5}$$

where $M_k = \max\{Y_1, \dots, Y_k\}$. Note that the denominator can be expressed as

$$\begin{aligned}
\mathbb{P}(M_{N^{(n)}} > a_n) &= \sum_{k=1}^{\infty} \mathbb{P}(M_k > a_n)\mathbb{P}(N^{(n)} = k) \\
&= \sum_{k=1}^{\infty} [1 - F_Y(a_n)^k]\mathbb{P}(N^{(n)} = k) \\
&= 1 - g_{N^{(n)}}(F_Y(a_n)),
\end{aligned}$$

where $g_{N^{(n)}}(t) = \mathbb{E}[t^{N^{(n)}}]$ is the generating function of $N^{(n)}$. Sufficient conditions for (3.5) to hold are given in [16, Theorem 3.1]. For instance, if $\bar{F}_Y$ is regularly varying at ∞ with index $\beta > 1$ and $N^{(n)}$ has a Poisson distribution with mean $\lambda_n \to \infty$ as $n \to \infty$, then (3.5) holds with $a_n = a\lambda_n$ for $a > 0$.

https://doi.org/10.1239/jap/1402578630 Published online by Cambridge University Press

Similarly to the nonrandom setting, a good candidate for $V^{(n)}$ is the conditional distribution

$$V^{(n)}(\cdot) = \mathbb{P}(\bar{\boldsymbol{Y}}^{(n)} \in \cdot \mid M_{N^{(n)}} > a_n).$$

Then $V^{(n)}$ has a known density with respect to $F^{(n)}(\cdot) = \mathbb{P}(\bar{\boldsymbol{Y}}^{(n)} \in \cdot)$ given by

$$\begin{aligned}\frac{\mathrm{d}V^{(n)}}{\mathrm{d}F^{(n)}}(k, y_1, \ldots, y_k) &= \frac{1}{\mathbb{P}(M_{N^{(n)}} > a_n)} 1\left\{(y_1, \ldots, y_k) \colon \bigvee_{j=1}^{k} y_j > a_n\right\} \\ &= \frac{1}{1 - g_{N^{(n)}}(F_Y(a_n))} 1\left\{(y_1, \ldots, y_k) \colon \bigvee_{j=1}^{k} y_j > a_n\right\}.\end{aligned}$$

The estimator of $q^{(n)} = \mathbb{P}(S_n > a_n)^{-1}$ is given by

$$\hat{q}_T^{(n)} = \frac{1}{T} \sum_{t=0}^{T-1} \frac{\mathrm{d}V^{(n)}}{\mathrm{d}F^{(n)}}(\bar{\boldsymbol{Y}}_t^{(n)}) = \frac{1}{g_{N^{(n)}}(F_Y(a_n))} \frac{1}{T} \sum_{t=0}^{T-1} 1\left\{\bigvee_{j=1}^{N_t} Y_{t,j} > a_n\right\}, \tag{3.6}$$

where $(\bar{\boldsymbol{Y}}_t^{(n)})_{t \geq 0}$ is generated by Algorithm 3.2.

Theorem 3.2. *Suppose that (3.5) holds. The estimator $\hat{q}_T^{(n)}$ in (3.6) has vanishing normalized variance. That is,*

$$\lim_{n \to \infty} (p^{(n)})^2 \operatorname{var}_{\pi_n}(\hat{q}_T^{(n)}) = 0,$$

where π_n denotes the conditional distribution $\mathbb{P}(\bar{\boldsymbol{Y}}^{(n)} \in \cdot \mid S_{N^{(n)}} > a_n)$.

The proof is practically identical to that of Theorem 3.1 and is therefore omitted.

Remark 3.4. Because the distribution of $N^{(n)}$ may depend on n, Theorem 3.2 covers a wider range of settings for random sums than those studied in [5] and [10] where the authors presented provably efficient importance sampling algorithms.

