KTH Engineering Sciences

Rare-event simulation with Markov chain Monte Carlo

THORBJÖRN GUDMUNDSSON

Doctoral Thesis
Stockholm, Sweden 2015

TRITA-MAT-A 2014:18
ISRN KTH/MAT/A-14/18-SE
ISBN 978-91-7595-404-2

Department of Mathematics
Royal Institute of Technology
100 44 Stockholm, Sweden

Akademisk avhandling som med tillstånd av Kungl Tekniska Högskolan framlägges till offentlig granskning för avläggande av teknologie doktorsexamen, fredagen den 23 januari 2015 klockan 14.00 i sal F3, Lindstedtsvägen 26, Kungl Tekniska Högskolan, Stockholm.

Print: Universitetsservice US AB, Stockholm, 2015

Til Rannveigar og Gyðu

"A reader lives a thousand lives before he dies."

-Jojen Reed. A Dance with Dragons.

Abstract

Stochastic simulation is a popular method for computing probabilities or expectations where analytical answers are difficult to derive. It is well known that standard methods of simulation are inefficient for computing rare-event probabilities and therefore more advanced methods are needed to those problems.

This thesis presents a new method based on Markov chain Monte Carlo (MCMC) algorithm to effectively 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 normalising constant. Using the MCMC methodology a Markov chain is simulated, with that conditional distribution as its invariant distribution, and information about the normalising constant is extracted from its trajectory.

In the first two papers of the thesis, the algorithm is described in full generality and applied to four problems of computing rare-event probability in the context of heavy-tailed distributions. The assumption of heavy-tails allows us to propose distributions which approximate the conditional distribution conditioned on the rare event. The first problem considers a random walk $Y_1 + \cdots + Y_n$ exceeding a high threshold, where the increments Y are independent and identically distributed and heavy-tailed. The second problem is an extension of the first one to a heavy-tailed random sum $Y_1 + \cdots + Y_N$ exceeding a high threshold, where the number of increments N is random and independent of $Y_1, Y_2, \ldots$. The third problem considers the solution X_m to a stochastic recurrence equation, $X_m = A_m X_{m-1} + B_m$, exceeding a high threshold, where the innovations B are independent and identically distributed and heavy-tailed and the multipliers A satisfy a moment condition. The fourth problem is closely related to the third and considers the ruin probability for an insurance company with risky investments.

In last two papers of this thesis, the algorithm is extended to the context of light-tailed distributions and applied to four problems. The light-tail assumption ensures the existence of a large deviation principle or Laplace principle, which in turn allows us to propose distributions which approximate the conditional distribution conditioned on the rare event. The first problem considers a random walk $Y_1 + \cdots + Y_n$ exceeding a high threshold, where the increments Y are independent and identically distributed and light-tailed. The second problem considers a discrete-time Markov chains and the computation of general expectation, of its sample path, related to rare-events. The third problem extends the the discrete-time setting to Markov chains in continuous-time. The fourth problem is closely related to the third and considers a birth-and-death process with spatial intensities and the computation of first passage probabilities.

An unbiased estimator of the reciprocal probability for each corresponding problem is constructed with efficient rare-event properties. The algorithms are illustrated numerically and compared to existing importance sampling algorithms.

Sammanfattning

Stokastisk simulering är polulär metod för beräkning av sannolikheter eller väntevärde när analytiska svar är svåra att härleda. Det är känt att standard metoder inom simulering är ineffektiva för att beräkna sannolikheter av en sällsynta händelser och därför behövs det mer avancerade metoder till de typen av problem.

I denna avhandling presenteras en ny metod baserad på Markov chain Monte Carlo (MCMC) för att effektivt beräkna sannolikheten av en sällsynt händelse. Den betingade fördelningen för den underliggande processen givet att den sällsynta händelsen inträffar har den sökta sannolikheten som sin normaliseringskonstant. Med hjälp av MCMC-metodiken skapas en Markovkedja med betingade fördelningen som sin invarianta fördelning och en skattning av normaliseringskonstanten baseras på den simulerade kedjan.

I de två första pappren i avhandlingen, beskrivs algoritmen i full generalitet och tillämpas på fyra exempelproblem för beräkning av små sannolikheter i tungsvansade sammanhang. Det tungsvansade antagandet innebär att vi kan föreslå en fördelning som approximerar den betingade fördelningen givit den sällsynta händelsen. Första problemet handlar om en slumpvandring $Y_1 + \cdots + Y_n$ som överskrider en hög tröskel, då stegen Y är oberoende, likafödelade med tungsvansad fördelning. Andra problemet är en utvidgning av det första till summa av ett stokastiskt antal termer $Y_1 + \cdots + Y_N$ som överskrider en hög tröskel, då antalet steg N är stokastiskt och oberoende av $Y_1, Y_2, \ldots$. Tredje problemet behandlar sannolikheten att lösningen X_m till en stokastisk rekurrensekvation, $X_m = A_m X_{m-1} + B_m$, överskrider en hög tröskel då innovationerna B är oberoende, likafördelade med tungsvansad fördelning och multiplikatorerna A satisfierar ett moment villkor. Sista problemet är nära kopplat till det tredje och handlar om ruinsannolikhet för ett försäkringsbolag med riskfyllda investeringar.

I de två senare pappren i avhandlingen, utvidgas algoritmen till lättsvansade sammanhang och tillämpas på fyra exempelprolem. Det lättssvansade antagandet säkerställer existensen av stora avvikelse princip eller Laplace princip som i sin tur innebär att vi kan föreslå fördelningar som approximerar den betingade fördelningen betingad på den sällsynta händelsen. Första problemet handlar om en slumpvandring $Y_1 + \cdots + Y_n$ som överskrider en hög tröskel, då stegen Y är oberoende, likafördelade med lättsvansad fördelning. Andra problemet handlar om Markovkedjor i diskret tid och beräkning på almänna väntevärde, av dess trajektoria, relaterad till sällsynta händelser. Det tredje problemet utvidgar den diskreta bakgrunden till Markovkedjor i kontinuerlig tid. Det fjärde problemet är nära kopplat till det tredje och handlar om en födelse-och-döds process med rumsberoende intensiteter och beräkning av första övergångs sannolikheter.

För varje exempelproblem konstrueras en väntevärdesriktig skattning av den reciproka sannolikheten. Algoritmerna illustreras numeriskt och jämförs med existerande importance sampling algoritmer.

Acknowledgments

I made it! But I would never have succeeded if it had not been for the immense support and help that I got from you all.

My deepest appreciation goes to Associate Professor Henrik Hult, who has been my advisor and co-author during my years at KTH. Your creative mind and mathematical expertise have been a constant source of motivation. You have helped me restructure entirely the way I think about research and mathematical problems, seeing beyond and understanding the key concepts. Separating the technical details from the fundamental ideas. I am thankful that you never lost confidence in me and gave me continuous encouragement, even when the progress was slow. You have not only been great collaborator but also a good friend.

I would like to thank Associate Professor Filip Lindskog for his valuable contribution to my research education. The courses that you gave me, significantly raised my level of thinking about mathematics. Your open-door policy was much appreciated.

Further, I would like to express my gratitude to Tobias Rydén, a former Professor at the Institution, for our discussions about the ergodic properties of Markov chains and for the long overdue loan of Meyn and Tweedie. Thanks go also to Associate Professor Jimmy Olsson for his helpful comments on how to simulate a Markov chain.

I want to thank the faculty members at the Institution. Special thanks to Professor Timo Koski, affiliated Professor Tomas Björk, Professor Anders Szepessy and Professor Krister Svanberg for the excellent courses they gave me. Thanks to Associate Professor Gunnar Englund and Associate Professor emeritus Harald Lang for their helpful pedagogical advises and lively discussions during lunches. I am grateful to all at the Institution for making me feel welcome to the group and their friendly attitude towards me.

A big thanks go to my office buddies, Björn Löfdahl, the maester of Game of Thrones theory, and Johan Nykvist, my personal tutor in the Swedish language and culture. It has been fun hanging with you, whether at the blackboard confounded over how to compute some integral, at the chess table or in the ping-pong corridor. Many thanks go to my former colleague Dr Pierre Nyquist for his endless help on math problems. Your comments and corrections on this thesis were also highly appreciated. Cheers to my fellow PhD students, this has been a really entertaining journey with all of you.

I am forever grateful to my parents and family. You have always been there for me, cheering me onwards and giving me the confidence I needed to cross the line. I am so lucky to have you on my team. A special thanks goes to my brother Skúli. Your enthusiasm for mathematics have formed me more than you might believe. Thanks for all of our discussions.

Finally, I want to give thanks to the ones who mean the world to me. Rannveig and Gyða, with your love and support, everything is possible.

Stockholm, December 2014

Thorbjörn Gudmundsson

Table of Contents

Introduction

The theme of this thesis is the study of efficient stochastic simulation algorithms, consisting of four scientific papers. It considers the special case when the property under investigation is governed by an event which is thought of as rare in the sense it occurs with a small probability. This case is of particular interest as the standard tools within stochastic simulation fail in the rare-event setting. The method presented in this thesis uses the theory of Markov chain Monte Carlo (MCMC) which, to the best of the author's knowledge, has not been applied in the rare event simulation context before. The main contribution is a new stochastic simulation method which we will name the MCMC method. The thesis has a natural split into two parts. The first two papers which are presented in Chapter 2 and 3 assume the setting of heavy-tailed random variables, whilst the latter two papers which are presented in Chapter 4 and 5 assume the setting of light-tailed random variables.

It is therefore important for the reader to be familiar with some of the underlying theory such as rare event simulation and Markov chain Monte Carlo before embarking on reading the thesis. The introduction starts with a motivation for simulation in real-world applications, presented in Section 1.1. In Section 1.2 we present key concepts to this thesis such as Markov chain Monte Carlo, rare-event simulation and importance sampling. In Section 1.3 we present the primary contribution of this thesis, namely, a general description to estimating rare-event probabilities using Markov chain Monte Carlo methodology. The method is explained and the crucial design choices which determine its performance are highlighted. Finally, in Section 1.4 we provide summaries of the four papers which build up the thesis.

1.1 Background

Mathematical modelling has been an fundamental part of scientific progress. We model complex phenomena to enhance our understanding and to better predict properties currently unknown to us. The possible applications of modelling are endless and come up in fields such as physics, chemistry, economics and finance to name but few. More often than not the structure of the model depends on unknown factors, parameters which specify the detailed behaviour of the model. These factors need to be assigned some values to use the model for numerical purposes. This assignment is usually done by estimating the value of the factors but it introduces the possibility of an error due to the stochastic fluctuation in the estimation. It is natural to take the random error into account when modelling and expanding the framework accordingly. These types of models are called stochastic models.

The computational capacity has increased significantly in recent decades with the relative easy access to supercomputers. This has in turn allowed for more and more complicated stochastic models for applications. Finer components, which previously were not included, can now incorporated in the models, thus increasing the complexity. Researchers and practitioners alike strive to enhance the current models and introduce more and more details to it. This has had a positive effects on modern modelling, albeit not without some cost. Many stochastic models today have become so involved that it is becoming difficult to derive any analytical answers to the questions posed to the model. This has given rise to alternative approach to handling such complex stochastic models, namely stochastic simulation.

Briefly, simulation is the process of sampling the underlying random factors to generate many instances of the model. These multiple instances of the model, called the sample, gives the investigator an insight into the object being modelled and is used to make inferences about its properties. This has proved to be a powerful tool for computation. Generating instances of highly advanced models, multi-dimensional, non-linear and stochastic models can be done in a few milliseconds. Stochastic simulation has thus played its part in the scientific progress of recent decades and simulation itself has grown into an academic field in its own right.

In physics, hypothesis are often tested and verified via a number of experiments. One experiment is carried out after another, and if sufficiently many of the experiments support the hypothesis then it acquires a certain validity and becomes a theory. This was for instance the case at CERN in the summer of 2012, when the existence of the Higgs boson was confirmed through experiments which supported the old and well known hypothesis. However, one can not always carry out experiments to validate hypotheses. Sometimes it is simply impossible to replicate the model in reality, as is the case when studying the effects of global warming. Obviously, since we can only generate a single physical instance of the Earth, any simulations need to be done via computer modelling. To better reflect reality, the resolution needs to be high and many different physical and meteorological factors need to be taken into account. The surface of the Earth is broken into 10km times 10km squares, each with its temperature, air pressure, moisture and more. The dynamics of these weather factors need to be simulated with small time steps, perhaps many years into the future. The Mathematics and Climate Research Network (MCRN) carries out extensive stochastic simulations, replicating the Earth using different types of scenarios to forecast possible climate changes. Clearly, this type of stochastic simulation is immensely computationally costly. This scientific work alone justifies the importance of continuing research and improvement in the field of stochastic simulation.

In some contexts the properties being investigated are highly affected by the occurrence of so-called rare-events. This is for instance the case for evaluation of most risk measures in finance or insurance. The capital requirements or Value-at-Risk typically represent how much money needs to be set aside to serve as a buffer in the worst case scenario. The standard methods of stochastic simulation have had problems handling these unlikely events of small probability. This is because generating an instance of a very unlikely event in a model typically involves generating a very large number of instances and thus requiring prohibitive amount of computer time. As a consequence the investigation of rare-events

is both time-consuming and ineffective, motivating the need for further research for these rare-event problems. The field of stochastic simulation which focuses on these problems is called rare-event simulation. The importance of expanding our boundaries to understanding rare-event simulation is ever present as the usage of stochastic simulation continues to increase. Effective simulation methods for rare-events could be highly beneficial in areas such as finance and insurance.

1.2 Stochastic simulation

In this section we give introduction to some of the key building blocks of stochastic simulation with emphasis on those topics important for this thesis.

1.2.1 Sampling a random variable

The foundations of stochastic simulation in computers is the generation of a pseudo random number. We present the general theory and how it can be used to sample a random variable via the inversion method, which is central to the Markov chain Monte Carlo method.

Most statistical software programs provide methods for generating a uniformly distributed pseudo random number on the interval, say, $[0, 1]$. These algorithms are deterministic, at its core, and can only imitate the properties and behaviour of a uniformly distributed random variable. The early designs of such algorithms showed flaws in the sense that the pseudo random numbers generated followed a pattern which could easily be identified and predicted. Nowadays there exists many highly advanced algorithms that generate pseudo random numbers, mimicking a true random number quite well. For the purposes of this thesis we assume the existence of an algorithm producing a uniformly distributed pseudo random number, and ignore any deficiencies and errors arising from the algorithm. In short, we assume that we can sample a perfectly uniformly distributed random variable in some computer program. For a more thorough and detailed discussion we refer to [29].

Now consider a random variable X and denote by F its probability distribution. Say we would like, via some computer software, to sample the random variable X. One approach is the inversion method. The inversion method involves only applying the quantile function to uniformly random variable. The algorithm is as follows.

1. Sample U from the standard uniform distribution.
2. Compute $Z = F^{-1}(U)$,

where $F^{-1}(U) = \inf_{x \in U}\{x \mid F(x) \geq p\}$. The random variable Z has the same distribution as X as the following display shows.

$$\mathbf{P}(Z \leq x) = \mathbf{P}(F^{-1}\{U\} \leq x) = \mathbf{P}(U \leq F(x)) = F(x).$$

The method can easily be extended to sampling X conditioned on being larger than some constant c. Meaning that we want to sample from the conditional distribution

$$\mathbf{P}(X \in \cdot \mid X > c).$$

The algorithm is formally as follows.

1. Sample U from the standard uniform distribution.
2. Compute $Z = F^{-1}\Big(\big(1 - F(c)\big)U + F(c)\Big)$.

The distribution of Z is given by,

$$\begin{aligned}\mathbf{P}(Z \leq x) &= \mathbf{P}\big((1 - F(c))U + F(c) \leq F(x)\big) = \mathbf{P}\Big(U \leq \frac{F(x) - F(c)}{1 - F(c)}\Big) \\ &= \frac{F(x) - F(c)}{1 - F(c)} = \frac{\mathbf{P}(c \leq X \leq x)}{\mathbf{P}(X > c)} = \mathbf{P}(X \leq x \mid X > c).\end{aligned}$$

Thus the inversion method provides a simple way of sampling a random variable, conditioned on being larger than c, based solely on the generation of a uniformly distributed random number.

1.2.2 Rare-event simulation

In stochastic simulation there are two convergence properties which determine if an estimator is efficient or not. Firstly we require the estimator to be large-sample efficient which means that the variance of the estimator tends to zero as the sample size increases. Secondly we want the estimator to be rare-event efficient meaning that the estimate is accurate even when the sought probability is very small. To be more precise let us consider the canonical example of stochastic simulation, the standard Monte Carlo estimator.

The power of Monte Carlo is its simplicity but it lacks rare-event efficiency. To explain this, let us describe what is meant by rare-event simulation. Consider a sequence of random variables $X^{(1)}, X^{(2)}, \ldots$ where each can be sampled repeatedly by a simulation algorithm. Suppose we want to compute the probability $p^{(n)} = \mathbf{P}(X^{(n)} \in A)$ for some Borel set A and large n where it is assumed that $p^{(n)} \to 0$ as $n \to \infty$. The idea of Monte Carlo is to sample independent and identically distributed copies of the random variable $X^{(n)}$ and count the number of times it hits the set A. For a sample $X_0^{(n)}, \ldots, X_{T-1}^{(n)}$ the Monte Carlo estimator is given by

$$\widehat{p}_{\text{MC}}^{(n)} = \frac{1}{T}\sum_{t=0}^{T-1} I\{X_t^{(n)} \in A\}.$$

Of course, the variance $\text{Var}(\widehat{p}^{(n)}) = \frac{1}{T}p^{(n)}(1 - p^{(n)})$ tends to zero as $T \to \infty$, ensuring the large-sample efficiency of the Monte Carlo estimator but that is not main concern here. For an unbiased estimator $\widehat{p}^{(n)}$ of $p^{(n)}$ the natural rare-event performance criteria is that the relative error is controlled as $n \to \infty$. In the case of the Monte Carlo estimator

$$\frac{\text{Var}(\widehat{p}_{\text{MC}}^{(n)})}{(p^{(n)})^2} = \frac{p^{(n)}(1 - p^{(n)})}{T(p^{(n)})^2} = \frac{1}{T}\Big(\frac{1}{p^{(n)}} - 1\Big) \to \infty, \quad \text{as } n \to \infty,$$

indicating that the performance deteriorates when the event is rare. For example, if a relative error at 1% is desired and the probability is of order 10^{-6} then we need to take T

such that $\sqrt{(10^6-1)/T} \leq 0.01$. This implies that $T \approx 10^{10}$ which is infeasible on most computer systems.

In Section 1.2.5 we give description to alternative rare-event simulation algorithms that are rare-event efficient. But before presenting these algorithms we first give a brief summary of what is meant by an efficient algorithm. Moreover, we explain the difference between the two classes of problems considered in this thesis, namely the heavy-tailed class and the light-tailed class.

1.2.3 Efficiency properties

There are roughly two types of efficiency properties desirable in the rare-event simulation context. Firstly there is the strong efficiency, which is characterised by the estimator's relative error $\mathrm{RE}(\widehat{p}^{(n)}) = \mathrm{Std}(\widehat{p}^{(n)})/p^{(n)}$. An estimator is said to have vanishing relative error if

$$\mathrm{RE}(\widehat{p}^{(n)}) \to 0, \quad \text{as } n \to \infty,$$

and bounded relative error if

$$\mathrm{RE}(\widehat{p}^{(n)}) < \infty, \quad \text{as } n \to \infty.$$

Secondly there is the slightly weaker performance criteria, called logarithmic efficiency. An estimator is said to be logarithmically efficient if

$$\frac{1}{n} \log \frac{\mathbf{E}\big[(p^{(n)})^2\big]}{(p^{(n)})^2} \to 0, \quad \text{as } n \to \infty.$$

The interpretation is that the exponential rate of decay of the second moment of $\widehat{p}^{(n)}$ coincides with that of $(p^{(n)})^2$. From a practical perspective bounded relative error implies that sample size needs to be increased by n to obtain a certain accuracy whereas logarithmic efficiency implies that samples size increases at most sub exponentially with n.

1.2.4 Heavy and light tails

In the first two chapters of this thesis we assume the heavy-tailed setting as opposed to the latter two chapters where we assume the light-tailed setting. This types of probabilistic assumptions is important as it determines how we approach the problem and design the algorithm. We use the asymptotic properties, derived from the tail assumption, to ensure the rare-event performance is good.

The notion of heavy tails refers to the rate of decay of the tail $\overline{F} = 1 - F$ of a probability distribution function F. A popular class of heavy-tailed distributions is the class of subexponential distributions. A distribution function F supported on the positive axis is said to belong to the subexponential distributions if

$$\lim_{x \to \infty} \frac{\mathbf{P}(X_1 + X_2 > x)}{\mathbf{P}(X_1 > x)} = 2,$$

for independent random variables X_1 and X_2 with distribution F. A subclass of the subexponential distributions is the regularly varying distributions. $\overline{F}$ is called regularly varying (at ∞) with index $-\alpha \leq 0$ if

$$\lim_{t\to\infty} \frac{\overline{F}(tx)}{\overline{F}(t)} = x^{-\alpha}, \quad \text{for all } x > 0.$$

The heavy-tailed distributions are often described with the one big jump analogy, meaning that the event of a sum of heavy-tailed random variables being large is dominated by the case of one of the variables being very large whilst the rest are relatively small. This is in sharp contrast to the case of light-tails, where the same event is dominated by the case of every variable contributing equally to the total. As a reference to the one big jump analogy we refer the reader to [19, 21, 11].

This one big jump phenomenon has been observed in empirical data. For instance, when we consider stock market indices such as Nasdaq, Dow Jones etc. it turns out that the distribution of daily log returns typically has a heavy left tail, see Hult et al. in [20]. Another example is the well studied Danish fire insurance data, which consists of real-life claims caused by industrial fires in Denmark. While the arrivals of claims is showed to be not far from Poisson, the claim size distribution shows clear heavy-tail behaviour. The data set is analysed by Mikosch in [27] and the tail of the claim size is shown to be fit well with a Pareto distribution.

Stochastic simulation in the presence of heavy-tailed distributions has been studied with much interest in recent years. The conditional Monte Carlo technique was applied on this setting by Asmussen et al. [1, 3]. Dupuis et al. [12] used importance sampling algorithm in a heavy-tailed setting. Finally we mention the work of Blanchet et al. considering heavy-tailed distributions in [8, 7].

Turning to the second class of tail probabilities, a distribution function F is said to be light-tailed if its tail decays at an exponential rate or faster. A more formal description is as follows. A random variable X is said to have light-tailed distribution F if $\Lambda(\theta) = \log \mathbf{E}[e^{\theta X}] < \infty$ for some $\theta > 0$. Typical light-tailed distributions are the normal distribution, exponential distribution, gamma distribution and the compound Poisson process.

Closely related to the light-tailed assumption is the theory of large deviation principles. The sequence $\{X^{(n)}\}$ is said to satisfy the large deviation principle on $\mathscr{E}$ with rate function I if for each closed $F \subseteq \mathscr{E}$

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}(X^{(n)} \in F) \leq -I(F),$$

and for each open $G \subseteq \mathscr{E}$

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}(X^{(n)} \in G) \geq -I(G).$$

1.2.5 Importance sampling

We take up the thread left in Section 1.2.2 and present an alternative to the Monte Carlo method for efficiently running rare-event simulation.

Many simulation techniques have been presented to the scientific community to improve the poor rare-event properties of the Monte Carlo. The main idea of the variance reduction methods is to introduce a control mechanism that steers the samples towards the relevant part of the state space, thereby increasing the relevance of each sample. Few have had same popularity as importance sampling which is undoubtedly one of the more successful rare-event simulation techniques explored until today. The method was first introduced by Siegmund in 1976, see [31] and has since been widely developed. The control mechanism behind importance sampling is the change of sampling distribution.

A formal description of importance sampling is as follows. Let $X^{(1)}, X^{(2)}, \ldots$ be a sequence of random variables each of which can be sampled repeatedly and suppose that we want to compute the probability $p^{(n)} = \mathbf{P}(X^{(n)} \in A)$ for some large n and that $p^{(n)} \to 0$ as $n \to \infty$. Denote by $F^{(n)}$ the distribution of a random variable $X^{(n)}$. Instead of sampling from the original distribution $F^{(n)}$ then the $X_0^{(n)}, \ldots, X_{T-1}^{(n)}$ are sampled from a so-called sampling distribution denoted by $G^{(n)}$ such that $F^{(n)} \ll G^{(n)}$ on A. The sampling distribution $G^{(n)}$ is chosen such that we obtain more samples where $\{X^{(n)} \in A\}$. The importance sampling estimator is the average of the hitting probabilities, weighted with the respective Radon-Nikodym derivative,

$$\widehat{p}_{\text{IS}}^{(n)} = \frac{1}{T} \sum_{t=0}^{T-1} \frac{dF^{(n)}}{dG^{(n)}}(X_t^{(n)}) I\{X_t^{(n)} \in A\}.$$

This is an unbiased and consistent estimator since

$$\mathbf{E}_{G^{(n)}}[\widehat{p}_{\text{IS}}] = \int_A \frac{dF^{(n)}}{dG^{(n)}}(X^{(n)}) dG^{(n)}(X^{(n)}) = \mathbf{P}(X^{(n)} \in A).$$

The difficult task of importance sampling is the design of the sampling distribution $G^{(n)}$. Informally $G^{(n)}$ is chosen with two objectives in mind, apart from the necessary condition that $F^{(n)}$ needs to be absolutely continuous with respect to $G^{(n)}$. Firstly we want many more samples hitting the event, meaning that $\{X^{(n)} \in A\}$ is more likely under $G^{(n)}$ than $F^{(n)}$. Secondly the Radon-Nykodym derivative $dF^{(n)}/dG^{(n)}$ may not become too large and thereby ruining the stability of the method. Choosing the sampling distribution to be equal to the conditional distribution

$$F_A^{(n)}(\cdot) = \mathbf{P}(X^{(n)} \in \cdot \mid X^{(n)} \in A),$$

implies that $\widehat{p}_{\text{IS}}^{(n)}$ has zero variance and is therefore called the zero-variance distribution. The problem of choosing $F_A^{(n)}$ as a sampling distribution is that the likelihood ratio is $p^{(n)}$ which is unknown. Therefore $F_A^{(n)}$ is infeasible in practice but we want to to choose $G^{(n)}$ as an approximation of $F_A^{(n)}$.

Importance sampling quickly gained approval for problems in light-tailed setting, using a technique named exponential tilting or exponential change of measure. To better explain the solution let us consider the problem of computing the probability that random walk $S_n = Y_1 + \cdots + Y_n$ exceeds a high threshold a, that is $p^{(n)} = \mathbf{P}(S_n/n > a)$. Suppose that the increment Y has light-tailed distribution in the sense that $\Lambda(\theta) = \log \mathbf{E}[e^{\theta Y}] < \infty$, for some $\theta > 0$ and there exists a large deviation principle of the form

$$-\lim_{n\to\infty} \frac{1}{n} \log p^{(n)} = \Lambda^*(a),$$

where $\Lambda^*(a) = \sup_{\theta\in\mathbb{R}}\{\theta a - \Lambda(\theta)\}$ is the Fenchel-Legendre transform of Λ. The interpretation is that $p^{(n)} \approx e^{-n\Lambda^*(a)}$. Consider sampling the Y's from the exponentially tilted probability distribution

$$F_\theta(dy) = e^{\theta y - \Lambda(\theta)} F(dy).$$

Then the second moment of the estimator is

$$\begin{aligned}
\mathbf{E}_{F_\theta}\big[(\widehat{p}_{\text{IS}}^{(n)})^2\big] &= \mathbf{E}_F\big[I\{S_n/n > a\} \prod_{i=1}^{n} e^{\Lambda(\theta)-\theta Y_i}\big] \\
&= \int_a^\infty e^{n\Lambda(\theta)-n\theta y} F(dy) \\
&\approx \int_a^\infty e^{-n[\theta y - \Lambda(\theta) + \Lambda^*(y)]} dy \\
&\approx e^{-n[\theta a - \Lambda(\theta) + \Lambda^*(a)]},
\end{aligned}$$

where the approximations are made precise asymptotically by Varadhan's theorem [10][Thm 4.3.1, p. 137]. Thus we obtain an upper bound

$$\frac{1}{n} \log \mathbf{E}_{F_\theta}\big[(\widehat{p}_{\text{IS}}^{(n)})^2\big] \le -\sup_{\theta\in\mathbb{R}}\{\theta a - \Lambda(\theta) + \Lambda^*(a)\} = -2\Lambda^*(a).$$

By Jensen's inequality we have a lower bound

$$\liminf_{n\to\infty} -\frac{1}{n} \log \mathbf{E}_{F_\theta}\big[(\widehat{p}_{\text{IS}}^{(n)})^2\big] \ge \liminf_{n\to\infty} -\frac{1}{n} \log (p^{(n)})^2 \ge -2\Lambda^*(a).$$

Combining the upper and lower bound shows us that $\mathbf{E}_{F_\theta}\big[(\widehat{p}_{\text{IS}}^{(n)})^2\big] \approx e^{-2\Lambda^*(a)}$, meaning that the exponential rate of growth coincides with the sought probability squared, ensuring logarithmic efficiency.

In the context of importance sampling for light-tailed problems then two main approaches have been developed recently; the subsolution approach, based on control theory, by Dupuis, Wang, and collaborators, see e.g. [14, 15, 13], and the approach based on Lyapunov functions and stability theory by Blanchet, Glynn, and others, see [4, 5, 6, 8].

The technique of using exponential tilting for the class of heavy-tailed problems has had limited success as the exponential moments in such context do not exist. Additionally,

the heavy-tailed setting is more complicated as the conditional distribution $F_A^{(n)}$ typically becomes singular with respect to $F^{(n)}$ as $n \to \infty$. The efficient importance sampling algorithms for heavy-tailed setting was developed much later.

The technique of conditional Monte Carlo simulation was presented by Asmuessen and Kroese, see [3], and was shown to always provide variance reduction in the estimator. An importance sampling algorithm using hazard rate twisting was introduced in 2002 by Juneja and Shahabuddin [23]. In the heavy-tailed case this method becomes equivalent to changing the tail index of the distribution. An efficient importance sampling estimator for the heavy-tailed case for random walk was presented in 2007 by Dupuis et al. [12], based on mixture and sequential sampling.

The only drawback of the importance sampling approach is that despite its efficiency the mathematical proofs are lengthy and complex. In this thesis we suggest a different approach using Markov chain Monte Carlo which is more simple and easier to implement while still equally efficient as the importance sampling in the models which we have considered.

1.2.6 Markov chain Monte Carlo

In this section we present a sampling technique called Markov chain Monte Carlo (MCMC) for sampling a random variable X despite only having limited information about its distribution. MCMC is typically useful when sampling a random variable X having a density f that is only known up to a constant, say

$$f(x) = \frac{\pi(x)}{c},$$

where π is known but $c = \int \pi(x)dx$ is unknown. An example of this type of setup can be found in Bayesian statistics and hidden Markov chains.

In short, the basic idea of sampling via MCMC is to generate a Markov chain $(Y_t)_{t\geq 0}$ whose invariant density is the same as X, namely f. There exists plentiful of MCMC algorithms but we shall only name two in this thesis, the Metropolis-Hastings algorithm and the Gibbs algorithm.

The method first laid out by Metropolis [25] and then extended by Hastings [18] is based on a proposal density, which we shall denote by g. Firstly the Markov chain $(Y_t)_{t\geq 0}$ is initialised with some $Y_0 = y_0$. The idea behind the Metropolis-Hastings algorithm is to generate a proposal state Z using the proposal density g. The next state of the Markov chain is then assigned the value Z with the acceptance probability α, otherwise the next state of the Markov chain stays unchanged (i.e. retains the same value as before). More formally the algorithm is as follows.

Algorithm 1.2.1. Set $Y_0 = y_0$. For a given state Y_t, for some $t = 0, 1, \ldots$, the next state Y_{t+1} is sampled as follows

1. Sample Z from the proposal density $g(Y_t, \cdot)$.

2. Let

$$Y_{t+1} = \begin{cases} Z & \text{with probability } \alpha(Y_t, Z) \\ Y_t & \text{otherwise} \end{cases}$$

where $\alpha(y,z) = \min\{1, r(y,z)\}$ and $r(y,z) = \frac{\pi(z)g(z,y)}{\pi(y)g(y,z)}$.

This algorithm produces a Markov chain $(Y_t)_{t\geq 0}$ whose invariant density is given by f. Fore more details on the Metropolis-Hastings algorithm we refer to [2] and [17].

Another method of MCMC sampling is the Gibbs sampler, which was originally introduced by Geman and Geman in [16]. If the random variable X is multi-dimensional $X = (X_1, \ldots, X_d)$, the Gibbs sampler updates each component at the time by sampling from the conditional marginal distributions. Let $f_{k|\not k}(x_k \mid x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_d), k = 1, \ldots, d$, denote the conditional density of X_k given $X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_d$. The Gibbs sampler can be viewed as a special case of the Metropolis-Hastings algorithm where, given $Y_t = (Y_{t,1}, \ldots, Y_{t,d})$, one first updates $Y_{t,1}$ from the conditional density $f_{1|\not 1}(\cdot \mid Y_{t,2}, \ldots, Y_{t,d})$, then $Y_{t,2}$ from the conditional density $f_{2|\not 2}(\cdot \mid Y_{t+1,1}, Y_{t,3}, \ldots, Y_{t,d})$, and so on. Until at last one updates $Y_{t,d}$ from $f_{d|\not d}(\cdot \mid Y_{t+1,1}, \ldots, Y_{t+1,d-1})$. By sampling from these proposal densities the acceptance probability is always equal to 1, so no acceptance step is needed.

An important property of a Markov chain is its ergodicity. Informally, ergodicity measures the how quickly the Markov chain mixes and thus how soon the dependency of the chain dies out. This is a highly desired property since good mixing speeds up the convergence of the Markov chain.

1.3 Markov chain Monte Carlo in rare-event simulation

In this section we describe a new rare-event simulation methodology based on Markov chain Monte Carlo (MCMC). The new methodology results in a new estimator for computing probabilities of rare events, called the MCMC estimator. This technique is the main contribution of this thesis and the estimator is proved to be efficient in several different settings.

1.3.1 Formulation

Let X be a real-valued random variable with distribution F and density f with respect to the Lebesgue measure. The problem is to compute the probability

$$p = \mathbf{P}(X \in A) = \int_A dF. \tag{1.1}$$

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 $X \in A$. The density of F_A is given by

$$\frac{dF_A}{dx}(x) = \frac{f(x)I\{x \in A\}}{p}. \tag{1.2}$$

Consider a Markov chain $(X_t)_{t\geq 0}$ with invariant density given by (1.2). 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. [2, 17].

To construct an estimator for the normalising constant p, consider a non-negative function v, which is normalised in the sense that $\int_A v(x)dx = 1$. The function v will be chosen later as part of the design of the estimator. For any choice of v the sample mean,

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{v(X_t)I\{X_t \in A\}}{f(X_t)},$$

can be viewed as an estimate of

$$\mathbf{E}_{F_A}\left[\frac{v(X)I\{X \in A\}}{f(X)}\right] = \int_A \frac{v(x)}{f(x)}\frac{f(x)}{p}dx = \frac{1}{p}\int_A v(x)dx = \frac{1}{p}.$$

Thus,

$$\widehat{q}_T = \frac{1}{T}\sum_{t=0}^{T-1} u(X_t), \quad \text{where} \quad u(X_t) = \frac{v(X_t)I\{X_t \in A\}}{f(X_t)}, \tag{1.3}$$

is an unbiased estimator of $q = p^{-1}$. Then $\widehat{p}_T = \widehat{q}_T^{-1}$ is an estimator of 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. Another remark is that it is theoretically possible that all the terms in the sum in (1.3) are zero, leading to the estimate $\widehat{q}_T = 0$ and then $\widehat{p}_T = \infty$. To avoid such nonsense one can simply take $\widehat{p}_T$ as the minimum of $\widehat{q}_T^{-1}$ and one.

There are two essential design choices that determine the performance of the algorithm: the choice of the function v and the design of the MCMC sampler. The function v influences the variance of $u(X_t)$ in (1.3) and is therefore of main concern for controlling the rare-event properties of the algorithm. It is desirable to take v such that the normalised variance of the estimator, given by $p^2\,\mathrm{Var}(\widehat{q}_T)$, is not too large. The design of the MCMC sampler, on the other hand, 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 simulation it is desirable that the Markov chain mixes fast so that the dependence dies out quickly.

1.3.2 Controlling the normalised variance

This section contains a discussion on how to control the performance of the estimator $\widehat{q}_T$ by controlling its normalised variance.

For the estimator $\widehat{q}_T$ to be useful it is of course important that its variance is not too large. When the probability p to be estimated is small it is reasonable to ask that $\mathrm{Var}(\widehat{q}_T)$ is of size comparable to $q^2 = p^{-2}$, or equivalently, that the standard deviation of the estimator is roughly of the same size as p^{-1}. To this end the normalised variance $p^2\,\mathrm{Var}(\widehat{q}_T)$ is studied.

Let us consider $\mathrm{Var}(\widehat{q}_T)$. With

$$u(x) = \frac{v(x)I\{x \in A\}}{f(x)},$$

it follows that

$$\begin{aligned} p^2 \,\mathrm{Var}_{F_A}(\widehat{q}_T) &= p^2 \,\mathrm{Var}_{F_A}\Big(\frac{1}{T}\sum_{t=0}^{T-1} u(X_t)\Big) \\ &= p^2\Big(\frac{1}{T}\,\mathrm{Var}_{F_A}(u(X_0)) + \frac{2}{T^2}\sum_{t=0}^{T-1}\sum_{s=t+1}^{T-1} \mathrm{Cov}_{F_A}(u(X_s), u(X_t))\Big), \end{aligned} \tag{1.4}$$

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

$$\begin{aligned} \frac{p^2}{T}\,\mathrm{Var}_{F_A}\big(u(X_0)\big) &= \frac{p^2}{T}\Big(\mathbf{E}_{F_A}\big[u(X_0)^2\big] - \mathbf{E}_{F_A}\big[u(X_0)\big]^2\Big) \\ &= \frac{p^2}{T}\Big(\int\Big(\frac{v(x)}{f(x)}I\{x \in A\}\Big)^2 F_A(dx) - \frac{1}{p^2}\Big) \\ &= \frac{p^2}{T}\Big(\int \frac{v^2(x)}{f^2(x)}I\{x \in A\}\frac{f(x)}{p}dx - \frac{1}{p^2}\Big) \\ &= \frac{1}{T}\Big(\int_A \frac{v^2(x)p}{f(x)}dx - 1\Big). \end{aligned}$$

Therefore, in order to control the normalised variance the function v must be chosen so that $\int_A \frac{v^2(x)}{f(x)}dx$ is close to p^{-1}. An important observation is that the conditional density (1.2) plays a key role in finding a good choice of v. Letting v be the conditional density in (1.2) leads to

$$\int_A \frac{v^2(x)}{f(x)}dx = \int_A \frac{f^2(x)I\{x \in A\}}{p^2 f(x)}dx = \frac{1}{p^2}\int_A f(x)dx = \frac{1}{p},$$

which implies,

$$\frac{p^2}{T}\,\mathrm{Var}_{F_A}\big(u(X)\big) = 0.$$

This motivates taking v as an approximation of the conditional density (1.2). This is similar to the ideology behind choosing an efficient importance sampling estimator.

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

$$v(x) = \frac{f(x)I\{x \in B\}}{\mathbf{P}(X \in B)},$$

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

$$\int_A \frac{v^2(x)}{f(x)}dx = \int_A \frac{f^2(x)I\{x \in B\}}{\mathbf{P}(X \in B)^2 f(x)}dx = \frac{1}{\mathbf{P}(X \in B)^2}\int_B f(x)dx = \frac{1}{\mathbf{P}(X \in B)},$$

which will be close to p^{-1}.

Now, let us shift emphasis to the covariance term in (1.4). As the samples $(X_t)_{t=0}^{T-1}$ form a Markov chain the X_t's are dependent. Therefore the covariance term in (1.4) is non-zero and may not be ignored. The crude upper bound

$$\mathrm{Cov}_{F_A}(u(X_s), u(X_t)) \leq \mathrm{Var}_{F_A}(u(X_0)),$$

leads to the upper bound

$$\frac{2p^2}{T^2} \sum_{t=0}^{T-1} \sum_{s=t+1}^{T-1} \mathrm{Cov}_{F_A}(u(X_s), u(X_t)) \leq p^2 \Big(1 - \frac{1}{T}\Big) \mathrm{Var}_{F_A}(u(X_0))$$

for the covariance term. This is a very crude upper bound as it does not decay to zero as $T \to \infty$. But, at the moment, the emphasis is on small p so we will proceed with this upper bound anyway. As indicated above the choice of v controls the term $p^2 \mathrm{Var}_{F_A}(u(X_0))$. We conclude that the normalised variance (1.4) of the estimator $\widehat{q}_T$ is controlled by the choice of v when p is small.

1.3.3 Ergodic properties

As we have just seen the choice of the function v controls the normalised 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 analysed within the theory of φ-irreducible Markov chains. Fundamental results for φ-irreducible Markov chains are given in [26, 28]. We will focus on conditions that imply a geometric convergence rate. The conditions given below are well studied in the context of MCMC samplers. Conditions for geometric ergodicity in the context of Gibbs samplers have been studied by e.g. [9, 32, 33], and for Metropolis-Hastings algorithms by [24].

A Markov chain $(X_t)_{t\geq 0}$ with transition kernel $p(x, \cdot) = \mathbf{P}(X_{t+1} \in \cdot \mid X_t = x)$ is φ-irreducible if there exists a measure φ such that $\sum_t p^{(t)}(x, \cdot) \ll \varphi(\cdot)$, where $p^{(t)}(x, \cdot) = \mathbf{P}(X_t \in \cdot \mid X_0 = x)$ denotes the t-step transition kernel and $\ll$ denotes absolute continuity. 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

$$\|p^{(t)}(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \leq M(x) r^t, \tag{1.5}$$

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 (1.5) may be difficult to establish directly and are therefore substituted by suitable minorisation or drift conditions. A minorisation 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. Minorisation conditions have been used for obtaining rigorous bounds on the convergence of MCMC samplers, see e.g. [30].

If the entire state space is small, then the Markov chain is uniformly ergodic. Uniform ergodicity does typically not hold for Metropolis samplers, see Mengersen and Tweedie in [24] Theorem 3.1. Therefore useful sufficient conditions for geometric ergodicity are often given in the form of drift conditions [9, 24]. Drift conditions, established through the construction of appropriate Lyapunov functions, are also useful for establishing central limit theorems for MCMC algorithms, see [22, 26] and the references therein.

1.3.4 Efficiency of the MCMC algorithm

Roughly speaking, the arguments given above lead to the following desired properties of the estimator.

1. *Rare event efficiency:* Construct an unbiased estimator $\widehat{q}_T$ of p^{-1} according to (1.3) by finding a function v which approximates the conditional density (1.2). The choice of v controls the normalised variance of the estimator.

2. *Large sample efficiency:* Design the MCMC sampler, by finding an appropriate Gibbs sampler or a proposal density in the Metropolis-Hastings algorithm, such that the resulting Markov chain is geometrically ergodic.

1.4 Summary of papers

Paper 1: Markov chain Monte Carlo for computing rare-event probabilities for a heavy-tailed random walk

In this paper we provide the general framework of the Markov chain Monte Carlo simulation to effectively compute rare-event probabilities. The two design choices are highlighted, which determine the performance of the MCMC estimator. The first one is the design of the MCMC sampler which ensures the large sample efficiency and the second one is the choice of the distribution $\int v(x)dx$ for the v described earlier in Section 1.3.

The MCMC methodology is exemplified in the paper with two applications. The first application is the heavy-tailed random walk $S_n = Y_1 + \cdots + Y_n$ and the problem of computing

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

where $a_n \to \infty$ sufficiently fast so that the probability tends to zero and the distribution of the increments Y is assumed to be heavy-tailed. We present a Gibbs sampler which generates a Markov chain $(\mathbf{Y}_t^{(n)})_{t\geq 0}$, where $\mathbf{Y}_t^{(n)} = (Y_{t,1}, \ldots, Y_{t,n})$, whose invariant distribution is the conditional

$$F_{a_n}^{(n)} = \mathbf{P}\big((Y_1, \ldots, Y_n) \in \cdot \mid S_n/n > a_n\big).$$

The Markov chain is proved to preserve stationarity and to be uniformly ergodic.