4. Numerical experiments

The theoretical results presented in this paper guarantee that $\hat{q}_T^{(n)}$ is an efficient estimator of $(p^{(n)})^{-1}$. However, for comparison with existing algorithms, the numerical experiments presented in this section are based on $\hat{p}_T^{(n)} = (\hat{q}_T^{(n)})^{-1}$ as an estimator for $p^{(n)}$. The literature already contains numerical comparison of many of the existing algorithms. In particular, in the setting of random sums. Numerical results for the algorithms by Dupuis *et al.* [10], the hazard rate twisting algorithm by Juneja and Shahabuddin [15], and the conditional Monte Carlo algorithm by Asmussen and Kroese [4] can be found in [10]. Additional numerical results for the algorithms by Blanchet and Li [5], Dupuis *et al.* [10], and Asmussen and Kroese [4] can be found in [5]. From the existing results, it appears as if the algorithm by Dupuis *et al.* [10] has the best performance. Therefore, we only include numerical experiments of the MCMC estimator and the estimator in [10], which uses importance sampling (IS) for sums.

By construction, each simulation run of the MCMC algorithm generates only a single random variable (one step of the random walk), while both IS and standard MC generate n random variables (n steps of the random walk) for the case of a deterministic number of steps. Therefore, the number of runs for the MCMC is scaled up by a factor of n so that all algorithms (MCMC,

https://doi.org/10.1239/jap/1402578630 Published online by Cambridge University Press

TABLE 1: The table displays the batch mean and standard deviation of the estimates of $\mathbb{P}(S_n > a_n)$ as well as the average runtime per batch. The number of batches run is 20, each consisting of $T = 10^5$ simulations for IS and MC, and Tn simulations for MCMC. The asymptotic approximation is $p_{\max} = \mathbb{P}(\max\{Y_1, \ldots, Y_n\} > a_n)$.

$\beta = 2,\ n = 5$	$a = 20,\ p_{\max} = 4.901\text{e}-4$			$a = 10^4,\ p_{\max} = 1.999\,92\text{e}-9$		
	MCMC	IS	MC	MCMC	IS	MC
Average estimate	5.340e − 4	5.343e − 4	5.380e − 4	2.000 25e − 9	2.000 91e − 9	—
Standard deviation	6e − 7	13e − 7	770e − 7	7e − 14	215e − 14	—
Average time	14.4	13.9	1.5	15.9	15.9	—
$\beta = 2,\ n = 20$	$a = 20,\ p_{\max} = 1.2437\text{e}-4$			$a = 10^4,\ p_{\max} = 5.0000\text{e}-10$		
	MCMC	IS	MC	MCMC	IS	MC
Average estimate	1.2614e − 6	1.2615e − 6	1.2000e − 6	5.0010e − 10	5.0006e − 10	—
Standard deviation	4e − 10	12e − 10	33 166e − 10	2e − 14	71e − 14	—
Average time	49.4	48.4	1.9	48.0	47.1	—