For a given sample $(\mathbf{Y}_t^{(n)})_{t\geq 0}$ the MCMC estimator $\widehat{q}_T^{(n)}$ is by construction defined by

$$\widehat{q}_T^{(n)} = \frac{1}{T}\sum_{t=0}^{T-1} u(\mathbf{Y}_t), \quad u(\mathbf{Y}^{(n)}) = \frac{dV^{(n)}}{dF^{(n)}}(\mathbf{Y}^{(n)}),$$

where the distribution $V^{(n)}$ is defined as

$$V^{(n)}(\cdot) = \mathbf{P}\big((Y_1,\ldots,Y_n)\in\cdot \mid \max\{Y_1,\ldots,Y_n\} > a_n\big).$$

This choice of $V^{(n)}$ is motivated by the heavy-tail assumption of Y since that implies that $\mathbf{P}(S_n/n > a_n)/\mathbf{P}(\max\{Y_1,\ldots,Y_n\} > a_n) \to 1$ as $n\to\infty$, making $V^{(n)}$ a good asymptotic approximation of the zero-variance distribution $F_{a_n}^{(n)}$. The MCMC estimator is rewritten as

$$\widehat{q}_T^{(n)} = \mathbf{P}\big(\max\{Y_1,\ldots,Y_n\} > a_n\big)^{-1}\frac{1}{T}\sum_{t=0}^{T-1} I\Big\{\bigvee_{i=1}^{n} Y_{t,i} > a_n\Big\},$$

where the first factor can be interpreted as the asymptotic approximation of $1/p^{(n)}$ and the second factor is the stochastic correction term.

The ergodicity of the Markov chain produced by the Gibbs sampler implies that

$$\mathrm{Var}_{F_a^{(n)}}(q_T^{(n)}) \to 0, \quad \text{as } T\to\infty,$$

thus ensuring the estimator's large sample efficiency. Moreover, the main result Theorem 2.4.6 shows that the estimator has vanishing relative error and is thereby rare-event efficient. Numerical experiments are performed on the MCMC estimator and compared to importance sampling algorithms and standard Monte Carlo.

The second application is the heavy-tailed random sum $S_N = Y_1 + \cdots Y_N$ and the problem of computing

$$p^{(n)} = \mathbf{P}(Y_1 + \cdots + Y_{N_n} > a_n),$$

where N is a random variable and $a_n\to\infty$ sufficiently fast so that the probability tends to zero. The distribution of the increments Y is assumed to be heavy-tailed. The solution approach is similar to that of random walk save for the design of the Gibbs sampler and the choice of $V^{(n)}$. The Gibbs sampler takes into the account the new random number N and an extra step is inserted into the algorithm sampling N correctly such that the resulting Markov chain both preserves stationarity and is uniformly ergodic. The $V^{(n)}$ in this application is chosen to be

$$\mathbf{P}\big((N,Y_1,\ldots,Y_N)\in\cdot \mid \max\{Y_1,\ldots,Y_N\} > a_n\big).$$

The ergodicity of the Markov chain again ensure that the variance of the estimator tends to zero as sample size increase. By Theorem 2.4.11 the estimator has vanishing relative error. Numerical experiments are performed and compared to that of importance sampling and standard Monte Carlo. From viewing the numerical results we draw the conclusion that for small probabilities, around 10^{-2} and smaller, the MCMC estimator outperforms the importance sampling estimator.

Paper 2: Markov chain Monte Carlo for rare-event simulation for stochastic recurrence equations with heavy-tailed innovations

This paper continues the development of rare-event simulation techniques based on MCMC sampling. Motivated by the positive results from Paper 1, we investigate the MCMC methodology for broader set of models. Although, the application in this paper restrict itself to the heavy-tailed setting as in the previous paper.

The MCMC methodology is developed to the setting of solutions to stochastic recurrence equations with heavy-tailed innovations. The MCMC methodology is applied to the problem of computing $p^{(n)} = \mathbf{P}(X_m > a_n)$, where

$$\begin{aligned} X_k &= A_k X_{k-1} + B_k, \quad \text{for } k = 1, \dots, m, \\ X_0 &= 0, \end{aligned}$$

and $a_n \to \infty$ as $n \to \infty$. The tail of the increments B are assumed to be regularly varying of index $-\alpha < 0$ and $\mathbf{E}[A^{\alpha+\epsilon}] < \infty$ for some $\epsilon > 0$. We present a Gibbs sampler which generates a Markov chain $(\mathbf{A}_t, \mathbf{B}_t)_{t\geq 0}$, where $(\mathbf{A}_t, \mathbf{B}_t) = (A_{t,1}, \dots, A_{t,m}, B_{t,1}, \dots, B_{t,m})$, whose invariant distribution is the conditional distribution

$$F_{a_n}^{(n)} = \mathbf{P}\big((A_2, \dots, A_m, B_1, \dots, B_m) \in \cdot \mid X_m > a_n\big).$$

The sampler updates the variables sequentially, ensuring in each step that the updating preserves $X_m > a_n$. The Markov chain is shown to be stationary and uniformly ergodic.

Motivated by the heavy-tail assumption on the distribution of the innovations B and the light-tail behaviour of A we use the following as an asymptotic approximation of $\{X_m > a_n\}$,

$$R^{(n)} = \bigcup_{k=1}^{m} R_k^{(n)}, \quad \text{where} \quad R_k^{(n)} = \{A_m \cdots A_{k+1} B_k > c_n, A_m, \dots, A_{k+1} > a\}.$$

The probability of this event $r^{(n)} = \mathbf{P}\big((\mathbf{A}, \mathbf{B}) \in R^{(n)}\big)$ can be computed explicitly using the inclusion-exclusion formula and for a given sample from the Markov chain $(\mathbf{A}_0, \mathbf{B}_0), \dots, (\mathbf{A}_{T-1}, \mathbf{B}_{T-1})$, the MCMC estimator is given by

$$\widehat{q}_T^{(n)} = (r^{(n)})^{-1} \frac{1}{T} \sum_{t=0}^{T-1} I\big\{(\mathbf{A}_t, \mathbf{B}_t) \in R^{(n)}\big\}.$$

The interpretation is again that the first term is the asymptotic approximation of the probability $1/p^{(n)}$ and the second term is the stochastic correction factor.

The ergodicity of the Markov chain ensures that the variance of the estimator tends to zero as sample size increases and the main result, Theorem 3.3.5, shows that the MCMC estimator has vanishing relative error as n tends to infinity. Numerical experiments are performed on the MCMC estimator and compared to existing importance sampling algorithms for the problem. The numerical performance of the MCMC estimator is reasonable but displays some problem with the burn-in of the MCMC sampling.

Moreover, this paper presents an example in the context of the capital requirements for an insurance company with risky investments. The company's reserves are modelled using a stochastic recurrence equations where the As are interpreted as the random return on the company's capital and the Bs are the claim amounts. The discounted loss process is described in terms of A and B and the MCMC methodology is applied to the problem of computing the probability of ruin.

Paper 3: Markov chain Monte Carlo for rare-event simulation for light-tailed random walk

Motivated by the success of designing effective MCMC algorithms for computing rare-event probabilities in the heavy-tailed context we considered how the methodology can be extended to cover the light-tailed setting. In the light-tailed setting we rely on logarithmic large deviations and therefore emphasis is on logarithmic efficiency rather than strong efficiency.

The application considered in this paper is a light-tailed random walk $S_n = Y_1 + \cdots + Y_n$ and the problem of efficiently computing

$$p^{(n)} = \mathbf{P}(S_n/n > a),$$

as $n \to \infty$. We consider two cases in this paper; when the increments are supported on $\mathbb{R}\hat{\text{E}}$and when they are supported only on R_+. The distribution of the increments Y is assumed to be light-tailed such that $\Lambda(\theta) = \log \mathbf{E}[e^{\theta Y}] < \infty$ for some $\theta > 0$ which implies the existence of a large deviation principle. In particular we assume that

$$-\lim_{n\to\infty} \frac{1}{n} \log p^{(n)} = I(a),$$

where I is the Fenchel-Legendre transform of the Λ, namely $I(a) = \sup_{\theta\in\mathbb{R}}\{\theta a - \Lambda(\theta)\}$. We identify the zero-variance distribution as the conditional distribution given by

$$F_a^{(n)}(\cdot) = \mathbf{P}\big((Y_1, \ldots, Y_n) \in \cdot \mid S_n/n > a\big),$$

and we present a Gibbs sampler that generates a Markov chain $(\mathbf{Y}_t^{(n)})_{t\geq 0}$, where $\mathbf{Y}_t^{(n)} = (Y_{t,1}, \ldots, Y_{t,n})$, whose invariant distribution is $F_a^{(n)}$. The Markov chain is shown to preserve stationarity and be uniformly ergodic.

The MCMC estimator is defined in terms of a distribution $V^{(n)}$ as follows,

$$\widehat{q}_T^{(n)} = \frac{1}{T}\sum_{t=0}^{T-1} u(\mathbf{Y}_t^{(n)}), \quad u(\mathbf{Y}^{(n)}) = \frac{dV^{(n)}}{dF^{(n)}}(\mathbf{Y}^{(n)}).$$

As motivated in Section 1.3 the distribution $V^{(n)}$ is chosen as an asymptotic approximation of $F_a^{(n)}$. For the case when the support of the increments is $\mathbb{R}$ then we define

$$V^{(n)}(\cdot) = \mathbf{P}\big((Z_1, \ldots, Z_n) \in \cdot \mid (Z_1 + \cdots + Z_n)/n > a\big),$$

where the Z's are independent and identically normal distributed variables with mean μ and standard deviation σ. The normalising constant of $V^{(n)}$ can be computed explicitly

$$r^{(n)} = \mathbf{P}\big((Z_1 + \cdots + Z_n)/n > a\big) = 1 - \Phi\Big(\frac{a-\mu}{\sigma/\sqrt{n}}\Big),$$

where Φ is the standard normal probability distribution. The parameters are set such that $\mu = \mathbf{E}[Y]$ and σ such that

$$-\lim_{n\to\infty} \frac{1}{n} \log r^{(n)} = I(a),$$

that is, the rate function of $\{Z_i\}$ coincides with that of $\{Y_i\}$. The MCMC estimator is rewritten as

$$\widehat{q}_T^{(n)} = (r^{(n)})^{-1} \frac{1}{T} \sum_{t=0}^{T-1} \prod_{i=1}^{n} \frac{\phi_{\mu,\sigma}(Y_{t,i})}{f_Y(Y_{t,i})} I\Big\{\frac{1}{n}\sum_{i=1}^{n} Y_{t,i} > a\Big\}.$$

For the case when the support of the increments is $\mathbb{R}_+$ we follow a similar procedure save the Z's are chosen as gamma distributed because the normalising constant can be computed explicitly since the sum of gamma distributed variables is itself gamma distributed.

The ergodicity of the Markov chain ensures that the variance of the MCMC estimator tends to zero as $T \to \infty$ and the main results, Theorem 4.3.4 and Theorem 4.3.5, characterise the efficiency of the MCMC estimator, giving exact which conditions need to be fullfilled to ensure logarithmic efficiency. Numerical experiments are performed and compare the output of the MCMC estimator to importance sampling estimator and Monte Carlo. The MCMC estimator performs comparably with the strongly efficient importance sampling algorithm, both for the $\mathbb{R}$ support case and the $\mathbb{R}_+$ support case.

Paper 4: Markov chain Monte Carlo for rare-event simulation for Markov chains

In this paper we continue the study of the MCMC methodology in the light-tailed setting. The models considered in this paper are discrete-time Markov chains and continuous-time Markov chains.

In the first part of the paper we consider the implementation of the MCMC methodology for a discrete-time Markov chain $(X_i^{(n)})_{i\geq 0}$ with a given initial value $X_0^{(n)} = x_0$. The problem considered is the computation of expectations on the form

$$\theta^{(n)} = \mathbf{E}[e^{-nh_0(X^{(n)})}],$$

where $X^{(n)}$ is the linearly interpolated version of $(X_i^{(n)})_{i=0}^n$ and h_0 is a bounded continuous mapping $\mathscr{C}([0,1];\mathbb{R}^d) \to \mathbb{R}$. We identify the zero-variance distribution given by

$$F_{h_0}^{(n)}(\cdot) = \frac{\mathbf{E}[I\{X^{(n)} \in \cdot\} e^{-nh_0(X^{(n)})}]}{\theta^{(n)}},$$

and present a Metropolis-Hastings algorithm which generates a $\mathscr{C}([0,1];\mathbb{R}^d)$-valued Markov chain whose invariant distribution is $F_{h_0}^{(n)}$.

Under certain conditions it follows that the Markov chain $X^{(n)}$ satisfies a Laplace principle which is used to define the MCMC estimator. By construction of the Metropolis-Hastings algorithm the estimator is large sample efficient and Theorem 5.3.2 characterise the efficiency of the estimator.

In the second part of the paper a similar analysis is completed for continuous-time Markov chain. A Metropolis-Hastings algorithm is designed such that it constructs a Markov chain having the zero-variance distributions as its invariant distribution. Then an MCMC estimator is defined and its efficiency is characterised.

The paper is conclude with the example of a birth-and-death process in continuous-time and spatial dependent intensities. We consider the first passage problem

$$p^{(n)} = \mathbf{P}(\sup_{t\in[0,T]} X^{(n)}(t) > a),$$

as $n \to \infty$. Numerical experiments are performed on an example of birth-and-death process and compared to standard Monte Carlo estimates.

1.5 References

[1] S. Asmussen and K. Binswanger. Simulation of ruin probabilities for subexponential claims. *Astin Bulletin*, 27(2):297–318, 1997.

[2] S. Asmussen and P. W. Glynn. *Stochastic Simulation*, volume 57 of *Stochastic Modelling and Applied Probability*. Springer, New York, 2007.

[3] S. Asmussen and D. P. Kroese. Improved algorithms for rare event simulation with heavy tails. *Adv. Appl. Probab.*, 38:545–558, 2006.

[4] J. Blanchet and P. W. Glynn. Efficient rare-event simulation for the maximum of a heavy-tailed random walk. *Ann. of Appl. Prob.*, 18(4):1351–1378, 2008.

[5] J. Blanchet and P.W. Glynn. Efficient rare event simulation for the single server queue with heavy tailed distributions. *Ann. Appl. Prob.*, 18(4):1351–1378, 2008.

[6] J. Blanchet, P.W. Glynn, and J. Liu. Fluid heuristics, Lyapunov bounds and efficient importance sampling for a heavy-tailed G/G/1 queue. *Queueing Syst.*, 56(3):99–113, 2007.

[7] J. Blanchet and C. Li. Efficient rare-event simulation for heavy-tailed compound sums. *ACM T. Model Comput. S.*, 21:1–10, 2011.

[8] J. Blanchet and J. C. Liu. State-dependent importance sampling for regularly varying random walks. *Adv. Appl. Probab.*, 40:1104–1128, 2008.

[9] K. J. Chan. Asymptotic behavior of the Gibbs sampler. *J. Am. Stat. Assoc.*, 88:320–326, 1993.

[10] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications, 2nd Ed.* Springer, New York, 1998.

[11] D. Denisov, A. B. Dieker, and V. Shneer. Large deviations for random walks under subexponentiality: The big-jump domain. *Ann. Probab.*, 36:1946–1991, 2008.

[12] P. Dupuis, K. Leder, and H. Wang. Importance sampling for sums of random variables with regularly varying tails. *ACM T. Model Comput. S.*, 17(3), 2007.

[13] P. Dupuis, A. Sezer, and H. Wang. Dynamic importance sampling for queueing networks. *Ann. Appl. Probab*, 17(4):1306–1346, 2007.

[14] P. Dupuis and H. Wang. Importance sampling, large deviations, and differential games. *Stoch. Stoch. Rep.*, 76(6):481–508, 2004.

[15] P. Dupuis and H. Wang. Subsolutions of an Isaacs equation and efficient schemes for importance sampling. *Math. Oper. Res.*, 32(3):1–35, 2007.

[16] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6:721–741, 1984.

[17] W. R. Gilks, S. Richardsson, and D. J. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman & Hall, 1996.

[18] W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[19] H. Hult and F. Lindskog. Extremal behavior of regularly varying stochastic processes. *Stoch. Proc. Appl.*, 115(2):249–274, 2005.

[20] H. Hult, F. Lindskog, O. Hammarlid, and C.J. Rehn. *Risk and Portfolio Analysis*. Springer, New York, 2012.

[21] H. Hult, F. Lindskog, T. Mikosch, and G. Samorodnitsky. Functional large deviations for multivariate regularly varying random walks. *Ann. Appl. Probab.*, 14(4):2651–2680, 2005.

[22] G. L. Jones. On the Markov chain central limit theorem. *Probability Surveys*, 1:299–320, 2004.

[23] S. Juneja and P. Shahabuddin. Simulation heavy-tailed processes using delayed hazard rate twisting. *ACM T. Model Comput. S.*, 12(2):94–118, April 2002.

[24] K. L. Mengersen and R. L. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Stat.*, 24:101–121, 1996.

[25] N. Metropolis, A.W. Rosebluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.* , 21, 1953.

[26] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, New York, 1993.

[27] T. Mikosch. *Non-life Insurance Mathematics*. Springer, New York, 2009.

[28] E. Nummelin. *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge University Press, 1984.

[29] B. D. Ripley. *Stochastic Simulation*. Wiley Series in Probability and Statistics, New Jersey, 1987.

[30] J. S. Rosenthal. Minorization conditions and convergence rates for markov chain monte carlo. *J. Am. Stat. Assoc.*, 90(430):558–566, June 1995.

[31] D. Siegmund. Importance sampling in the monte carlo study of sequential tests. *Ann. Statist.*, 4(4):673–684, 1976.

[32] A. F. M. Smith and A. E. Gelfand. Bayesian statistics without tears: a sampling-resampling perspective. *Am. Stat.*, 46:84–88, 1992.

[33] L. Tierney. Markov chains for exploring posterior distributions (with discussion). *Ann. Stat.*, 22:1701–1762, 1994.

Markov chain Monte Carlo for computing rare-event probabilities for a heavy-tailed random walk

Markov chain Monte Carlo for computing rare-event probabilities for a heavy-tailed random walk

by

Thorbjörn Gudmundsson and Henrik Hult

Condensed version published in
Journal of Applied Probability, Vol. 61, No. 2, June 2014

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 normalising constant. Using the MCMC methodology a Markov chain is simulated, with that conditional distribution as its invariant distribution, and information about the normalising 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 normalised variance vanishes asymptotically. The algorithm is extended to random sums and its performance is illustrated numerically and compared to existing importance sampling algorithms.

2.1 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 the event of interest and then extract the probability of the event as the normalising 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 $X^{(1)}, X^{(2)}, \ldots$, possibly multidimensional, each of which can be sampled repeatedly by a simulation algorithm. The objective is to estimate $p^{(n)} = \mathbf{P}(X^{(n)} \in A)$, for some large n, based on a sample $X_0^{(n)}, \ldots, X_{T-1}^{(n)}$. It is assumed that

the probability $\mathbf{P}(X^{(n)} \in A) \to 0$, as $n \to \infty$, so that the event $\{X^{(n)} \in A\}$ 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 $\widehat{p}_T^{(n)}$ of $p^{(n)}$ a useful performance measure is the relative error:

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

An algorithm is said to have *vanishing relative error* if the relative error tends to zero 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 algorithm is inefficient for computing rare-event probabilities. As an illustration, consider the standard Monte Carlo estimate

$$\widehat{p}_T^{(n)} = \frac{1}{T} \sum_{t=0}^{T-1} I\{X_t^{(n)} \in A\},$$

of $p^{(n)} = \mathbf{P}(X^{(n)} \in A)$ based on independent replicates $X_0^{(n)}, \ldots, X_{T-1}^{(n)}$. The relative error of the Monte Carlo estimator is

$$\frac{\mathrm{Var}(\widehat{p}_T^{(n)})}{(p^{(n)})^2} = \frac{p^{(n)}(1-p^{(n)})}{T(p^{(n)})^2} = \frac{1}{Tp^{(n)}} - \frac{1}{T} \to \infty,$$

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 a weighted empirical estimator,

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

where $L^{(n)} = dF^{(n)}/dG^{(n)}$ is the likelihood ratio, which is assumed to exist on A. The importance sampling estimator $\widehat{p}_T^{(n)}$ 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,

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

In this case the likelihood ratio weights $L^{(n)}$ are equal to $p^{(n)}$ which implies that $\widehat{p}_T^{(n)}$ 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 $\{X^{(n)} \in A\}$ 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 $F_A^{(n)}$. Because $F_A^{(n)}$ is known up to the normalising constant $p^{(n)}$ it is possible to sample from $F_A^{(n)}$ using an MCMC algorithm such as a Gibbs sampler or Metropolis-Hastings algorithm. The idea is to generate samples $X_0^{(n)}, \dots, X_{T-1}^{(n)}$ from a Markov chain with stationary distribution $F_A^{(n)}$ and construct an estimator of the normalising constant $p^{(n)}$. An unbiased estimator of $(p^{(n)})^{-1}$ is constructed from a known probability density $v^{(n)}$ on A and the original density $f^{(n)}$ of $X^{(n)}$ by

$$\widehat{q}_T^{(n)} = \frac{1}{T} \sum_{t=0}^{T-1} \frac{v^{(n)}(X_t^{(n)}) I\{X_t^{(n)} \in A\}}{f^{(n)}(X_t^{(n)})}. \tag{2.1}$$

The performance of the estimator depends both on the choice of the density $v^{(n)}$ 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 normalising constants and ratios of normalising 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 the MCMC methodology we consider the problem of computing the probability that a random walk $S_n = Y_1 + \cdots + Y_n$, where $Y_1, \dots, Y_n$ are nonnegative, independent, and heavy-tailed random variables, exceeds a high threshold a_n, as number of summands n increases. This problem has received some attention in the context of conditional Monte Carlo algorithms [2, 4] and importance sampling algorithms [15, 10, 6, 5], most notably in the setting of fixed number of summands.

In this paper a Gibbs sampler is presented for sampling from the conditional distribution $\mathbf{P}((Y_1, \dots, Y_n) \in \cdot \mid S_n > a_n)$. The resulting Markov chain is proved to be uniformly ergodic. An estimator for $(p^{(n)})^{-1}$ of the form (2.1) is suggested with $v^{(n)}$ as the conditional density of $(Y_1, \dots, Y_n)$ given $\max\{Y_1, \dots, Y_n\} > a_n$. The estimator is proved to have vanishing normalised 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 [10, 6, 5]. An extension of the algorithm to a sum with a random number of steps is also presented.

Here follows an outline of the paper. The basic methodology and a heuristic efficiency analysis for computing rare-event probabilities is described in Section 2.2. The general

formulation for computing expectations is given in Section 2.3 along with a precise formulation of the efficiency criteria. Section 2.4 contains the design and efficiency results for the estimator for computing hitting probabilities for a heavy-tailed random walk, with deterministic and random number of steps. Section 2.5 presents numerical experiments and compares the efficiency of the MCMC estimator against an existing importance sampling algorithm and standard Monte Carlo. The MCMC estimator has strikingly better performance than existing importance sampling algorithms.

2.2 Computing rare-event probabilities by Markov chain Monte Carlo

In this section an algorithm for computing rare-event probabilities using Markov chain Monte Carlo (MCMC) is presented and conditions that ensure good convergence are discussed in a heuristic fashion. A more general version of the algorithm, for computing expectations, is provided in Section 2.3 along with a precise asymptotic efficiency criteria.

2.2.1 Formulation of the algorithm

Let X be a real-valued random variable with distribution F and density f with respect to the Lebesgue measure. The problem is to compute the probability

$$p = \mathbf{P}(X \in A) = \int_A dF. \tag{2.2}$$

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 $X \in A$. The density of F_A is given by

$$\frac{dF_A}{dx}(x) = \frac{f(x)I\{x \in A\}}{p}. \tag{2.3}$$

Consider a Markov chain $(X_t)_{t\geq 0}$ density is given by (2.3). 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, 12].

To construct an estimator for the normalising constant p, consider a non-negative function v, which is normalised in the sense that $\int_A v(x)dx = 1$. The function v will be chosen later as part of the design of the estimator. For any choice of v the sample mean,

$$\frac{1}{T}\sum_{t=0}^{T-1} \frac{v(X_t)I\{X_t \in A\}}{f(X_t)},$$

can be viewed as an estimate of

$$\mathbf{E}_{F_A}\left[\frac{v(X)I\{X \in A\}}{f(X)}\right] = \int_A \frac{v(x)}{f(x)}\frac{f(x)}{p}dx = \frac{1}{p}\int_A v(x)dx = \frac{1}{p}.$$

Thus,

$$\widehat{q}_T = \frac{1}{T}\sum_{t=0}^{T-1} u(X_t), \quad \text{where} \quad u(X_t) = \frac{v(X_t)I\{X_t \in A\}}{f(X_t)}, \tag{2.4}$$

is an unbiased estimator of $q = p^{-1}$. Then $\widehat{p}_T = \widehat{q}_T^{-1}$ is an estimator of 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. Another remark is that it is theoretically possible that all the terms in the sum in (2.4) are zero, leading to the estimate $\widehat{q}_T = 0$ and then $\widehat{p}_T = \infty$. To avoid such nonsense one can simply take $\widehat{p}_T$ as the minimum of $\widehat{q}_T^{-1}$ and one.

There are two essential design choices that determine the performance of the algorithm: the choice of the function v and the design of the MCMC sampler. The function v influences the variance of $u(X_t)$ in (2.4) and is therefore of main concern for controlling the rare-event properties of the algorithm. It is desirable to take v such that the normalised variance of the estimator, given by $p^2 \operatorname{Var}(\widehat{q}_T)$, is not too large. The design of the MCMC sampler, on the other hand, 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 simulation it is desirable that the Markov chain mixes fast so that the dependence dies out quickly.

2.2.2 Controlling the normalised variance

This section contains a discussion on how to control the performance of the estimator $\widehat{q}_T$ by controlling its normalised variance.

For the estimator $\widehat{q}_T$ to be useful it is of course important that its variance is not too large. When the probability p to be estimated is small it is reasonable to ask that $\operatorname{Var}(\widehat{q}_T)$ is of size comparable to $q^2 = p^{-2}$, or equivalently, that the standard deviation of the estimator is roughly of the same size as p^{-1}. To this end the normalised variance $p^2 \operatorname{Var}(\widehat{q}_T)$ is studied.

Let us consider $\operatorname{Var}(\widehat{q}_T)$. With

$$u(x) = \frac{v(x)I\{x \in A\}}{f(x)},$$

it follows that

$$\begin{aligned} p^2 \operatorname{Var}_{F_A}(\widehat{q}_T) &= p^2 \operatorname{Var}_{F_A}\Big(\frac{1}{T}\sum_{t=0}^{T-1} u(X_t)\Big) \\ &= p^2\Big(\frac{1}{T}\operatorname{Var}_{F_A}(u(X_0)) + \frac{2}{T^2}\sum_{t=0}^{T-1}\sum_{s=t+1}^{T-1} \operatorname{Cov}_{F_A}(u(X_s), u(X_t))\Big). \end{aligned} \tag{2.5}$$

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

$$\begin{aligned}\frac{p^2}{T}\operatorname{Var}_{F_A}\big(u(X_0)\big) &= \frac{p^2}{T}\Big(\mathbf{E}_{F_A}\big[u(X_0)^2\big]-\mathbf{E}_{F_A}\big[u(X_0)\big]^2\Big)\\ &= \frac{p^2}{T}\Big(\int\Big(\frac{v(x)}{f(x)}I\{x\in A\}\Big)^2F_A(dx)-\frac{1}{p^2}\Big)\\ &= \frac{p^2}{T}\Big(\int\frac{v^2(x)}{f^2(x)}I\{x\in A\}\frac{f(x)}{p}dx-\frac{1}{p^2}\Big)\\ &= \frac{1}{T}\Big(\int_A\frac{v^2(x)p}{f(x)}dx-1\Big).\end{aligned}$$

Therefore, in order to control the normalised variance the function v must be chosen so that $\int_A \frac{v^2(x)}{f(x)}dx$ is close to p^{-1}. An important observation is that the conditional density (2.3) plays a key role in finding a good choice of v. Letting v be the conditional density in (2.3) leads to

$$\int_A\frac{v^2(x)}{f(x)}dx=\int_A\frac{f^2(x)I\{x\in A\}}{p^2f(x)}dx=\frac{1}{p^2}\int_A f(x)dx=\frac{1}{p},$$

which implies,

$$\frac{p^2}{T}\operatorname{Var}_{F_A}\big(u(X)\big)=0.$$

This motivates taking v as an approximation of the conditional density (2.3). This is similar to the ideology behind choosing an efficient importance sampling estimator.

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

$$v(x)=\frac{f(x)I\{x\in B\}}{\mathbf{P}(X\in B)};$$

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

$$\int_A\frac{v^2(x)}{f(x)}dx=\int_A\frac{f^2(x)I\{x\in B\}}{\mathbf{P}(X\in B)^2f(x)}dx=\frac{1}{\mathbf{P}(X\in B)^2}\int_B f(x)dx=\frac{1}{\mathbf{P}(X\in B)},$$

which will be close to p^{-1}.

Now, let us shift emphasis to the covariance term in (2.5). As the samples $(X_t)_{t=0}^{T-1}$ form a Markov chain the X_t's are dependent. Therefore the covariance term in (2.5) is non-zero 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

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

for the covariance term. This is a very crude upper bound as it does not decay to zero as $T \to \infty$. But, at the moment, the emphasis is on small p so we will proceed with this upper bound anyway. As indicated above the choice of v controls the term $p^2 \operatorname{Var}_{F_A}(u(X_0))$. We conclude that the normalised variance (2.5) of the estimator $\widehat{q}_T$ is controlled by the choice of v when p is small.

2.2.3 Ergodic properties

As we have just seen the choice of the function v controls the normalised 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 analysed within the theory of φ-irreducible Markov chains. Fundamental results for φ-irreducible Markov chains are given in [18, 19]. We will focus on conditions that imply a geometric convergence rate. The conditions given below are well studied in the context of MCMC samplers. Conditions for geometric ergodicity in the context of Gibbs samplers have been studied by e.g. [7, 21, 22], and for Metropolis-Hastings algorithms by [17].

A Markov chain $(X_t)_{t\geq 0}$ with transition kernel $p(x,\cdot) = \mathbf{P}(X_{t+1} \in \cdot \mid X_t = x)$ is φ-irreducible if there exists a measure φ such that $\sum_t p^{(t)}(x,\cdot) \ll \varphi(\cdot)$, where $p^{(t)}(x,\cdot) = \mathbf{P}(X_t \in \cdot \mid X_0 = x)$ denotes the t-step transition kernel and $\ll$ denotes absolute continuity. 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

$$\|p^{(t)}(x,\cdot) - \pi(\cdot)\|_{\mathrm{TV}} \leq M(x) r^t, \tag{2.6}$$

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 (2.6) may be difficult to establish directly and are therefore substituted by suitable minorisation or drift conditions. A minorisation 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. Minorisation conditions have been used for obtaining 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. Uniform ergodicity does typically not hold for Metropolis samplers, [17] Theorem 3.1. Therefore useful sufficient conditions for geometric ergodicity are often given in the form of drift conditions [7, 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 2.4, we will encounter Gibbs samplers that are uniformly ergodic.

2.2.4 Heuristic efficiency criteria

To summarise, the heuristic arguments given above lead to the following desired properties of the estimator.

1. *Rare event efficiency:* Construct an unbiased estimator $\widehat{q}_T$ of p^{-1} according to (2.4) by finding a function v which approximates the conditional density (2.3). The choice of v controls the normalised variance of the estimator.

2. *Large sample efficiency:* Design the MCMC sampler, by finding an appropriate Gibbs sampler or a proposal density in the Metropolis-Hastings algorithm, such that the resulting Markov chain is geometrically ergodic.

2.3 The general formulation of the algorithm

In the previous section an estimator, based on Markov chain Monte Carlo, was introduced for computing the probability of a rare event. In this section the same ideas are applied to the problem of computing an expectation. Here the setting is somewhat more general. For instance, there is no assumption that densities with respect to Lebesgue measure exist.

Let X be a random variable with distribution F and h be a non-negative F-integrable function. The problem is to compute the expectation

$$\theta = \mathbf{E}\big[h(X)\big] = \int h(x)dF(x).$$

In the special case when F has density f and $h(x) = I\{x \in A\}$ this problem reduces to computing the probability in (2.2).

The analogue of the conditional distribution in (2.3) is the distribution F_h given by

$$F_h(B) = \frac{1}{\theta}\int_B h(x)dF(x), \quad \text{for measurable sets } B.$$

Consider a Markov chain $(X_t)_{t\geq 0}$ having F_h as its invariant distribution. To define an estimator of θ^{-1}, consider a probability distribution V with $V \ll F_h$. Then it follows that $V \ll F$ and it is assumed that the density dV/dF is known. Consider the estimator of $\zeta = \theta^{-1}$ given by

$$\widehat{\zeta}_T = \frac{1}{T}\sum_{t=0}^{T-1} u(X_t), \quad \text{where} \quad u(x) = \frac{1}{\theta}\frac{dV}{dF_h}(x). \tag{2.7}$$

Note that u does not depend on θ because $V \ll F_h$ and therefore

$$u(x) = \frac{1}{\theta}\frac{dV}{dF_h}(x) = \frac{1}{h(x)}\frac{dV}{dF}(x),$$

for x such that $h(x) > 0$. The estimator (2.7) is a generalisation of the estimator (2.4) where one can think of v as the density of V with respect to Lebesgue measure. An estimator of θ can then constructed as $\widehat{\theta}_T = \widehat{\zeta}_T^{-1}$.

The variance analysis of $\widehat{\zeta}_T$ follows precisely the steps outlined above in Section 2.2. The normalised variance is

$$\theta^2 \operatorname{Var}_{F_h}(\widehat{\zeta}_T) = \frac{\theta^2}{T} \operatorname{Var}_{F_h}\big(u(X_0)\big) + \frac{2\theta^2}{T^2} \sum_{t=0}^{T-1} \sum_{s=t+1}^{T-1} \operatorname{Cov}_{F_h}\big(u(X_s), u(X_t)\big), \quad (2.8)$$

where the first term can be rewritten, similarly to the display below (2.5), as

$$\frac{\theta^2}{T} \operatorname{Var}_{F_h}\big(u(X_0)\big) = \frac{1}{T}\Big(\mathbf{E}_V\Big[\frac{dV}{dF_h}\Big] - 1\Big).$$

The analysis above indicates that an appropriate choice of V is such that $\mathbf{E}_V[\frac{dV}{dF_h}]$ is close to 1. Again, the ideal choice would be taking $V = F_h$ leading to zero variance. This choice is not feasible but nevertheless suggests selecting V as an approximation of F_h. As already noted this is similar to the ideology behind choosing an efficient importance sampling estimator. The difference being that here $V \ll F$ is required whereas in importance sampling F needs be absolutely continuous with respect to the sampling distribution. The crude upper bound for the covariance term in (2.8) is valid, just as in Section 2.2.

2.3.1 Asymptotic efficiency criteria

Asymptotic efficiency can be conveniently formulated in terms of a limit criteria as a large deviation parameter tends to infinity. As is customary 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 formalised as follows.

Let $(X^{(n)})_{n\geq 1}$ be a sequence of random variables with $X^{(n)}$ having distribution $F^{(n)}$. Let h be a non-negative function, integrable with respect to $F^{(n)}$, for each n. Suppose

$$\theta^{(n)} = \mathbf{E}\big[h(X^{(n)})\big] = \int h(x) dF^{(n)}(x) \to 0,$$

as $n \to \infty$. The problem is to compute $\theta^{(n)}$ for some large n.

Denote by $F_h^{(n)}$ the distribution with $dF_h^{(n)}/dF^{(n)} = h/\theta^{(n)}$. For the nth problem, a Markov chain $(X_t^{(n)})_{t=0}^{T-1}$ with invariant distribution $F_h^{(n)}$ is generated by an MCMC algorithm. The estimator of $\zeta^{(n)} = (\theta^{(n)})^{-1}$ is based on a probability distribution $V^{(n)}$, such that $V^{(n)} \ll F_h^{(n)}$, with known density with respect to $F^{(n)}$. An estimator $\widehat{\zeta}_T^{(n)}$ of ζ is given by

$$\widehat{\zeta}_T^{(n)} = \frac{1}{T} \sum_{t=0}^{T-1} u^{(n)}(X_t^{(n)}),$$

where

$$u^{(n)}(x) = \frac{1}{h(x)} \frac{dV^{(n)}}{dF^{(n)}}(x).$$

The heuristic efficiency criteria in Sections 2.2 can now be rigorously formulated as follows:

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

$$(\theta^{(n)})^2 \operatorname{Var}_{F_h^{(n)}}(u^{(n)}(X)) \to 0, \text{ as } n \to \infty.$$

2. *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 2.3.1. The rare-event efficiency criteria is formulated in terms of the efficiency of estimating $(\theta^{(n)})^{-1}$ by $\widehat{\zeta}_T^{(n)}$. If one insists on studying the mean and variance of $\widehat{\theta}_T^{(n)} = (\widehat{\zeta}_T^{(n)})^{-1}$, then the effects of the transformation $x \mapsto x^{-1}$ must be taken into account. For instance, the estimator $\widehat{\theta}_T^{(n)}$ 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 $(\theta^{(n)})^{-1}$ by $\widehat{\zeta}_T^{(n)}$ there is a bias coming from the fact that the Markov chain not being perfectly stationary.

2.4 A random walk with heavy-tailed steps

In this section the estimator introduced in Section 2.2 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 and density f_Y with respect to some reference measure μ. Consider the random walk $S_n = Y_1 + \cdots + Y_n$ and the problem of computing the probability

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

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

It is convenient to denote by $\mathbf{Y}^{(n)}$ the n-dimensional random vector

$$\mathbf{Y}^{(n)} = (Y_1, \ldots, Y_n)^{\mathrm{T}},$$

and the set

$$A_n = \{\mathbf{y} \in \mathbb{R}^n : \mathbf{1}^{\mathrm{T}}\mathbf{y} > a_n\},$$

where $\mathbf{1} = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^n$ and $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$. With this notation

$$p^{(n)} = \mathbf{P}(S_n > a_n) = \mathbf{P}(\mathbf{1}^{\mathrm{T}}\mathbf{Y}^{(n)} > a_n) = \mathbf{P}(\mathbf{Y}^{(n)} \in A_n).$$

The conditional distribution

$$F_{A_n}^{(n)}(\cdot) = \mathbf{P}(\mathbf{Y}^{(n)} \in \cdot \mid \mathbf{Y}^{(n)} \in A_n),$$

has density

$$\frac{dF_{A_n}^{(n)}}{d\mu}(y_1, \ldots, y_n) = \frac{\prod_{j=1}^n f_Y(y_j) I\{y_1 + \cdots + y_n > a_n\}}{p^{(n)}}. \tag{2.9}$$

The first step towards defining the estimator of $p^{(n)}$ is to construct the Markov chain $(\mathbf{Y}_t^{(n)})_{t\geq 0}$ whose invariant density is given by (2.9) using a Gibbs sampler. In short, the Gibbs sampler updates one element of $\mathbf{Y}_t^{(n)}$ at a time keeping the other elements constant. Formally the algorithm proceeds as follows.

Algorithm 2.4.1. Start at an initial state $\mathbf{Y}_0^{(n)} = (Y_{0,1}, \ldots, Y_{0,n})^{\mathrm{T}}$ where $Y_{0,1} + \cdots + Y_{0,n} > a_n$. Given $\mathbf{Y}_t^{(n)} = (Y_{t,1}, \ldots, Y_{t,n})^{\mathrm{T}}$, for some $t = 0, 1, \ldots$, the next state $\mathbf{Y}_{t+1}^{(n)}$ is sampled as follows:

1. Draw $j_1, \ldots, j_n$ from $\{1, \ldots, n\}$ without replacement and proceed by updating the components of $\mathbf{Y}_t^{(n)}$ 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

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

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

   $$\mathbf{P}(Y'_{t,j} \in \cdot \mid \mathbf{Y}_{t,-j}) = \mathbf{P}\Big(Y \in \cdot \mid Y + \sum_{k\neq j} Y_{t,k} > a_n\Big).$$

   b) Put $\mathbf{Y}'_t = (Y_{t,1}, \ldots, Y_{t,j-1}, Y'_{t,j}, Y_{t,j+1}, \ldots, Y_{t,n})^{\mathrm{T}}$.

3. Draw a random permutation π of the numbers $\{1, \ldots, n\}$ from the uniform distribution and put $\mathbf{Y}_{t+1}^{(n)} = (Y'_{t,\pi(1)}, \ldots, Y'_{t,\pi(n)})^{\mathrm{T}}$.

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

Remark 2.4.2. (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 (the 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 2.4.1 is not really needed when considering the probability $\mathbf{P}(S_n > a_n)$. 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 $\mathbf{P}(Y \in \cdot \mid Y > c)$ 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 $\mathbf{P}(Y \in \cdot \mid Y > c)$ may be difficult, see [13, Section 2.2].

The following proposition confirms that the Markov chain $(\mathbf{Y}_t^{(n)})_{t\geq 0}$, generated by Algorithm 2.4.1, has $F_{A_n}^{(n)}$ as its invariant distribution.

Proposition 2.4.3. *The Markov chain $(\mathbf{Y}_t^{(n)})_{t\geq 0}$, generated by Algorithm 2.4.1, has the conditional distribution $F_{A_n}^{(n)}$ as its invariant distribution.*

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

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

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

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

$$\begin{aligned}
F_{A_n}^{(n)}(B_1 \times \cdots \times B_n) &= \mathbf{E}\Big[\prod_{k=1}^{n} I\{Y_k \in B_k\} \mid S_n > a_n\Big] \\
&= \frac{\mathbf{E}[I\{Y_j \in B_j\} I\{S_n > a_n\} \prod_{k\neq j} I\{Y_k \in B_k\}]}{\mathbf{P}(S_n > a_n)} \\
&= \frac{\mathbf{E}\Big[\frac{\mathbf{E}[I\{Y_j \in B_j\} | Y_j > a_n - S_{n,-j}, \mathbf{Y}_{-j}^{(n)}] \prod_{k\neq j} I\{Y_k \in B_k\}}{\mathbf{P}(Y_j > a_n - S_{n,-j} | \mathbf{Y}_{-j}^{(n)})}\Big]}{\mathbf{P}(S_n > a_n)} \\
&= \frac{\mathbf{E}[P_j(\mathbf{Y}^{(n)}, B_1 \times \cdots \times B_n) \prod_{k\neq j} I\{Y_k \in B_k\}]}{P(S_n > a_n)} \\
&= \mathbf{E}[P_j(\mathbf{Y}^{(n)}, B_1 \times \cdots \times B_n) \mid S_n > a_n] \\
&= \mathbf{E}_{F_{A_n}^{(n)}}[P_j(\mathbf{Y}, B_1 \times \cdots \times B_n)],
\end{aligned}$$

with the conventional notation of writing $\mathbf{Y}^{(n)} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $S_n = Y_1 + \cdots + Y_n$, $\mathbf{Y}_{-j}^{(n)} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, Y_n)^{\mathrm{T}}$ and $S_{n,-j} = Y_1 + \cdots + Y_{j-1} + Y_{j+1} + \cdots + Y_n$. □

As for the ergodic properties, Algorithm 2.4.1 produces a Markov chain which is uniformly ergodic.

Proposition 2.4.4. *For each $n \geq 1$, the Markov chain $(\mathbf{Y}_t^{(n)})_{t\geq 0}$ is uniformly ergodic. In particular, it satisfies the following minorisation condition: there exists $\delta > 0$ such that*

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

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

Proof. Take an arbitrary $n \geq 1$. Uniform ergodicity can be deduced from the following minorisation condition (see [19]): there exists a probability measure ν, $\delta > 0$, and an integer t_0 such that

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

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

For any $\mathbf{x} \in A_n$ 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

$$\begin{aligned} g(\mathbf{x} \mid \mathbf{y}) \geq \frac{1}{n!} & \frac{f_Y(x_{j_1}) I\{x_{j_1} \geq a_n - \sum_{i \neq j_1} y_i\}}{\overline{F}_Y(a_n - \sum_{i \neq j_1} y_i)} \\ & \times \frac{f_Y(x_{j_2}) I\{x_{j_2} \geq a_n - \sum_{i \neq j_1, j_2} y_i - x_{j_1}\}}{\overline{F}_Y(a_n - \sum_{i \neq j_1, j_2} y_i - x_{j_1})} \\ & \vdots \\ & \times \frac{f_Y(x_{j_n}) I\{x_{j_n} \geq a_n - x_{j_1} - \ldots x_{j_{n-1}}\}}{\overline{F}_Y(a_n - x_{j_1} - \ldots x_{j_{n-1}})}. \end{aligned}$$

By construction of the ordering $j_1, \ldots, j_n$ all the indicators are equal to 1 and the expression in the last display is bounded from below by

$$\frac{1}{n!} \prod_{j=1}^{n} f_Y(x_j) = \frac{p^{(n)}}{n!} \cdot \frac{\prod_{j=1}^{n} f_Y(x_j) I\{x_1 + \cdots + x_n > a_n\}}{p^{(n)}}.$$

The proof is completed by integrating both sides of the inequality over any Borel set $B \subset A_n$. □

Remark 2.4.5. To keep the proof of Proposition 2.4.4 simple, we have not used the permutation step of the algorithm in the proof and not tried to optimise 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 of 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 a heavy-tailed random walk is considered. To be precise, assume that the variables

$Y_1, \ldots, Y_n$ 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

$$\lim_{n\to\infty} \frac{\mathbf{P}(S_n > a_n)}{\mathbf{P}(M_n > a_n)} = 1, \tag{2.10}$$

where M_n denotes the maximum of $Y_1, \ldots, Y_n$. The class of distributions for which (2.10) holds is large and includes the subexponential distributions. General conditions on the sequence (a_n) for which (2.10) holds are given in [9], see also [8]. For instance, if $\overline{F}_Y$ is regularly varying at ∞ with index $\beta > 1$ then (2.10) holds with $a_n = an$, for $a > 0$.