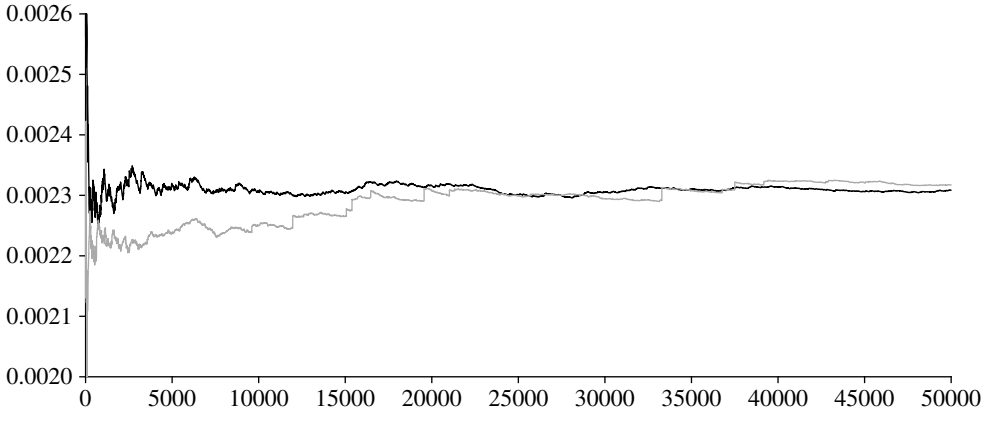

FIGURE 1: An illustration of the point estimate of $\mathbb{P}(S_n > a_n)$ as a function of the number of simulation steps, with $n = 5$, $a = 10$, and $\beta = 2$. The estimate generated by the MCMC approach is shown with a black line and the estimate generated by importance sampling is shown with a grey line.

MC, and IS) generate essentially the same number of random numbers. This gives a fair comparison of the computer runtime between the three approaches.

First consider estimating $\mathbb{P}(S_n > a_n)$, where $S_n = Y_1 + \cdots + Y_n$ with Y_1 having a Pareto distribution with density $f_Y(x) = \beta(x+1)^{-\beta-1}$ for $x \geq 0$. Let $a_n = an$. Each estimate is calculated using 20 batches, each consisting of $T = 10^5$ simulations in the case of IS and standard MC, and Tn simulations in the case of MCMC. The batch sample mean and sample standard deviation is recorded as well as the average runtime per batch. The results are presented in Table 1. The convergence of the algorithms can also be visualized by considering the point estimate as a function of number of simulation steps. This is presented in Figure 1. The MCMC algorithm appears to perform comparably with the IS algorithm for p up to order 10^{-4} which is a relevant range in, say, insurance and finance. However, for smaller p the MCMC algorithm appears to performs better. The improvement over IS appears to increase as the event becomes

https://doi.org/10.1239/jap/1402578630 Published online by Cambridge University Press

TABLE 2: The table displays the batch mean and standard deviation of the estimates of $\mathbb{P}(S_N > a_\rho)$ as well as the average runtime per batch. The number of batches run is 20, each consisting of $T = 10^5$ simulations for IS and MC, and $T\mathbb{E}[N]$ simulations for MCMC. The asymptotic approximation is $p_{\max} = \mathbb{P}(\max\{Y_1, \ldots, Y_N\} > a_\rho)$.

$\beta = 1,\ \rho = 0.2$	$a = 10^3,\ p_{\max} = 0.999\text{e}-3$			$a = 5 \times 10^9,\ p_{\max} = 2.0000\text{e}-10$		
	MCMC	IS	MC	MCMC	IS	MC
Average estimate	1.019e − 3	1.012e − 3	1.037e − 3	2.000 003e − 8	1.999 325e − 8	—
Standard deviation	1e − 6	3e − 6	76e − 6	6e − 14	1114e − 14	—
Average time	25.8	11.1	1.2	385.3	139.9	—
$\beta = 1,\ \rho = 0.05$	$a = 10^3,\ p_{\max} = 0.999\text{e}-3$			$a = 5 \times 10^5,\ p_{\max} = 1.9999\text{e}-6$		
	MCMC	IS	MC	MCMC	IS	MC
Average estimate	1.027e − 3	1.017e − 3	1.045e − 3	2.0002e − 6	2.0005e − 6	3.2000e − 6
Standard deviation	1e − 6	4e − 6	105e − 6	1e − 10	53e − 10	55 678e − 10
Average time	61.5	44.8	1.3	60.7	45.0	1.3

more rare. This is due to the fact that the asymptotic approximation becomes better and better as the event becomes more rare.