Next consider the choice of $V^{(n)}$. As observed in Section 2.2 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 the assumption (2.10) a good candidate for $V^{(n)}$ is the conditional distribution,

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

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

$$\frac{dV^{(n)}}{dF^{(n)}}(\mathbf{y}) = \frac{1}{\mathbf{P}(M_n > a_n)} I\{\mathbf{y} : \vee_{j=1}^n y_j > a_n\} = \frac{1}{1 - F_Y(a_n)^n} I\{\mathbf{y} : \vee_{j=1}^n y_j > a_n\}.$$

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

$$\widehat{q}_T^{(n)} = \frac{1}{T} \sum_{t=0}^{T-1} \frac{dV^{(n)}}{dF^{(n)}}(\mathbf{Y}_t^{(n)}) = \frac{1}{1 - F_Y(a_n)^n} \cdot \frac{1}{T} \sum_{t=0}^{T-1} I\{\vee_{j=1}^n Y_{t,j} > a_n\} \tag{2.11}$$

where $(\mathbf{Y}_t^{(n)})_{t\geq 0}$ is generated by Algorithm 2.4.1. Note that the estimator (2.11) can be viewed as the asymptotic approximation $(1 - F_Y(a_n)^n)^{-1}$ of $(p^{(n)})^{-1}$ multiplied by the random correction factor $\frac{1}{T}\sum_{t=0}^{T-1} I\{\vee_{j=1}^n Y_{t,j} > a_n\}$. The efficiency of this estimator is based on the fact that the random correction factor is likely to be close to 1 and has small variance.

Theorem 2.4.6. *Suppose that* (2.10) *holds. Then the estimator* $\widehat{q}_T^{(n)}$ *in* (2.11) *has vanishing normalised variance for estimating* $(p^{(n)})^{-1}$. *That is,*

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

Proof. With $u^{(n)}(\mathbf{y}) = \frac{1}{1-F_Y(a_n)^n} I\{\vee_{j=1}^n y_j > a_n\}$ it follows from (2.10) that

$$\begin{aligned}
&(p^{(n)})^2 \operatorname{Var}_{F^{(n)}_{A_n}}(u^{(n)}(\mathbf{Y}^{(n)})) \\
&= \frac{\mathbf{P}(S_n > a_n)^2}{\mathbf{P}(M_n > a_n)^2} \operatorname{Var}_{F^{(n)}_{A_n}}(I\{\mathbf{Y} : \vee_{j=1}^n Y_j > a_n\}) \\
&= \frac{\mathbf{P}(S_n > a_n)^2}{\mathbf{P}(M_n > a_n)^2} \mathbf{P}(M_n > a_n \mid S_n > a_n)\mathbf{P}(M_n \le a_n \mid S_n > a_n) \\
&= \frac{\mathbf{P}(S_n > a_n)}{\mathbf{P}(M_n > a_n)} \Big(1 - \frac{\mathbf{P}(M_n > a_n)}{\mathbf{P}(S_n > a_n)}\Big) \to 0.
\end{aligned}$$

This completes the proof. □

Remark 2.4.7. Theorem 2.4.6 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. [10, 6, 5]. It must be mentioned, though, that Theorem 2.4.6 proves efficiency for computing $(p^{(n)})^{-1}$, whereas the authors of [10, 6, 5] prove efficiency for a direct computation of $p^{(n)}$.

2.4.1 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 2.4.1 is modified to efficiently estimate hitting probabilities for heavy-tailed random sums.

Let $Y_1, Y_2, \ldots$ be non-negative independent random variables with common distribution F_Y and density f_Y. Let $(N^{(n)})_{n\ge1}$ 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

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

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

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

$$\begin{aligned}
&\mathbf{P}(N^{(n)} = k, (Y_1, \ldots, Y_k) \in \cdot \mid S_{N^{(n)}} > a_n) \\
&\quad = \frac{\mathbf{P}((Y_1, \ldots, Y_k) \in \cdot, S_k > a_n)\mathbf{P}(N^{(n)} = k)}{p^{(n)}}.
\end{aligned}$$

A Gibbs sampler for sampling from the above conditional distribution can be constructed essentially as in Algorithm 2.4.1. 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. To ease the notation the superscript n is suppressed in the description of the algorithm.

Algorithm 2.4.8. To initiate, draw N_0 from $\mathbf{P}(N \in \cdot)$ 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 $\overline{\mathbf{Y}}_t = (k_t, y_{t,1}, \ldots, y_{t,k_t})$ 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

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

   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 $\mathbf{Y}_t^{(1)} = (Y_{t,1}, \ldots, Y_{t,k_t}, Y_{t+1,k_t+1}, \ldots, Y_{t+1,N_{t+1}})$.

2. Proceed by updating all the individual steps as in Algorithm 2.4.1.

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

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

      i. Let $j = j_k$ be the index to be updated and write

$$\mathbf{Y}_{t,-j}^{(1)} = (Y_{t,1}^{(1)}, \ldots, Y_{t,j-1}^{(1)}, Y_{t,j+1}^{(1)}, \ldots, Y_{t,N_{t+1}}^{(1)}).$$

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

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

      ii. Put $\mathbf{Y}_t^{(2)} = (Y_{t,1}^{(1)}, \ldots, Y_{t,j-1}^{(1)}, Y_{t,j}^{(2)}, Y_{t,j+1}^{(1)}, \ldots, Y_{t,N_{t+1}}^{(1)})^{\mathrm{T}}$.

   c) Draw a random permutation π of the numbers $\{1, \ldots, N_{t+1}\}$ from the uniform distribution and put $\overline{\mathbf{Y}}_{t+1} = (N_{t+1}, Y_{t,\pi(1)}^{(2)}, \ldots, Y_{t,\pi(N_{t+1})}^{(2)})$.

Iterate until the entire Markov Chain $(\overline{\mathbf{Y}}_t)_{t=0}^{T-1}$ is constructed.

Proposition 2.4.9. *The Markov chain $(\overline{\mathbf{Y}}_t)_{t \geq 0}$ generated by Algorithm 2.4.8 has the conditional distribution $\mathbf{P}((N, Y_1, \ldots, Y_N) \in \cdot \mid Y_1 + \ldots Y_N > a_n)$ as its invariant distribution.*

Proof. The only essential difference from Algorithm 2.4.1 is the first step of the algorithm, where the number of steps and possibly the additional steps are updated. Therefore, 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 follows.

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

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 $\mathbf{P}((N, Y_1, \dots, Y_N) \in \cdot \mid Y_1 + \dots Y_N > a_n)$, it holds that

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

By conditioning on N and using independence of N and $Y_1, Y_2, \dots$ the expression in the last display equals

$$\begin{aligned}
&\frac{1}{\mathbf{P}(S_N > a_n)} \sum_{k_t=1}^{\infty} \mathbf{P}(N = k_t) \\
&\qquad \times \mathbf{E}\Big[P^{(1)}(k_t, Y_1, \dots, Y_{k_t}; k_{t+1}, A_1 \times \cdots \times A_{k_{t+1}}) I\{S_{k_t} > a_n\}\Big].
\end{aligned}$$

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

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

Inserting the expression for $P^{(1)}$ the last expression equals

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

Changing the order of summation the last expression equals

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

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

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

Summing over k_t the last expression equals

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

From the definition of $p(k_{t+1} \mid k^*)$ it follows that the last expression equals

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

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

Proposition 2.4.10. *The Markov chain $(\overline{\mathbf{Y}}_t)_{t \geq 0}$ generated by Algorithm 2.4.8 is uniformly ergodic. In particular, it satisfies the following minorisation condition: there exists $\delta > 0$ such that*

$$\mathbf{P}(\overline{\mathbf{Y}}_1 \in B \mid \overline{\mathbf{Y}}_0 = \overline{\mathbf{y}}) \geq \delta \, \mathbf{P}((N, Y_1, \ldots, Y_N) \in B \mid Y_1 + \cdots + Y_N > a_n),$$

for all $\overline{\mathbf{y}} \in A = \cup_{k \geq 1} \{(k, y_1, \ldots, y_k) : y_1 + \cdots + y_k > a_n\}$ and all Borel sets $B \subset A$.

The proof requires only a minor modification from the non-random case, Proposition 2.4.4, and is therefore omitted.

Next consider the distributional assumptions and the design of $V^{(n)}$. The main focus is on the rare event properties of the estimator and therefore the large deviation parameter n will be suppressed to ease notation. Let the distribution of the number of steps $\mathbf{P}(N^{(n)} \in \cdot)$ to depend on n. By a similar reasoning as in the case of non-random number of steps the following assumption are imposed: the variables $N^{(n)}, Y_1, Y_2, \ldots$ and the numbers a_n are such that

$$\lim_{n \to \infty} \frac{\mathbf{P}(Y_1 + \cdots + Y_{N^{(n)}} > a_n)}{\mathbf{P}(M_{N^{(n)}} > a_n)} = 1, \tag{2.12}$$

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

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

where $g_{N^{(n)}}(t) = \mathbf{E}[t^{N^{(n)}}]$ is the generating function of $N^{(n)}$. Sufficient conditions for (2.12) to hold are given in [16], Theorem 3.1. For instance, if $\overline{F}_Y$ is regularly varying at

∞ with index $\beta > 1$ and $N^{(n)}$ has Poisson distribution with mean $\lambda_n \to \infty$, as $n \to \infty$, then (2.12) holds with $a_n = a\lambda_n$, for $a > 0$.

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

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

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

$$\begin{aligned}\frac{dV^{(n)}}{dF^{(n)}}(k, y_1, \ldots, y_k) &= \frac{1}{\mathbf{P}(M_{N^{(n)}} > a_n)} I\{(y_1, \ldots, y_k) : \vee_{j=1}^k y_j > a_n\} \\ &= \frac{1}{1 - g_{N^{(n)}}(F_Y(a_n))} I\{(y_1, \ldots, y_k) : \vee_{j=1}^k y_j > a_n\}.\end{aligned}$$

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

$$\widehat{q}_T^{(n)} = \frac{1}{T} \sum_{t=0}^{T-1} \frac{dV^{(n)}}{dF^{(n)}}(\overline{\mathbf{Y}}_t^{(n)}) = \frac{1}{g_{N^{(n)}}(F_Y(a_n))} \cdot \frac{1}{T} \sum_{t=0}^{T-1} I\{\vee_{j=1}^{N_t} Y_{t,j} > a_n\}, \qquad (2.13)$$

where $(\overline{\mathbf{Y}}_t^{(n)})_{t \geq 0}$ is generated by Algorithm 2.4.8.

Theorem 2.4.11. *Suppose* (2.12) *holds. The estimator* $\widehat{q}_T^{(n)}$ *in* (2.13) *has vanishing normalised variance. That is,*

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

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

Remark 2.4.12. Because the distribution of $N^{(n)}$ may depend on n Theorem 2.4.11 covers a wider range of settings for random sums than those studied in [10, 5] where the authors present provably efficient importance sampling algorithms.

Proof. Since $p^{(n)} = \mathbf{P}(S_{N^{(n)}} > a_n)$ and

$$u^{(n)}(k, y_1, \ldots, y_k) = \frac{I\{\vee_{j=1}^k y_j > a_n\}}{\mathbf{P}(M_{N^{(n)}} > a_n)},$$

it follows that

$$\begin{aligned}&[p^{(n)}]^2 \operatorname{Var}_{\pi_n}(u^{(n)}(\overline{\mathbf{Y}}^{(n)})) \\ &\quad = \frac{\mathbf{P}(S_{N^{(n)}} > a_n)^2}{\mathbf{P}(M_{N^{(n)}} > a_n)^2} \operatorname{Var}_{\pi_n}(I\{\vee_{j=1}^{N^{(n)}} Y_j > a_n\}) \\ &\quad = \frac{\mathbf{P}(S_{N^{(n)}} > a_n)^2}{\mathbf{P}(M_{N^{(n)}} > a_n)^2} \mathbf{P}(M_{N^{(n)}} > a_n \mid S_{N^{(n)}} > a_n) \mathbf{P}(M_{N^{(n)}} \leq a_n \mid S_{N^{(n)}} > a_n) \\ &\quad = \frac{\mathbf{P}(S_{N^{(n)}} > a_n)}{\mathbf{P}(M_{N^{(n)}} > a_n)} \Big(1 - \frac{\mathbf{P}(M_{N^{(n)}} > a_n)}{\mathbf{P}(S_{N^{(n)}} > a_n)}\Big) \to 0,\end{aligned}$$

by (2.12). This completes the proof. □

2.5 Numerical experiments

The theoretical results guarantee that $\widehat{q}_T^{(n)}$ is an efficient estimator of $(p^{(n)})^{-1}$. However, for comparison of existing algorithms the numerical experiments are based on $\widehat{p}_T^{(n)} = (\widehat{q}_T^{(n)})^{-1}$ as an estimator for $p^{(n)}$. The literature includes numerical comparison for 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 is labelled IS.

By construction each simulation run of the MCMC algorithm only generates a single random variable (one simulation step) while both importance sampling and standard Monte Carlo generate n number of random variables (n simulation steps) for the case of fixed number of steps ($N + 1$ in the random number of steps case). Therefore the number of runs for the MCMC is scaled up by a factor of n so that all of the algorithms (MCMC, Monte Carlo and importance sampling) generate essentially the same number of random numbers. Thus getting a fairer comparison of the computer runtime between the three approaches.

First consider estimating $\mathbf{P}(S_n > a_n)$ 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 b number of batches, each consisting of T simulations in the case of importance sampling and standard Monte Carlo and Tn 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 2.1. The convergence of the algorithms can also be visualised by considering the point estimate as a function of number of simulation steps. This is presented in Figure 2.1. The MCMC algorithm appears to perform comparably with the importance sampling algorithm for p up to order 10^{-4} which is a relevant range in, say, insurance and finance. However for smaller p the MCMC appears to performs better. The improvement over importance sampling appears to increase as the event becomes more rare. This is due to the fact that the asymptotic approximation becomes better and better as the event becomes more rare.

Secondly consider estimating $\mathbf{P}(S_N > a_\rho)$ where $S_N = Y_1 + \cdots + Y_N$ with N geometrically distributed $\mathbf{P}(N = k) = (1-\rho)^{k-1}\rho$ for $k = 1, 2, \ldots$ and $a_\rho = a\mathbf{E}[N] = a/\rho$. The estimator considered here is $\widehat{p}_T = (\widehat{q}_T)^{-1}$ with $\widehat{q}_T$ as in (2.13). Again, each estimate is calculated using b number of batches, each consisting of T simulations in the case of importance sampling and standard Monte Carlo and $T\mathbf{E}[N]$ in the case of MCMC. The results are presented in Table 2.2. Also in the case of random number of steps the MCMC algorithm appears to outperform the importance sampling algorithm consistently for different choices of the parameters.

We remark that in our simulation with $\rho = 0.2$, $a = 5 \cdot 10^9$ the sample standard

deviation of the MCMC estimate is zero. This is because we did not observe any indicators $I\{\vee_{j=1}^{n} y_{t,j} > a_\rho\}$ being equal to 0 in this case.

Table 2.1: The table displays the batch mean and standard deviation of the estimates of $\mathbf{P}(S_n > a_n)$ as well as the average runtime per batch for time comparison. The number of batches run is b, each consisting of T simulations for importance sampling (IS) and standard Monte Carlo (MC) and Tn simulations for Markov chain Monte Carlo (MCMC). The asymptotic approximation is $p_{\max} = \mathbf{P}(\max\{Y_1, \ldots, Y_n\} > a_n)$.

$b = 25, T = 10^5, \beta = 2, n = 5, a = 5, p_{\max} =$ 0.737e-2			
	MCMC	IS	MC
Avg. est.	1.050e-2	1.048e-2	1.053e-2
Std. dev.	3e-5	9e-5	27e-5
Avg. time per batch(s)	12.8	12.7	1.4
$b = 25, T = 10^5, \beta = 2, n = 5, a = 20, p_{\max} =$ 4.901e-4			
	MCMC	IS	MC
Avg. est.	5.340e-4	5.343e-4	5.380e-4
Std. dev.	6e-7	13e-7	770e-7
Avg. time per batch(s)	14.4	13.9	1.5
$b = 20, T = 10^5, \beta = 2, n = 5, a = 10^3, p_{\max} =$ 1.9992e-7			
	MCMC	IS	
Avg. est.	2.0024e-7	2.0027e-7	
Std. dev.	3e-11	20e-11	
Avg. time per batch(s)	15.9	15.9	
$b = 20, T = 10^5, \beta = 2, n = 5, a = 10^4, p_{\max} =$ 1.99992e-9			
	MCMC	IS	
Avg. est.	2.00025e-9	2.00091e-9	
Std. dev.	7e-14	215e-14	
Avg. time per batch(s)	15.9	15.9	
$b = 25, T = 10^5, \beta = 2, n = 20, a = 20, p_{\max} =$ 1.2437e-4			
	MCMC	IS	MC
Avg. est.	1.375e-4	1.374e-4	1.444e-4
Std. dev.	2e-7	3e-7	492e-7
Avg. time per batch(s)	52.8	50.0	2.0
$b = 25, T = 10^5, \beta = 2, n = 20, a = 200, p_{\max} =$ 1.2494e-6			
	MCMC	IS	MC
Avg. est.	1.2614e-6	1.2615e-6	1.2000e-6
Std. dev.	4e-10	12e-10	33,166e-10
Avg. time per batch(s)	49.4	48.4	1.9
$b = 20, T = 10^5, \beta = 2, n = 20, a = 10^3, p_{\max} =$ 4.9995e-8			
	MCMC	IS	
Avg. est.	5.0091e-8	5.0079e-8	
Std. dev.	7e-12	66e-12	
Avg. time per batch(s)	53.0	50.6	
$b = 20, T = 10^5, \beta = 2, n = 20, a = 10^4, p_{\max} =$ 5.0000e-10			
	MCMC	IS	
Avg. est.	5.0010e-10	5.0006e-10	
Std. dev.	2e-14	71e-14	
Avg. time per batch(s)	48.0	47.1	

Acknowledgments

Henrik Hult's research was supported by the Göran Gustafsson Foundation. The authors thank the referee and Editor of Journal of Applied Probability for useful comments that

Table 2.2: The table displays the batch mean and standard deviation of the estimates of $\mathbf{P}(S_N > a_\rho)$ as well as the average runtime per batch for time comparison. The number of batches run is b, each consisting of T simulations for importance sampling (IS) and standard Monte Carlo (MC) and $T\,\mathbf{E}[N]$ simulations for Markov chain Monte Carlo (MCMC). The asymptotic approximation is $p_{\max} = \mathbf{P}(\max\{Y_1, \ldots, Y_N\} > a_\rho)$.

$b = 25, T = 10^5, \beta = 1, \rho = 0.2, a = 10^2, p_{\max} = 0.990\text{e-}2$			
	MCMC	IS	MC
Avg. est.	1.149e-2	1.087e-2	1.089e-2
Std. dev.	4e-5	6e-5	35e-5
Avg. time per batch(s)	25.0	11.0	1.2
$b = 25, T = 10^5, \beta = 1, \rho = 0.2, a = 10^3, p_{\max} = 0.999\text{e-}3$			
	MCMC	IS	MC
Avg. est.	1.019e-3	1.012e-3	1.037e-3
Std. dev.	1e-6	3e-6	76e-6
Avg. time per batch(s)	25.8	11.1	1.2
$b = 20, T = 10^6, \beta = 1, \rho = 0.2, a = 5 \cdot 10^7, p_{\max} = 2.000000\text{e-}8$			
	MCMC	IS	
Avg. est.	2.000003e-8	1.999325e-8	
Std. dev.	6e-14	1114e-14	
Avg. time per batch(s)	385.3	139.9	
$b = 20, T = 10^6, \beta = 1, \rho = 0.2, a = 5 \cdot 10^9, p_{\max} = 2.0000\text{e-}10$			
	MCMC	IS	
Avg. est.	2.0000e-10	1.9998e-10	
Std. dev.	0	13e-14	
Avg. time per batch(s)	358.7	130.9	
$b = 25, T = 10^5, \beta = 1, \rho = 0.05, a = 10^3, p_{\max} = 0.999\text{e-}3$			
	MCMC	IS	MC
Avg. est.	1.027e-3	1.017e-3	1.045e-3
Std. dev.	1e-6	4e-6	105e-6
Avg. time per batch(s)	61.5	44.8	1.3
$b = 25, T = 10^5, \beta = 1, \rho = 0.05, a = 5 \cdot 10^5, p_{\max} = 1.9999\text{e-}6$			
	MCMC	IS	MC
Avg. est.	2.0002e-6	2.0005e-6	3.2000e-6
Std. dev.	1e-10	53e-10	55,678e-10
Avg. time per batch(s)	60.7	45.0	1.3

helped improve the manuscript. The authors are is grateful to Tobias Rydén for his helpful discussion throughout the work of this paper.

2.6 References

[1] S. Asmussen. *Applied Probability and Queues*, volume 51 of *Stochastic Modelling and Applied Probability*. Springer, New York, 2003.

[2] S. Asmussen and K. Binswanger. Simulation of ruin probabilities for subexponential claims. *Astin Bulletin*, 27(2):297–318, 1997.

[3] S. Asmussen and P. W. Glynn. *Stochastic Simulation*, volume 57 of *Stochastic Modelling and Applied Probability*. Springer, New York, 2007.

[4] S. Asmussen and D. P. Kroese. Improved algorithms for rare event simulation with heavy tails. *Adv. Appl. Probab.*, 38:545–558, 2006.