Now, consider estimating $\mathbb{P}(S_N > a_\rho)$, where $S_N = Y_1 + \cdots + Y_N$ with N geometrically distributed $\mathbb{P}(N = k) = (1-\rho)^{k-1}\rho$ for $k = 1, 2, \ldots$ and $a_\rho = a\mathbb{E}[N] = a/\rho$. The numerical results presented here are for the estimator $\hat{p}_T = (\hat{q}_T)^{-1}$ with $\hat{q}_T$ as in (3.6). Again, each estimate is calculated using 20 batches, each consisting of $T = 10^5$ simulations in the case of IS and standard MC, and $T\mathbb{E}[N]$ simulations in the case of MCMC. The results are presented in Table 2. Again, the MCMC algorithm appears to have performance comparable to the IS algorithm for p up to the order 10^{-4} and better performance as p gets smaller.

Acknowledgements

Henrik Hult's research was supported by the Göran Gustafsson Foundation. The authors thank the anonymous referee and Editor for useful comments that helped improve the manuscript.

References

[1] ASMUSSEN, S. (2003). *Applied Probability and Queues* (Stoch. Modelling Appl. **51**). Springer, New York.
[2] ASMUSSEN, S. AND BINSWANGER, K. (1997). Simulation of ruin probabilities for subexponential claims. *Astin Bull.* **27,** 297–318.
[3] ASMUSSEN, S. AND GLYNN, P. W. (2007). *Stochastic Simulation* (Stoch. Modelling Appl. **57**). Springer, New York.
[4] ASMUSSEN, S. AND KROESE, D. P. (2006). Improved algorithms for rare event simulation with heavy tails. *Adv. Appl. Prob.* **38,** 545–558.
[5] BLANCHET, J. AND LI, C. (2011). Efficient rare-event simulation for heavy-tailed compound sums. *ACM Trans. Model. Comput. Simul.* **21,** 23pp.
[6] BLANCHET, J AND LIU, J. C. (2008). State-dependent importance sampling for regularly varying random walks. *Adv. Appl. Prob.* **40,** 1104–1128.
[7] CHAN, K. S. (1993). Asymptotic behavior of the Gibbs sampler. *J. Amer. Statist. Assoc.* **88,** 320–326.
[8] CLINE, D. B. H. AND HSING, T. (1994). Large deviation probabilities for sums of random variables with heavy or subexponential tails. Tech. Rep., Texas A&M University.
[9] DENISOV, D., DIEKER, A. B. AND SHNEER, V. (2008). Large deviations for random walks under subexponentiality: the big-jump domain. *Ann. Prob.* **36,** 1946–1991.

[10] Dupuis, P., Leder, K. and Wang, H. (2007). Importance sampling for sums of random variables with regularly varying tails. *ACM Trans. Model. Comput. Simul.* **17,** 21pp.
[11] Gelman, A. and Meng, X. L. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13,** 163–185.
[12] Gilks, W. R., Richardsson, S. and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.
[13] Hult, H. and Svensson, J. (2012). On importance sampling with mixtures for random walks with heavy tails. *ACM Trans. Model. Comput. Simul.* **22,** 21pp.
[14] Jones, G. L. (2004). On the Markov chain central limit theorem. *Prob. Surveys* **1,** 299–320.
[15] Juneja, S. and Shahabuddin, P. (2002). Simulation heavy tailed processes using delayed hazard rate twisting. *ACM Trans. Model. Comput. Simul.* **12,** 25pp.
[16] Klüppelberg, C. and Mikosch, T. (1997). Large deviations for heavy-tailed random sums with applications to insurance and finance. *J. Appl. Prob.* **37,** 293–308.
[17] Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* **24,** 101–121.
[18] Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer, London.
[19] Nummelin, E. (1984). *General Irreducible Markov Chains and Non-negative Operators*. Cambridge University Press.
[20] Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **90,** 558–566.
[21] Smith, A. F. M. and Gelfand, A. E. (1992). Bayesian statistics without tears: a sampling-resampling perspective. *Amer. Statist.* **46,** 84–88.
[22] Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22,** 1701–1762.