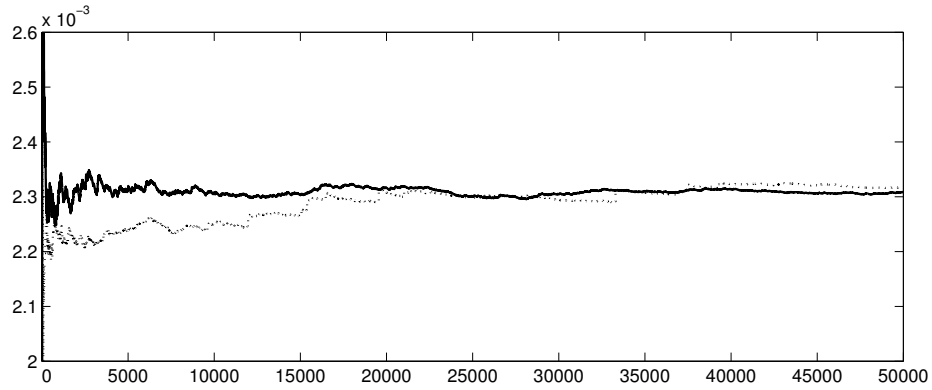


Figure 2.1: The figure illustrates the point estimate of $\mathbf{P}(S_n > a_n)$ as a function of the number of simulation steps, with $n = 5$, $a = 10$, $\beta = 2$. The estimate generated via the MCMC approach is drawn by a *solid line* and the estimate generated via IS is drawn by a *dotted line*.

[5] J. Blanchet and C. Li. Efficient rare-event simulation for heavy-tailed compound sums. *ACM T. Model Comput. S.*, 21:1–10, 2011.

[6] J. Blanchet and J. C. Liu. State-dependent importance sampling for regularly varying random walks. *Adv. Appl. Probab.*, 40:1104–1128, 2008.

[7] K. J. Chan. Asymptotic behavior of the Gibbs sampler. *J. Am. Stat. Assoc.*, 88:320–326, 1993.

[8] D. B. H. Cline and T. Hsing. Large deviation probabilities for sums of random variables with heavy or subexponential tails. Technical report, Texas A&M University, 1994.

[9] D. Denisov, A. B. Dieker, and V. Shneer. Large deviations for random walks under subexponentiality: The big-jump domain. *Ann. Probab.*, 36:1946–1991, 2008.

[10] P. Dupuis, K. Leder, and H. Wang. Importance sampling for sums of random variables with regularly varying tails. *ACM T. Model Comput. S.*, 17(3), 2007.

[11] A: Gelman and X. L. Meng. Aimulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Stat. Sci.*, 13(2):721–741, 1998.

[12] W. R. Gilks, S. Richardsson, and D. J. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman & Hall, 1996.

[13] H. Hult, F. Lindskog, O. Hammarlid, and C.J. Rehn. *Risk and Portfolio Analysis*. Springer, New York, 2012.

[14] G. L. Jones. On the Markov chain central limit theorem. *Probability Surveys*, 1:299–320, 2004.

[15] S. Juneja and P. Shahabuddin. Simulation heavy-tailed processes using delayed hazard rate twisting. *ACM T. Model Comput. S.*, 12(2):94–118, April 2002.

[16] C. Klüppelberg and T. Mikosch. Large deviations for heavy-tailed random sums with applications to insurance and finance. *J. Appl. Probab.*, 37(2):293–308, l 1997.

[17] K. L. Mengersen and R. L. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Stat.*, 24:101–121, 1996.

[18] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, New York, 1993.

[19] E. Nummelin. *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge University Press, 1984.

[20] J. S. Rosenthal. Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Am. Stat. Assoc.*, 90(430):558–566, June 1995.

[21] A. F. M. Smith and A. E. Gelfand. Bayesian statistics without tears: a sampling-resampling perspective. *Am. Stat.*, 46:84–88, 1992.

[22] L. Tierney. Markov chains for exploring posterior distributions (with discussion). *Ann. Stat.*, 22:1701–1762, 1994.

Markov chain Monte Carlo for rare-event simulation for stochastic recurrence equations with heavy-tailed innovations

Markov chain Monte Carlo for rare-event simulation for stochastic recurrence equations with heavy-tailed innovations

by

Thorbjörn Gudmundsson and Henrik Hult

Abstract

This paper extends a rare-event simulation technique based on sampling via Markov chain Monte Carlo (MCMC) introduced in [13]. Consider a stochastic recurrence equation of the form

$$X_{k+1} = A_{k+1}X_k + B_{k+1}, \quad X_0 = 0,$$

where $(A_k)_{k=1}^m$ and $(B_k)_{k=1}^m$ are independent sequences of non-negative independent and identically distributed random variables. The innovations B's are assumed to have regularly varying distribution with index $-\alpha < 0$ and that the A's satisfy the Breiman condition $\mathbf{E}[A^{\alpha+\epsilon}] < \infty$ for some $\epsilon > 0$. The problem considered is to efficiently compute $\mathbf{P}(X_m > c)$ for large c using Markov chain Monte Carlo techniques. The paper presents an estimator with uniformly bounded relative error as c tends to infinity.

The technique is also developed in the context of the risk reserves of an insurance company investing in risky investments and the problem of computing the probability of ruin.

3.1 Introduction

One of the key objectives of rare-event simulation is the design of efficient estimators for computing small probabilities. The ultimate goal is to construct a strongly efficient estimator meaning that its relative error is bounded as the probability tends to zero. Such an estimator thus provides a uniform variational bound on the desired probability.

The classic example of stochastic simulation is the standard technique of Monte Carlo. Its inefficiency for computation of rare-event probabilities is well documented and can be illustrated as follows. Consider the problem of computing $p = \mathbf{P}(X \in A)$ for a random variable X having probability distribution F. Given an independent random sample $X_0, \ldots, X_{T-1}$ from F a Monte Carlo estimator is defined by

$$\widehat{p}_T^{\,\mathrm{MC}} = \frac{1}{T}\sum_{t=0}^{T-1} I\{X_t \in A\}.$$

The Monte Carlo estimator is both simple and unbiased and its variance is given by

$$\mathrm{Var}(\widehat{p}_T^{\mathrm{MC}}) = \frac{1}{T}p(1-p).$$

While the variance tends to zero, for fixed p as $T \to \infty$, the normalised variance is unbounded, for fixed T as $p \to 0$:

$$\frac{\mathrm{Var}(\widehat{p}_T^{\mathrm{MC}})}{p^2} = \frac{1}{T}\Big(\frac{1}{p} - 1\Big).$$

This makes the Monte Carlo technique costly when it comes to rare-event simulation. For example, if a relative error at 1% is desired and the probability is of order 10^{-6} then we need to take T such that $\sqrt{(10^6-1)/T} \leq 0.01$. This implies that $T \approx 10^{10}$ which is infeasible on most computer systems. The intuitive reason for the inefficiency of the Monte Carlo is that very few hits are registered when the event is rare.

This paper assumes a heavy-tailed settings for the rare-event simulation. Heavy-tailed settings have been carefully investigated in recent years in the context of random walks and random sums and a number of methods have been suggested for efficient simulation. A variance reduction technique based on conditional Monte Carlo was presented in 1997 by Asmussen and Binswanger in [1] and later improved in 2006 by Asmussen and Kroese in [4]. However, the most popular technique for designing efficient estimators is importance sampling.

In importance sampling the random variables $X_0, \ldots, X_{T-1}$ are sampled independently from a different probability distribution, say G, instead of the original distribution F, with $G \ll F$ on A. The new distribution G is called the sampling distribution. The basic idea behind the importance sampling technique is to shift the probability mass, via the sampling distribution G, to the area where it is more likely that the rare event occurs and multiply with a weight function to correct for the change of measure. The importance sampling estimator is given by

$$\widehat{p}_T^{\mathrm{IS}} = \frac{1}{T}\sum_{t=0}^{T-1}\frac{dF}{dG}I\{X_t \in A\}.$$

The importance sampling estimator is unbiased and its performance depends on the choice of sampling distribution G. Choosing the sampling distribution to be equal to the conditional distribution

$$F_A(\cdot) = \mathbf{P}(X \in \cdot \mid X \in A),$$

implies that $\widehat{p}_T^{\mathrm{IS}}$ has zero variance and is therefore called the zero-variance distribution. It serves as guidance in choosing a sampling distribution. A good candidate for G fullfills the following:

(i) The sampling distribution G approximates F_A and is such that that the random variable X can be easily sampled from G.

(ii) The event $\{X \in A\}$ is more likely under G than the original distribution F.

(iii) The likelihood ratio dF/dG is unlikely to become too large.

The standard importance sampling technique for the light-tailed case is based on an exponential change of measure, see for instance [3] or [8], but that approach has had limited success in the heavy-tailed case since the exponential moments do not exist. Moreover, as illustrated by Asmussen et al. in [2] the situation in the heavy-tailed setting is more complicated because the conditional distribution $F_A(\cdot)$ typically becomes singular with respect to the original distribution as the desired probability tends to zero.

An efficient importance sampling estimator in a heavy-tailed settings for a random walk was presented in 2007 by Dupuis et al. [10], based on mixtures and sequential sampling. Efficient importance sampling techniques have also been presented in e.g. [6], [7] and [15]. However, proving efficiency for an importance sampling estimator can be technically cumbersome and often requires extensive analysis.

A new technique for computing rare-event probabilities based on sampling via Markov chain Monte Carlo is proposed in [13]. The main idea of this method is to use MCMC algorithm to sample from the conditional distribution given the event of interest and then extract the probability of the event as the normalising constant. In [13] we apply the MCMC method on the examples of random walk and random sum. The goal of this paper is to continue the development of the MCMC method by extending it to the solution of a recurrence equation.

Consider the solution $(X_k)_{k=0}^m$ of the stochastic recurrence equation

$$X_{k+1} = A_{k+1}X_k + B_{k+1}, \quad X_0 = 0, \tag{3.1}$$

where $\mathbf{A} = (A_k)_{k=1}^m$ and $\mathbf{B} = (B_k)_{k=1}^m$ are independent sequences of non-negative independent and identically distributed random variables. The innovations B's are assumed to have regularly varying distribution with index $-\alpha < 0$,

$$\frac{\mathbf{P}(B > tx)}{\mathbf{P}(B > t)} \to x^{-\alpha}, \quad \text{as } t \to \infty, \quad \text{for } x > 0,$$

and the A's satisfy the Breiman condition

$$\mathbf{E}[A^{\alpha+\epsilon}] < \infty, \quad \text{for some } \epsilon > 0.$$

Thorough studies of the solution $(X_k)_{k=0}^m$ can be found in the literature, e.g. in [17], in the setting of insurance company with risky investment [21], and in an economic environment with focus on financial processes [9]. The MCMC method presented in this paper will be exemplified in that context as well. Recently an efficient importance sampling estimator for computing the probability that the solution exceeds a high threshold was introduced by Blanchet, Hult and Leder in [14].

This paper extends the MCMC methodology first established in [13] to study a rare-event simulation based on MCMC for a solution to a stochastic recurrence equations given by (3.1). An estimator $\widehat{q}$ of the reciprocal of the rare-event probability that the solution X_m exceeds a high threshold is presented, namely $1/p = \mathbf{P}(X_m > c)^{-1}$. The estimator is unbiased and proven to be rare-event efficient in the sense that

$$p^2 \operatorname{Var}(\widehat{q}) \to 0, \quad \text{as } c \to 0.$$

The technique introduced here is also applied to compute the ruin probability for the setting of an insurance company with risky investments.

The proof of efficiency is elementary and completed in in just a few lines. This is in sharp contrast to efficiency proofs for importance sampling algorithms for the same problem, which require more restrictive assumptions and tend to be long and technical, see for instance [14].

Here follows the outline of the paper. In Section 3.2 the fundamental idea of the MCMC technique is presented. The key design choices that control the efficiency of the algorithm are highlighted. In Section 3.3 the MCMC method is applied to a solution of a stochastic recurrence equation and an efficient MCMC estimator derived. Numerical experiments demonstrate the performance and the algorithm is compared to existing importance sampling techniques. Finally, in Section 3.4 the method is applied to compute the ruin probability of an insurance company with risky investments.

3.2 Markov chain Monte Carlo methodology

In this section the Markov chain Monte Carlo (MCMC) methodology in rare-event simulation is introduced and the important design choices highlighted. The approach is later used in Section 3.3 to develop efficient estimator for the solution to stochastic recurrence equations and in Section 3.4 for an example of an insurance company.

Let $X^{(1)}, X^{(2)}, \ldots$ be a sequence of random elements, each of which can be sampled via a simulation algorithm, and consider the problem of computing $p^{(n)} = \mathbf{P}(X^{(n)} \in C)$ where the event $\{X^{(n)} \in C\}$ is rare in the sense that $p^{(n)} \to 0$ as $n \to \infty$. Let $\widehat{p}_T^{(n)}$ be an estimator of $p^{(n)}$ based on a sample $X_0^{(n)}, \ldots, X_{T-1}^{(n)}$. The efficiency of $\widehat{p}_T^{(n)}$ is characterised by its relative error. An estimator is said to be strongly efficient if it has bounded relative error. An efficient estimator $\widehat{p}_T^{(n)}$ of $p^{(n)}$ is said to have bounded relative error if

$$\frac{\mathrm{Var}(\widehat{p}_T^{(n)})}{\left(p^{(n)}\right)^2} < \infty, \quad \text{as } n \to \infty.$$

An estimator has vanishing relative error if

$$\frac{\mathrm{Var}(\widehat{p}_T^{(n)})}{\left(p^{(n)}\right)^2} \to 0, \quad \text{as } n \to \infty.$$

Consider a Markov chain $(X_t^{(n)})_{t\geq 0}$ constructed via an MCMC algorithm, such as a Gibbs sampler or Metropolis-Hastings sampler, whose invariant distribution is the conditional distribution

$$F_C^{(n)}(\cdot) = \mathbf{P}(X^{(n)} \in \cdot \mid X^{(n)} \in C).$$

To construct an unbiased estimator $\widehat{q}_T^{(n)}$ of $1/p^{(n)}$, consider a probability distribution $V^{(n)}$ with $V^{(n)} \ll F_C^{(n)}$ and define u, a function of the Markov chain $(X_t^{(n)})_{t\geq 0}$, as

follows

$$u(X^{(n)}) = \frac{dV^{(n)}}{dF^{(n)}}(X^{(n)}). \tag{3.2}$$

Taking expectation with respect to the conditional distribution of the sample

$$\mathbf{E}_{F_C^{(n)}}[u(X^{(n)})] = \int \frac{dV^{(n)}}{dF^{(n)}}(X^{(n)}) dF_C^{(n)}(X^{(n)}) = \frac{1}{p^{(n)}},$$

thus motivating taking the MCMC estimator to be given by

$$\widehat{q}_T^{(n)} = \frac{1}{T}\sum_{t=0}^{T-1} u(X_t^{(n)}). \tag{3.3}$$

There are two important design choices that determine the performance of this method: the choice of the distribution $V^{(n)}$ and the design of the MCMC algorithm. The distribution $V^{(n)}$ controls the variance of $u(X^{(n)})$ in (3.2) and is thus crucial when ensuring good rare-event properties of the method. It is desirable to take $V^{(n)}$ such that the normalised variance of the estimator, given by $\left(p^{(n)}\right)^2 \operatorname{Var}(\widehat{q}_T^{(n)})$, is not too large. The design of the MCMC algorithm , on the other hand, controls the dependence of the Markov chain and thereby the convergence rate of the algorithm as sample size grows. In order to speed up convergence it is desirable that the Markov chain mixes fast so that the dependence dies out quickly.

Firstly consider the choice of the probability distribution $V^{(n)}$ which determines the rare-event efficiency. The variance of the estimator under the invariant distribution $F_C^{(n)}$ of the Markov chain is controlled by

$$\operatorname{Var}_{F_C^{(n)}}\left(u(X^{(n)})\right) = \ldots = \frac{1}{(p^{(n)})^2}\left(\mathbf{E}_{V^{(n)}}\left[\frac{dV^{(n)}}{dF_C^{(n)}}\right] - 1\right),$$

for detailed computations, the reader is referred to [13]. Observe that letting $V^{(n)}$ be equal to the conditional distribution $F_C^{(n)}$ implies $\operatorname{Var}_{F_C^{(n)}}\left(u(X^{(n)})\right) = 0$. This motivates taking $V^{(n)}$ as an approximation of $F_C^{(n)}$, similar to the ideology behind choosing an efficient importance sampling estimator.

For any $R \subset C$, for which $\mathbf{P}(X^{(n)} \in R)$ can be computed explicitly, a candidate for $V^{(n)}$ is given by

$$V^{(n)}(\cdot) = \mathbf{P}(X^{(n)} \in \cdot \mid X^{(n)} \in R).$$

Such a choice is likely to perform well if $\mathbf{P}(X^{(n)} \in R)$ is close to $\mathbf{P}(X^{(n)} \in C)$ since then

$$\mathbf{E}_{V^{(n)}}\left[\frac{dV^{(n)}}{dF_C^{(n)}}\right] = \frac{\mathbf{P}(X^{(n)} \in C)}{\mathbf{P}(X^{(n)} \in R)^2}\mathbf{E}_{V^{(n)}}\left[I\{X^{(n)} \in R\}\right] = \frac{\mathbf{P}(X^{(n)} \in C)}{\mathbf{P}(X^{(n)} \in R)},$$

which will be close to 1.

Secondly, consider the design choice of the MCMC-sampler which determines the strength of the dependence in the Markov chain. High dependence implies slow convergence and therefore high computational cost. The algorithm should thus be designed so that the Markov chain mixes fast and that the dependence dies out quickly. The minimum requirement is that the Markov chain is geometric ergodic, which guarantees large-sample efficiency. Again, the reader is referred to [13] for more details.

To summarise, the following are desired properties of the estimator.

(i) *Rare event efficiency:* Construct an unbiased estimator $\widehat{q}_T^{(n)}$ of $1/p^{(n)}$ according to (3.3) by finding a probability distribution function $V^{(n)}$ which approximates the conditional distribution $F_C^{(n)}$. The choice of $V^{(n)}$ controls $(p^{(n)})^2 \operatorname{Var}_{F_C^{(n)}}\big(u(X^{(n)})\big)$, which in turn controls the normalised variance of the estimator.

(ii) *Large sample efficiency:* Design the MCMC sampler, by finding an appropriate Gibbs sampler or a proposal density in the Metropolis-Hastings algorithm, such that the resulting Markov chain is geometrically ergodic.

3.3 Stochastic recurrence equation

In this section the MCMC estimator introduced in Section 3.2 is applied for computing the probability that a solution to a stochastic recurrence equation exceeds a high threshold. The estimator has vanishing normalised variance and the associated Markov chain is uniformly ergodic.

Fix a positive integer m and let $\mathbf{A} = (A_1, \ldots, A_m)$ and $\mathbf{B} = (B_1, \ldots, B_m)$ be independent sequences of independent and identically distributed random variables. Let A be a generic random variable for an element of the sequence $\mathbf{A}$ and likewise B for an element of the sequence $\mathbf{B}$. Observe that $(\mathbf{A}, \mathbf{B})$ plays the role of X in the previous section.

Consider the solution $(X_k)_{k=0}^m$ of the stochastic recurrence equation of the form

$$\begin{aligned} X_k &= A_k X_{k-1} + B_k, \quad \text{for } k = 1, \ldots, m, \\ X_0 &= 0, \end{aligned}$$

and the problem of computing

$$p^{(n)} = \mathbf{P}(X_m > c_n),$$

where $c_n \to \infty$ as $n \to \infty$.

The solution $(X_k)_{k=0}^m$ can be written as a randomly weighted random walk

$$X_k = B_k + A_k B_{k-1} + \cdots + A_k A_{k-1} \cdots A_2 B_1, \quad \text{for } k = 1, \ldots, m. \tag{3.4}$$

The first step is to design a Gibbs sampler that produces a Markov chain with the conditional distribution

$$F_{c_n}(\cdot) = \mathbf{P}\big((\mathbf{A}, \mathbf{B}) \in \cdot \mid X_m > c_n\big),$$

as its invariant distribution. In addition, a probability distribution $V^{(n)}$ will be suggested having good asymptotic properties.

The Markov chain $(\mathbf{A}_t, \mathbf{B}_t)_{t\geq 0}$ is constructed by the following algorithm, where the elements are updated sequentially in such a way that the weighted random walk exceeds the threshold after each individual update. Formally the algorithm is given as follows. An empty product, such as $\prod_{j=m+1}^{m} A_j$, is interpreted as 1.

Algorithm 3.3.1. Start with initial state $(\mathbf{A}_0, \mathbf{B}_0) = (A_{0,1}, \ldots, A_{0,m}, B_{0,1}, \ldots, B_{0,m})$ where $X_0^{(m)} = B_{0,m} + \sum_{i=1}^{m-1} B_{0,i} \prod_{j=i+1}^{m} A_{0,j} > c_n$. Given $(\mathbf{A}_t, \mathbf{B}_t)$, for some $t \geq 0$, the next state $(\mathbf{A}_{t+1}, \mathbf{B}_{t+1})$ is sampled as follows:

1. Draw a randomised ordering $j_1, \ldots, j_{2m}$ of $\{1, \ldots, 2m\}$ and proceed updating $(\mathbf{A}_t, \mathbf{B}_t)$ in the order thus obtained.

2. For $l = 1, \ldots, 2m$, set $k = j_l$ and do the following:

   i. If $k \in \{1, \ldots, m\}$ then $A_{t,k}$ is to be updated. Sample A' from the conditional distribution
   $$\mathbf{P}(A' \in \cdot \mid A' > s),$$
   where
   $$s = \max\left\{ \frac{c_n - \sum_{i=k}^{m} B_{t,i} \prod_{j=i+1}^{m} A_{t,j}}{\sum_{i=1}^{k-1} B_{t,i} \prod_{j=i+1, \neq k}^{m} A_{t,j}}, 0 \right\}.$$
   Put $\mathbf{A}_{t+1} = (A_{t,1}, \ldots, A_{t,k-1}, A', A_{t,k+1}, \ldots, A_{t,m})$ and $\mathbf{B}_{t+1} = \mathbf{B}_t$.

   ii. If $k \in \{m+1, \ldots, 2m\}$ then B_{t,k^*}, where $k^* = k - m$, is to be updated. Sample B' from the conditional distribution
   $$\mathbf{P}(B' \in \cdot \mid B' > s),$$
   where
   $$s = \max\left\{ \frac{c_n - \sum_{i=1, \neq k^*}^{m} B_{t,i} \prod_{j=i+1}^{m} A_{t,j}}{A_{t,m} \cdots A_{t,k^*+1}}, 0 \right\}.$$
   Put $\mathbf{A}_{t+1} = \mathbf{A}_t$ and $\mathbf{B}_{t+1} = (B_{t,1}, \ldots, B_{t,k^*-1}, B', B_{t,k^*+1}, \ldots, B_{t,m})$.

Iterate steps 1 and 2 until the entire Markov chain $(\mathbf{A}_t, \mathbf{B}_t)_{t=0}^{T-1}$ is constructed.

Proposition 3.3.2. *The Markov chain $(\boldsymbol{A}_t, \boldsymbol{B}_t)_{t\geq 0}$ generated by Algorithm 3.3.1, has the conditional distribution F_{c_n} as its invariant distribution.*

Proof. Note that it is sufficient to show that each updating step (Step 2i and 2ii in the Algorithm) preserves stationarity.

Consider the updating steps (Step 2i and 2ii). Let m be given and set $P_k^A(\mathbf{a}, \mathbf{b}, \cdot)$ and $P_k^B(\mathbf{a}, \mathbf{b}, \cdot)$ to be the transition probability of the Markov chain $(\mathbf{A}_t, \mathbf{B}_t)_{t\geq 0}$ where the kth element of $\mathbf{A}_t$ and $\mathbf{B}_t$ is updated, respectively. Let

$$C^{(n)} = \big\{(A_1, \ldots, A_m, B_1, \ldots, B_m) \mid X_m > c_n\},$$

and observe that if A_k is to be updated conditioned on $X_m > c_n$ then

$$A_k > \frac{c_n - \sum_{i=k}^{m} B_{t,i} \prod_{j=i+1}^{m} A_{t,j}}{\sum_{i=1}^{k-1} B_{t,i} \prod_{j=i+1,\neq k}^{m} A_{t,j}} =: s_{A_k},$$

and similarly, if B_k is to be updated conditioned on $X_m > c_n$ then

$$B_k > \frac{c_n - \sum_{i=1,\neq(k-m)}^{m} B_{t,i} \prod_{j=i+1}^{m} A_{t,j}}{A_{t,m} \cdots A_{t,(k-m)+1}} =: s_{B_k}.$$

To prove that stationarity is preserved under updating via Step 2i it is sufficient to show that for arbitrary $k \in \{1, \ldots, m\}$ and $D_1 \times \cdots \times D_m \times E_1 \times \cdots \times E_m \subset C^{(n)}$ then it holds that

$$\begin{aligned} &F_{c_n}(D_1 \times \cdots \times D_m \times E_1 \times \cdots \times E_m) \\ &\quad = \mathbf{E}_{F_{c_n}}\big[P_k^A(A_1, \ldots, A_m, B_1, \ldots, B_m, D_1 \times \ldots \times D_m \times E_1 \times \ldots \times E_m)\big]. \end{aligned} \tag{3.5}$$

Similarly to prove that stationarity is preserved under updating via Step 2ii it is sufficient to show

$$\begin{aligned} &F_{c_n}(D_1 \times \cdots \times D_m \times E_1 \times \cdots \times E_m) \\ &\quad = \mathbf{E}_{F_{c_n}}\big[P_k^B(A_1, \ldots, A_m, B_1, \ldots, B_m, D_1 \times \cdots \times D_m \times E_1 \times \cdots \times E_m)\big]. \end{aligned} \tag{3.6}$$

The following computation shows that (3.5) holds.

$$\begin{aligned} &F_{c_n}(D_1 \times \cdots \times D_m \times E_1 \times \cdots \times E_m) \\ &\quad = \mathbf{E}_{F_{c_n}}\Big[\prod_{j=1}^{m} I\{A_j \in D_j\} \prod_{i=1}^{m} I\{B_i \in E_i\}\Big] \\ &\quad = \frac{\mathbf{E}\big[I\{A_k \in D_k\} I\{X_m > c_n\} \cdot \prod_{j=1,\neq k}^{m} I\{A_j \in D_j\} \prod_{i=1}^{m} I\{B_i \in E_i\}\big]}{\mathbf{P}(X_m > c_n)} \\ &\quad = \frac{\mathbf{E}\Big[\frac{\mathbf{E}[I\{A_k \in D_k\} | A_k > s_{A_k}, \mathbf{A}_{-k}, \mathbf{B}]}{\mathbf{P}(A_k > s_{A_k})} \cdot \prod_{j=1,\neq k}^{m} I\{A_j \in D_j\} \prod_{i=1}^{m} I\{B_i \in E_i\}\Big]}{\mathbf{P}(X_m > c_n)} \\ &\quad = \mathbf{E}\big[P_k^A(\mathbf{A}, \mathbf{B}, D_1 \times \cdots \times D_m \times E_1 \times \cdots \times E_m) \\ &\qquad \times \frac{\prod_{j=1,\neq k}^{m} I\{A_j \in D_j\} \prod_{i=1}^{m} I\{B_i \in E_i\}\big]}{\mathbf{P}(X_m > c_n)} \\ &\quad = \mathbf{E}\big[P_k^A(\mathbf{A}, \mathbf{B}, D_1 \times \cdots \times D_m \times E_1 \times \cdots \times E_m) \mid X_m > c_n\big] \\ &\quad = \mathbf{E}_{F_{c_n}}\big[P_k^A(\mathbf{A}, \mathbf{B}, D_1 \times \cdots \times D_m \times E_1 \times \cdots \times E_m)\big], \end{aligned}$$

with the conventional notation $\mathbf{A}_{-k} = (A_1, \ldots, A_{k-1}, A_{k+1}, \ldots, A_m)$.

The proof is completed by showing that (3.6) holds with similar computation as above. □

Proposition 3.3.3. *For any $m \geq 1$, the Markov chain $(\boldsymbol{A}_t, \boldsymbol{B}_t)_{t\geq 0}$ is uniformly ergodic.*

Proof. Let $m \geq 1$ be given and set

$$C^{(n)} = \big\{(A_1, \ldots, A_m, B_1, \ldots, B_m) \mid X_m > c_n\big\}.$$

Uniform ergodicity follows from the minorisation condition, see [20]: there exists a probability measure ν, $\delta > 0$ and $t_0 \in \mathbb{N}$ such that

$$\mathbf{P}\big((\mathbf{A}_{t_0}, \mathbf{B}_{t_0}) \in D \times E \mid (\mathbf{A}_0, \mathbf{B}_0) = (\mathbf{a}, \mathbf{b})\big) \geq \delta\nu(D \times E),$$

for any $(\mathbf{a}, \mathbf{b})$ and $D \times E \subset C^{(n)}$. The goal is to prove this inequality for $t_0 = 1$, $\delta = p^{(n)}/(2m)!$ and $\nu = F_{c_n}$.

Take $\mathbf{c} = (\mathbf{a}, \mathbf{b})$ and let $g(\cdot \mid \mathbf{a}, \mathbf{b})$ be the density of $\mathbf{P}(\mathbf{A}_1, \mathbf{B}_1 \in \cdot \mid \mathbf{A}_0, \mathbf{B}_0 = \mathbf{a}, \mathbf{b})$. Observe that for any $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in C^{(n)}$ there exists an ordering $j_1, \ldots, j_{2m}$ of $\{1, \ldots, 2m\}$ such that

$$\begin{aligned} c_{j_1} \leq z_{j_1}, \ldots, c_{j_k} \leq z_{j_k} \\ c_{j_{k+1}} \geq z_{j_{k+1}}, \ldots, c_{j_{2m}} \geq z_{j_{2m}}, \end{aligned}$$

for some k. When updating from $\mathbf{c}$ to $\mathbf{z}$ using this particular ordering, then first all of elements in $\mathbf{z}$ which are larger than their counterparts in $\mathbf{c}$ are updated, and then all of the elements in $\mathbf{z}$ which are smaller are updated. This guarantees that after every updating step, the updated vector belongs to $C^{(n)}$.

The probability for this particular ordering is $1/(2m)!$. To simplify notation, introduce

$$Z_k = \begin{cases} A_i & \text{if update } j_k \text{ corresponds to updating } A_i \text{ for some } i \\ B_i & \text{if update } j_k \text{ corresponds to updating } B_i \text{ for some } i \end{cases}$$

and

$$s_{Z_k} = \begin{cases} s_{A_i} & \text{if update } j_k \text{ corresponds to updating } A_i \text{ for some } i \\ s_{B_i} & \text{if update } j_k \text{ corresponds to updating } B_i \text{ for some } i \end{cases}$$

Therefore

$$\begin{aligned} g(\mathbf{x}, \mathbf{y}) \quad = \quad & \frac{1}{(2m)!} \frac{f_{Z_1}(z_{j_1}) I\{Z_1 > s_{Z_1}\}}{\mathbf{P}(Z > s_{Z_1})} \\ & \times \frac{f_{Z_2}(z_{j_2}) I\{Z_2 > s_{Z_2}\}}{\mathbf{P}(Z > s_{Z_2})} \\ & \vdots \\ & \times \frac{f_{Z_{2m}}(z_{j_{2m}}) I\{Z_{2m} > s_{Z_{2m}}\}}{\mathbf{P}(Z > s_{Z_{2m}})}. \end{aligned}$$

By construction all of the indicator functions are equal to 1 and the normalising probabilities are bounded by 1 so the last display is bounded from below by

$$\frac{1}{(2m)!} \prod_{k=1}^{2m} f_{Z_k}(z_k) = \frac{p^{(n)}}{(2m)!} \cdot \frac{\prod_{k=1}^{2m} f_{Z_k}(z_k) I\{\mathbf{z} \in C^{(n)}\}}{p^{(m)}}.$$

The proof is completed by integrating both sides. □

Remark 3.3.4. In order to keep the proof of Proposition 3.3.3 simple and short the choice of δ is not optimised. Taking advantage of the permutation step (Step 1 of the Algorithm 3.3.1) the constant δ could, with some additional effort, be chosen to be larger.

As mentioned in Section 3.2 a good candidate for $V^{(n)}$ is a probability distribution

$$\mathbf{P}\big((\mathbf{A},\mathbf{B}) \in \cdot \mid (\mathbf{A},\mathbf{B}) \in R^{(n)}\big),$$

where $r^{(n)} = \mathbf{P}\big((\mathbf{A},\mathbf{B}) \in R^{(n)}\big)$ is asymptotically close to $p^{(n)} = \mathbf{P}(X_m > c_n)$ in the sense that $r^{(n)}/p^{(n)} \to 1$ as $n \to \infty$.

Observe that so far no limitation have been set on the probabilistic properties of $\mathbf{A}$ and $\mathbf{B}$. The distributional assumptions have been very general. For the design of $V^{(n)}$ the probabilistic properties of $\mathbf{A}$ and $\mathbf{B}$ are of central importance and here they come into play. This paper considers the setting where the innovations B are most likely responsible for extreme values of the solution to the stochastic recurrence equation. The following is assumed.

1. The generic random variables A and B are nonnegative.
2. The generic random variable B has a regularly varying tail, with index $-\alpha < 0$. Formally,
$$\lim_{t\to\infty} \frac{\mathbf{P}(B > xt)}{\mathbf{P}(B > t)} = x^{-\alpha}, \text{ for all } x > 0.$$
3. The Breiman condition holds for the generic random variable A. That is, there exists $\epsilon > 0$ such that
$$\mathbf{E}[A^{\alpha+\epsilon}] < \infty.$$

Under the assumptions (1)-(3) it is possible to derive the asymptotic decay of $p^{(n)}$. Indeed, observe that by the weighted random walk representation of (3.4) it follows from [5], that

$$\frac{\mathbf{P}(X_m > c_n)}{\mathbf{P}(B > c_n)} \to \sum_{k=0}^{m-1} E[A^\alpha]^k, \quad \text{as } n \to \infty,$$

so that

$$p^{(n)} \sim \overline{F}_B(c_n) \sum_{k=0}^{m-1} \mathbf{E}[A^\alpha]^k, \quad \text{as } n \to \infty. \tag{3.7}$$

Now consider the choice of $V^{(n)}$. Let $V^{(n)}$ be defined as the probability distribution

$$V^{(n)}(\cdot) = \mathbf{P}\big((\mathbf{A},\mathbf{B}) \in \cdot \mid (\mathbf{A},\mathbf{B}) \in R^{(n)}\big),$$

with

$$R^{(n)} = \bigcup_{k=1}^{m} R_k, \text{ where } \quad R_k = \{A_m \cdots A_{k+1} B_k > c_n,\ A_m, \ldots, A_{k+1} > a\},$$

for some constant $a > 0$. The probability of this conditioning event can be computed explicitly using the inclusion-exclusion formula

$$\begin{aligned} r^{(n)} \quad &= \quad \sum_{k=1}^{m} \mathbf{P}(R_k) - \sum_{k_1=1}^{m-1} \sum_{k_2=k_1+1}^{m} \mathbf{P}(R_{k_1} \cap R_{k_2}) \\ &+ \quad \sum_{k_1=1}^{m-2} \sum_{k_2=k_1+1}^{m-1} \sum_{k_3=k_2+1}^{m} \mathbf{P}(R_{k_1} \cap R_{k_2} \cap R_{k_3}) \\ &+ \ldots \quad + (-1)^m \mathbf{P}(R_1 \cap \cdots \cap R_m), \end{aligned}$$

where each term can be computed as follows

$$\mathbf{P}\Big(\bigcap_{j \in J} R_{k_j} \Big) = \overline{F}_A(a)^{m-k} \prod_{j \in J} \overline{F}_B(c_n / a^{m-k_j}), \quad \text{where } k = \min\{k_j : j \in J\}.$$

From the regular variation property of the distribution of B, assumption (2), it follows that

$$r^{(n)} \sim \overline{F}_B(c_n) \sum_{k=0}^{m-1} \overline{F}_A(a)^k a^{k\alpha}, \quad \text{as } n \to \infty. \tag{3.8}$$

A convenient choice of the level $a = a_n$ is such that $r^{(n)}/p^{(n)} \to 1$, as $n \to \infty$. Comparing equation (3.7) and (3.8), a may be chosen as the solution to

$$\sum_{k=0}^{m-1} \overline{F}_A(a)^k a^{k\alpha} = \sum_{k=0}^{m-1} E[A^\alpha]^k. \tag{3.9}$$

The distribution $V^{(n)}$ has a known density with respect to $F(\cdot) = \mathbf{P}\big((\mathbf{A}, \mathbf{B}) \in \cdot\big)$ given by

$$\frac{dV^{(n)}}{dF}(\mathbf{a}, \mathbf{b}) = \frac{1}{r^{(n)}} I\big\{(\mathbf{a}, \mathbf{b}) \in R^{(n)}\big\}.$$

Thus the MCMC estimator $\widehat{q}_T^{(n)}$ of $1/p^{(n)}$ is given by

$$\widehat{q}_T^{(n)} = \frac{1}{r^{(n)}} \frac{1}{T} \sum_{t=0}^{T-1} I\big\{(\mathbf{A}_t, \mathbf{B}_t) \in R^{(n)}\big\}, \tag{3.10}$$

where $(\mathbf{A}_t, \mathbf{B}_t)_{t=0}^{T-1}$ is generated via Algorithm 3.3.1 and $r^{(n)}$ given explicitly by equation (3.8). Observe that the estimator first factor of the estimator $\widehat{q}_T^{(n)}$ may be interpreted as the asymptotic approximation $1/r^{(n)}$ multiplied by a stochastic correction factor.

Theorem 3.3.5. *The estimator $\widehat{q}_T^{(n)}$ given by 3.10 has vanishing normalised variance for estimating $1/p^{(n)}$,*

$$\lim_{n \to \infty} \big(p^{(n)}\big)^2 \operatorname{Var}_{F_{c_n}}(\widehat{q}_T^{(n)}) \to 0.$$

Proof. With $u^{(n)}(\mathbf{a},\mathbf{b}) = \frac{1}{r^{(n)}} I\{(\mathbf{a},\mathbf{b}) \in R\}$ it follows from assumptions (1)-(3) above that

$$\begin{aligned}
&(p^{(n)})^2 \operatorname{Var}_{F_{c_n}} \Big(\frac{1}{r^{(n)}} I\{(\mathbf{a},\mathbf{b}) \in R\} \Big) \\
&= \frac{(p^{(n)})^2}{(r^{(n)})^2} \operatorname{Var}_{F_{c_n}} \big(I\{(\mathbf{a},\mathbf{b}) \in R\} \big) \\
&= \frac{(p^{(n)})^2}{(r^{(n)})^2} \mathbf{P}\big(I\{(\mathbf{a},\mathbf{b}) \in R\} \mid X_m > c_n \big) \mathbf{P}\big(I\{(\mathbf{a},\mathbf{b}) \notin R\} \mid X_m > c_n \big) \\
&= \frac{p^{(n)}}{r^{(n)}} \mathbf{P}\Big(1 - \frac{r^{(n)}}{p^{(n)}} \Big) \to 0.
\end{aligned}$$

This completes the proof. □

3.3.1 Numerical experiments

The theoretical results guarantee that $\widehat{q}_T^{(n)}$ is an efficient estimator of $1/p^{(n)}$. However, for comparison with existing importance sampling algorithms the numerical experiments are based on $\widehat{p}_T^{(n)} = (\widehat{q}_T^{(n)})^{-1}$ as an estimator for $p^{(n)}$. The literature includes numerical experiments for the technique proposed by Blanchet, Hult and Leder in [14], who propose a sequential importance sampling algorithm based on conditional mixtures.

We consider the problem of computing $\mathbf{P}(X_m > c)$ where X_m is the solution to the recurrence equation $X_k = A_k X_{k-1} + B_k$, for $k \geq 1$ with $X_0 = 0$. The innovation B is assumed to be non-negative Pareto distributed variable with tail, $\overline{F}_B(x) = (x+1)^{-2}$ for $x \geq 0$.

The performance of the MCMC estimator is compared to the importance sampling estimator in [14], which is labeled IS. By construction each simulation run of the MCMC algorithm only generates a single random variable (one simulation step) while the importance sampling algorithm generate $2m$ number of random variables ($2m$ simulation steps). Therefore the number of runs for the MCMC is scaled up by a factor of $2m$ so that the two algorithms generate essentially the same number of random numbers, and thus getting a fairer comparison of the computer runtime.

Table 3.1 presents the estimates based on 10 batches, each consisting of 10^5 simulations in the case of the importance sampling and $2m \cdot 10^5$ in the case of MCMC. The results are divided into two classes. In the first one, the log-normal setting, the A is assumed to be log-normally distributed with $\sigma = 0.1$ and $\mu = \log(1.05) - \sigma^2/2$. In the second one, the exponential setting, the A is assumed to be exponentially distributed with mean 0.25. Table 3.1 shows that MCMC appears to perform reasonably well compared to importance sampling, but there seems to be a bias, which is believed to come from slow convergence of the MCMC sampler.

The numerical experiments show that the convergence of the MCMC algorithm is sensitive to the asymptotic choice of a in equation (3.9). In addition burn-in time tends to be extensive because the chain gets stuck where B_m is large. The slow convergence

due to long burn-in time is illustrated in Table 3.2 which presents the estimates based on 10 batches, each consisting of T simulations. This problem could be overcome with an MCMC sampler with better mixing properties than the one given by Algorithm 3.3.1.

Table 3.1: Numerical comparison of computing $\mathbf{P}(X_4 > c)$.

	Log-normal setting		Exponential setting	
$c = 10$	MCMC	IS	MCMC	IS
Estimate	2.390e-02	6.645e-02	1.044e-02	1.042e-02
Std. deviation	5.033e-04	15.422e-04	2.477e-04	1.838e-04
Rel. error	0.0211	0.0232	0.0237	0.0176
Comp. time(s)	4.79	4.09	4.29	4.63
$c = 100$	MCMC	IS	MCMC	IS
Estimate	1.788e-04	4.947e-04	1.148e-04	1.134e-04
Std. deviation	1.984e-05	1.614e-05	18.290e-06	5.744e-06
Rel. error	0.1109	0.0326	0.1593	0.0507
Comp. time(s)	4.71	4.08	4.14	4.53
$c = 1,000$	MCMC	IS	MCMC	IS
Estimate	1.084e-06	4.756e-06	1.045-06	1.140e-06
Std. deviation	21.184e-08	3.921e-08	8.880e-08	1.459e-08
Rel. error	0.1953	0.0082	0.0850	0.0128
Comp. time(s)	4.09	4.02	4.04	4.53
$c = 10,000$	MCMC	IS	MCMC	IS
Estimate	1.017e-08	4.734e-08	1.034e-08	1.142e-08
Std. deviation	4.169e-10	4.037e-10	8.487e-13	1100.195e-13
Rel. error	0.0410	0.0085	8.205e-05	963.3e-05
Comp. time(s)	4.17	4.15	4.02	4.53
$c = 100,000$	MCMC	IS	MCMC	IS
Estimate	1.012e-10	4.738e-10	1.034e-10	1.143e-10
Std. deviation	7.822e-15	4053.249e-15	1.983e-15	1390.922e-15
Rel. error	7.726e-05	855.431e-05	1.917e-05	1217.112e-05
Comp. time(s)	4.15	4.16	4.16	4.67

Table 3.2: Burn-in time effect of computing $\mathbf{P}(X_4 > c)$ in Exponential setting.

	MCMC	IS	MCMC	IS	MCMC	IS
$c = 1,000$	$T = 10,000$		$T = 100,000$		$T = 1,000,000$	
Estimate	1.033e-06	1.144e-06	1.033e-06	1.141e-06	1.197e-06	1.141e-06
Std. deviation	2.405e-10	1.154e-08	1.044e-10	3.531e-09	1.734e-07	1.009e-09
Rel. error	2.327e-04	1.008e-02	1.010e-04	3.094e-03	1.449e-01	8.840e-04

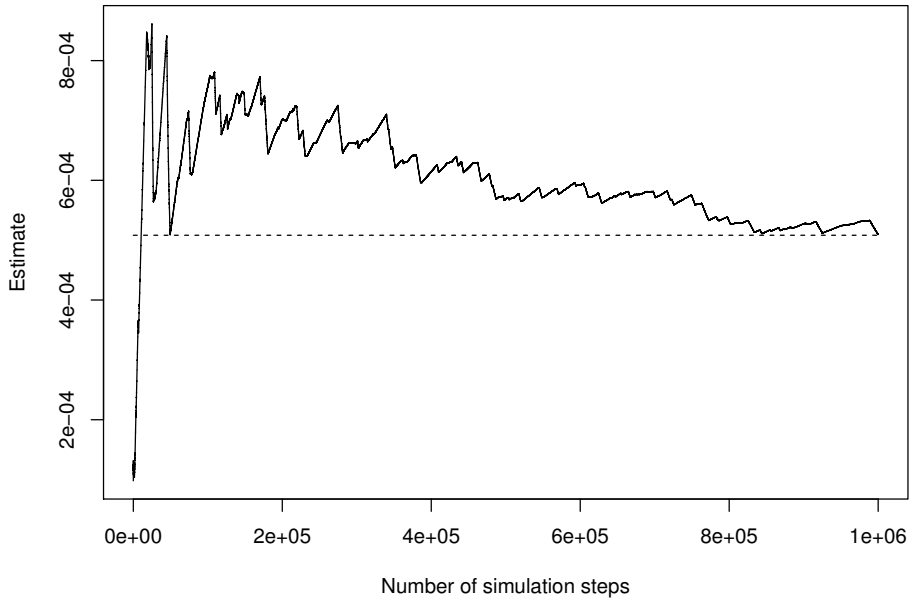
Finally, the convergence of the MCMC technique is visualised in Figure 3.1 which illustrates the point estimate of $\mathbf{P}(X_4 > 25)$ as a function of number of simulation steps. In the figure, the A is assumed to be log-normally distributed with $\sigma = 0.1$ and $\mu = \log(1.05) - \sigma^2/2$.

3.4 Insurance company with risky investments

In this section the MCMC estimator introduced in Section 3.2 is applied for computing the probability of ruin for an insurance company with risky investments.

The ruin problem with investment is reasonably well studied. A recent overview is given by Paulsen in [22]. In the infinite horizon setting there are two asymptotic regimes. Power tail asymptotics can arise either as the cumulative effect of negative returns on the

Figure 3.1: The point estimate of $\mathbf{P}(X_4 > 25)$ as a function of simulations (solid line) compared to the Monte Carlo estimate (dotted line) using 10^7 simulations.



investment asset or because of power tails of the claim size distribution. In the first case the power tail asymptotics can be derived by expressing the risk reserve as the solution to a stochastic recurrence equation whose stationary solution has a power tail, see e.g. [21], [16], [11], [23] and [18]. In the second case the power tail asymptotics of the ruin probability is more directly inferred from the power tail of the claim size distribution, see e.g. [12] and [18]. This paper assumes the latter of the two settings and in finite time. Here the occurrence of the ruin event is dominated by the power tail asymptotics of the claim size distribution.

Consider a discrete time model for the risk reserve of an insurance company motivated by the Solvency II regulatory framework. Given a fixed integer $m \geq 1$, let $0 = t_0 < t_1 < \cdots < t_m = 1$ be a partition of the time interval $[0, 1]$. Assume that the company receives its premium c at the start of the period $t = 0$ and is therefore not affected by the stochastic return during the period. Denote by B_k the claim loss during the kth subperiod $(t_{k-1}, t_k]$. Suppose that the insurance company invests the risk reserve in a risky asset and denote by R_k the stochastic return on the risky asset over the kth subperiod $(t_{k-1}, t_k]$. It is assumed that $\{B_k\}_{k=1}^m$ and $\{R_k\}_{k=1}^m$ are independent sequences, each consisting of independent and identically distributed random variables. The risk reserve, U_k, at the end of the kth

period, given an initial wealth of $U_0 = u_n$, is modeled as

$$\begin{aligned} U_k &= R_k(Z_{k-1} - B_k), \quad \text{for } k = 1, \ldots, m, \\ U_0 &= u_n. \end{aligned}$$

Iterating the relation above yields

$$U_m = R_m \cdots R_1 u_n - \big(R_m \cdots R_1 B_1 + R_m \cdots R_2 B_2 + \cdots + R_m B_m\big).$$

Assume that $R_k > 0$ almost surely for all $k = 1, \ldots, m$ and put $A_k = 1/R_k$. The last display is equivalent to

$$A_1 \cdots A_m U_m = u_n - W_m,$$

where

$$W_m = B_1 + A_1 B_2 + \cdots + A_1 \cdots A_{m-1} B_m.$$

Observe that W_m represents the discounted losses that have accumulated up until time $t_m = 1$. Since the only difference between the losses W_m, and X_m from the previous section, is the labeling of the identically distributed and independent B's, then it holds that $W_m \stackrel{d}{=} X_m$. The event of ruin up until time $t_m = 1$ is equivalent to

$$\left\{ \inf_{0 \le k \le m} U_k < 0 \right\} = \left\{ \sup_{0 \le k \le m} W_k > u_n \right\} = \{W_m > u_n\}.$$

The objective is to compute the ruin probability

$$p^{(n)} = \mathbf{P}(W_m > u_n),$$

where the last display is equal to $\mathbf{P}(X_m > u_n)$. Applying the method presented in Section 3.3 thus enables us to compute the probability of ruin with minor notational adjustments.

3.4.1 Numerical experiments

In this section the MCMC estimator from Section 3.3 is exemplified on the the problem of computing the probability of ruin $p^{(n)} = \mathbf{P}(W_m > u_n)$.

It is assumed that the premium c is received at the start of the period $t = 0$ and hence not affected by the stochastic return. This translates into the condition that $u_n \ge c$. Let the accumulated claim size distribution be given by,

$$B_k = \lambda Z_k, \quad \text{for } k = 1, \ldots, m,$$

where $\lambda \in (0, \infty)$ represents the intensity of the number of claims and Z_k the claim amount. Assume that the distribution of Z, the generic representative of the collection $\{Z_k\}$, has the following distribution

$$\mathbf{P}(Z \le z) = 1 - \kappa^\alpha (\kappa + z)^{-\alpha}, \quad \text{for } z \ge 0. \tag{3.11}$$

Then the distribution of B has a regularly varying tail with index $-\alpha < 0$. More precisely,

$$\lim_{t\to\infty} \frac{\mathbf{P}(B > t)}{\mathbf{P}(Z > t)} = \lambda.$$

For the annual returns R, a log-normal distribution will be assumed with mean $\mu - \frac{\sigma^2}{2}$ and standard deviation σ. In this case, for a standard normal variable Y:

$$\mathbf{E}[R^{-\alpha}] = \mathbf{E}\Big[\exp\Big\{-\alpha\Big(\mu - \frac{\sigma^2}{2} - \sigma Y\Big)\Big\}\Big] = \exp\Big\{\alpha\Big(\frac{\sigma^2}{2}(\alpha+1) - \mu\Big)\Big\}.$$

In the numerical experiments the time interval is one year, partitioned into $m = 12$ time units. The intensity of the number of claims is $\lambda = 197$ per year, and the parameters for the claim size distribution (3.11) are $\kappa = 2.20$, $\alpha = 1.7$. These parameter values are consistent with the well studied Danish Fire Insurance data, see [19], where the claim amounts are in Million Danish Kroner. Let the stochastic return be specified with yearly return $\mu = 1.06$ and yearly standard deviation of $\sigma = 0.2$.

The results of the numerical experiments using the MCMC technique are provided in Table 3.3. All the simulations are based on 10 batches, each consisting of $10,000$ replications.

From the table it is clear that the relative error stays bounded and that the computational runtime is not correlated to the threshold u.

Table 3.3: Estimation of $\mathbf{P}(W_{12} > u)$ Log-normal setting.

u	Estimate	Std. deviation	Rel. error	Comp. time(s)
1,000	4.381e-03	1.766e-04	0.0403	20.03
5,000	3.180e-04	9.472e-05	0.2978	20.17
10,000	9.085e-05	3.789e-05	0.4171	19.55
100,000	1.406e-06	2.482e-10	0.0002	17.77

In addition, the convergence of the MCMC technique for computing the probability is visualised in Figure 3.2 which illustrates the point estimate of $\mathbf{P}(W_{12} > 10^3)$ as a function of number of simulation steps.

Acknowledgments

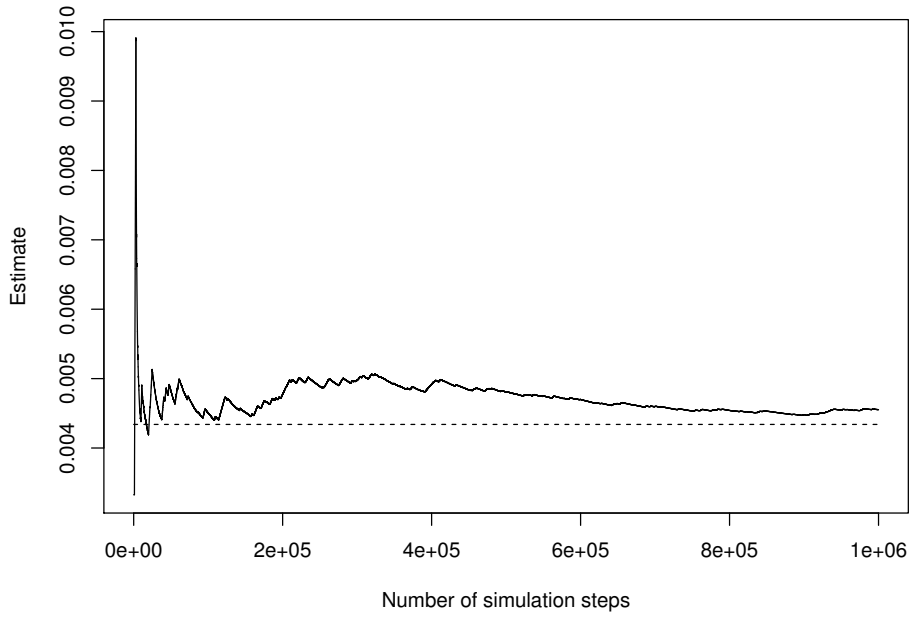
Henrik Hult's research was supported by the Göran Gustafsson Foundation

3.5 References

[1] S. Asmussen and K. Binswanger. Simulation of ruin probabilities for subexponential claims. *Astin Bulletin*, 27(2):297–318, 1997.

[2] S. Asmussen, K. Binswanger, and B. Hojgaard. Rare events simulation for heavy-tailed distributions. *Bernoulli*, 6(2):303–322, 2000.

Figure 3.2: The point estimate of $\mathbf{P}(W_{12} > 10^3)$ as a function of simulations (solid line) compared to the Monte Carlo estimate (dotted line) using 10^6 simulations.



[3] S. Asmussen and P. W. Glynn. *Stochastic Simulation*, volume 57 of *Stochastic Modelling and Applied Probability*. Springer, New York, 2007.

[4] S. Asmussen and D. P. Kroese. Improved algorithms for rare event simulation with heavy tails. *Adv. Appl. Probab.*, 38:545–558, 2006.

[5] B. Basrak, R.A. Davis, and T. Mikosch. Regular variation of GARCH processes. *Stoch. Proc. Appl.*, 99:95–115, 2002.

[6] J. Blanchet and C. Li. Efficient rare-event simulation for heavy-tailed compound sums. *ACM T. Model Comput. S.*, 21:1–10, 2011.

[7] J. Blanchet and J. C. Liu. State-dependent importance sampling for regularly varying random walks. *Adv. Appl. Probab.*, 40:1104–1128, 2008.

[8] J.A. Bucklew. *Introduction to Rare Event Simulation*. Springer, New York, 2004.

[9] J.F. Collamore. Random recurrence equations and ruin in a Markov-dependent stochastic economic environment. *Ann. Appl. Probab*, 19(4):1404–1458, 2009.

[10] P. Dupuis, K. Leder, and H. Wang. Importance sampling for sums of random variables with regularly varying tails. *ACM T. Model Comput. S.*, 17(3), 2007.

[11] A. Frolova, Y. Kabanov, and S. Pergamenshchikov. In the insurance business risky investments are dangerous. *Finance Stoch.*, 6:227–235, 2002.

[12] J. Gaier and P. Grandits. Ruin probabilities in the presence of regularly varying tails and optimal investment. *Insur. Math. Econ.*, 30:211–217, 2002.

[13] T. Gudmundsson and H. Hult. Markov chain Monte Carlo for computing rare-event probabilities for a heavy-tailed random walk. *J. Appl. Probab.*, 51(2), June 2014.

[14] H. Hult, Blanchet. J., and K. Leder. Rare-event simulation for stochastic recurrence equations with heavy-tailed innovations. *ACM T. Model. Comput. S.*, 2013.

[15] S. Juneja and P. Shahabuddin. Simulation heavy-tailed processes using delayed hazard rate twisting. *ACM T. Model Comput. S.*, 12(2):94–118, April 2002.

[16] V. Kalashnikov and R. Norberg. Power tailed ruin probabilities in the presence of risky investments. *Stoch. Process. Appl.*, 98:211–228, 2002.

[17] H. Kesten. Random difference equations and renewal theory for products of random matrices. *Acta Math.*, 131(1):207–248, 1973.

[18] C. Klüppelberg and R. Kostadinova. Integrated risk models with exponential Lévy investment. *Insur. Math. Econ.*, 42:560–577, 2008.

[19] T. Mikosch. *Non-life Insurance Mathematics*. Springer, New York, 2009.

[20] E. Nummelin. *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge University Press, 1984.

[21] H. Nyrhinen. On the ruin probabilities in a general economic environment. *Stoch. Process. Appl.*, 83:319–330, 1999.

[22] J. Paulsen. Ruin models with investment income. *Probab. Surv.*, 5:416–434, 2008.

[23] S. Pergamenshchikov and O. Zeitouni. Ruin probability in the presence of risky investments. *Stoch. Process. Appl.*, 116:267–278, 2006.

Markov chain Monte Carlo for rare-event simulation for light-tailed random walk

Markov chain Monte Carlo for rare-event simulation for light-tailed random walk

by

Thorbjörn Gudmundsson and Henrik Hult

Abstract

In this paper the Markov chain Monte Carlo (MCMC) algorithm for computing rare-event probabilities is developed and extended to the setting of light-tailed random walk. A Markov chain is generated using a Gibbs sampler, having as its invariant distribution the conditional distribution given that the event of interest occurs. The sought probability is the normalising constant of that conditional distribution and is estimated from the sample of the Markov chain.

The MCMC algorithm is presented for two light-tailed random walk problems, when the support of the distribution of the increments is $\mathbb{R}$ and $\mathbb{R}_+$, respectively. The logarithmic efficiency of the estimator is characterised and the main result of this paper is theorem which states that under certain condition the MCMC estimator is logarithmically efficient. Numerical experiments are performed to compare the MCMC algorithm to an existing strongly efficient importance sampling algorithm.

4.1 Introduction

This paper provides a Markov chain Monte Carlo (MCMC) algorithm for computing rare-event probabilities for a light-tailed random walk. The fundamental idea is to construct a Markov chain via an MCMC sampler, having as its invariant distribution the conditional distribution given that the event of interest occurs. The probability of the event of interest then appears as the normalising constant of that conditional distribution and it is estimated from a sample of the Markov chain.

The importance of evaluating the risk of disastrous events grows in many areas such as economics, finance and insurance. The evaluation often boils down to the computation of probabilities or estimation of quantiles. For complex models, there is often no analytical solution known to calculate such probabilities. This has motivated the use of stochastic simulation as an alternative for computing probabilities.

Consider a sequence of random variables $X^{(1)}, X^{(2)}, \ldots$, each of which can be sampled repeatedly by a simulation algorithm. The objective is to compute $p^{(n)} = \mathbf{P}(X^{(n)} \in A)$ for some large n and it is assumed that $p^{(n)} \to 0$ as $n \to \infty$. In this sense the event $\{X^{(n)} \in A\}$ can be thought of as a rare-event. For a sample $X_0^{(n)}, \ldots, X_{T-1}^{(n)}$ the Monte Carlo estimate is $\widehat{p}_T^{(n)} = \frac{1}{T}\sum_{t=0}^{T-1} I\{X_t^{(n)} \in A\}$ and has the relative error

$$\frac{\mathrm{Var}(\widehat{p}^{(n)})}{(p^{(n)})^2} = \frac{p^{(n)}(1-p^{(n)})}{T(p^{(n)})^2} = \frac{1}{T}\Big(\frac{1}{p^{(n)}} - 1\Big) \to \infty, \quad \text{as } n \to \infty,$$

thus indicating that the performance deteriorates when the event is rare.

The rare-event performance of an estimator is quantified by its relative error. A desired property is strong efficiency. An estimator is said to be strongly efficient if its relative error is bounded or tends to zero as $n \to \infty$. A slightly weaker property is logarithmic efficiency. Suppose, for simplicity, $\{p^{(n)}\}$ satisfies a large deviation principle with rate function I, in particular

$$\lim_{n\to\infty} -\frac{1}{n}\log p^{(n)} = I(A).$$

Informally this can interpreted that $p^{(n)} \approx e^{-nI(A)}$. Logarithmic efficiency means that the variance of the estimator (dominated by its second moment) is roughly of the same size as the probability squared, that is $\mathbf{E}[(\widehat{p}^{(n)})^2] \approx e^{-2nI(A)}$. More formally, an estimator is said to be logarithmically efficient if

$$\lim_{n\to\infty} \frac{1}{n}\log\frac{\mathbf{E}[(\widehat{p}^{(n)})^2]}{(p^{(n)})^2} = 0.$$

Importance sampling (IS) is a popular method for improving the rare-event efficiency of the standard Monte Carlo algorithm. The basic idea is to sample the variables $X_0^{(n)}, \ldots, X_{T-1}^{(n)}$ from a different distribution, say $G^{(n)}$, rather than the original distribution $F^{(n)}$. The IS estimate is $\widehat{p}_T^{(n)} = \frac{1}{T}\sum_{t=0}^{T-1} \frac{dF^{(n)}}{dG^{(n)}}(X_t^{(n)}) I\{X_t^{(n)} \in A_n\}$ and its performance is determined by the choice of the sampling distribution $G^{(n)}$. The optimal sampling distribution is the conditional distribution

$$F_A^{(n)}(\cdot) = \mathbf{P}(X^{(n)} \in \cdot \mid X^{(n)} \in A).$$

For this choice of a sampling distribution, the indicators in the IS estimate are all equal to one, thus implying that $\widehat{p}_T^{(n)}$ has zero variance. But $F_A^{(n)}$ is infeasible as a sampling distribution since it requires the knowledge of $p^{(n)}$. This motivates choosing for a sampling distribution $G^{(n)}$ a distribution which is 'close to' $F_A^{(n)}$. It has long been known that the design of the sampling distribution is closely related to the large deviations asymptotics of the underlying model in the sense that the change of measure should be such that under the sampling distribution the system tends to follow the most likely trajectory to the rare event. Recent developments have lead to a more systematic approach towards designing efficient IS algorithms. Indeed, the design of an efficient sampling distribution for an IS algorithm

is closely related to finding a classical subsolution to the partial differential equation, of Hamilton-Jacobi type, that characterises the rate function, see e.g. [3] and [4].

In previous papers we have presented the MCMC methodology for problems in heavy-tailed settings such as random walk and random sums with increments having heavy-tailed distributions and stochastic recurrence equations with heavy-tailed innovations. The MCMC method has proven to be very efficient, in particular for random walk and random sums. In addition, efficiency proofs for these examples have been short and simple, in contrast to the lengthy and technical proofs of the IS counterpart. The motivation for this paper is to investigate the applicability of the MCMC methodology in light-tailed settings. A natural starting point is to consider light-tailed random walk. Since efficient importance sampling algorithms exists and are well established the setting of random walks allows for comparison with state-of-the-art techniques.

This paper presents the MCMC methodology for computing rare-event probabilities and how logarithmic efficiency can be attained. The algorithm is exemplified by considering the problem of computing the probability that a random walk exceeds a high threshold. The probabilistic assumptions made in this paper are that the increments of the random walk are independent and identically light-tailed distributed random variables. The logarithmic efficiency is characterised for two separate cases, when the distribution of the increments is supported on $\mathbb{R}$ and $\mathbb{R}_+$, respectively. Numerical experiments are performed on the proposed MCMC algorithm and compared to the strongly efficient IS algorithm presented in [1]. Comparison shows that MCMC estimator performs comparably with the importance sampling algorithm, both when the increments are supported on $\mathbb{R}$ and on $\mathbb{R}_+$.

The literature includes several examples of logarithmically efficient importance sampling algorithms. These techniques are based on finding the optimal exponential tilting of the sampling distribution, from which all of the summands are sampled, see for instance Siegmund [7] and Sadowsky [6]. Blanchet, Leder and Glynn [1], provide an importance sampling algorithm for a light-tailed random walk which is proved to be strongly efficient. This is acquired by designing a state-dependent optimal exponential tilting, whereas the summands need not to be sampled from the same tilted sampling distribution.

The paper is organised as follows. The general MCMC methodology for computing rare-event probabilities and the first steps for proving logarithmic efficiency is presented in Section 4.2. Section 4.3 provides two examples of the algorithm in the setting of a light-tailed random walk, where the increments have support on the whole $\mathbb{R}$ and where the increments only have support on $\mathbb{R}_+$. Numerical experiments are presented in Section 4.4 to compare the numerical efficiency of the MCMC estimator against the strongly efficient IS estimator provided in [1].

4.2 Logarithmically efficient MCMC simulation

In this section the MCMC simulation method for computing rare-event probabilities is presented and the first steps towards for proving logarithmic efficiency explained.

Let $X^{(n)}$ be a random element with probability distribution $F^{(n)}$ and let A be a mea-

surable set. Consider the problem of computing

$$p^{(n)} = \mathbf{P}(X^{(n)} \in A),$$

where $\{X^{(n)} \in A\}$ is rare-event in the sense that $p^{(n)}$ is small. Denote by $F_A^{(n)}$ the probability distribution conditioned on the rare-event, namely

$$F_A^{(n)}(\cdot) = \mathbf{P}(X^{(n)} \in \cdot \mid X^{(n)} \in A).$$

The first step is designing an MCMC sampler which produces a Markov chain $(X_t^{(n)})_{t\geq 0}$ having $F_A^{(n)}$ as its invariant distribution. Assuming that such a sampler exists, let $X_0^{(n)}, \ldots, X_{T-1}^{(n)}$ be given having $F_A^{(n)}$ as its invariant distribution.

The next step is to construct an unbiased estimator $\widehat{q}_T^{(n)}$ of $1/p^{(n)}$ based on $X_0^{(n)}, \ldots, X_{T-1}^{(n)}$. Consider a probability distribution $V^{(n)}$, such that $V^{(n)} \ll F_A^{(n)}$, and define u, a function of $X^{(n)}$, by

$$u(X^{(n)}) = \frac{dV^{(n)}}{dF^{(n)}}(X^{(n)}).$$

Taking expectation with respect to F_A then

$$\mathbf{E}_{F_A^{(n)}}\big[u(X^{(n)})\big] = \int \frac{dV^{(n)}}{dF^{(n)}}(X^{(n)}) dF_A^{(n)}(X^{(n)}) = \frac{1}{p^{(n)}} \int dV^{(n)}(X^{(n)}) = \frac{1}{p^{(n)}},$$

thus motivating the following definition of the MCMC estimator $\widehat{q}_T^{(n)}$ of $1/p^{(n)}$

$$\widehat{q}_T^{(n)} = \frac{1}{T} \sum_{t=0}^{T-1} u(X_t^{(n)}),$$

based on the sample of a Markov chain $(X_t^{(n)})_{t=0}^{T-1}$ having $F_A^{(n)}$ as its invariant distribution.

For an unbiased estimator $\widehat{q}_T^{(n)}$ of $1/p^{(n)}$ we emphasise two efficiency concepts. Firstly, it is desirable that the relative error of the estimator is controlled. In the light-tailed setting of this paper the estimator $\widehat{q}_T^{(n)}$ of $1/p^{(n)}$ is said to be logarithmically efficient if

$$\lim_{n\to\infty} \frac{1}{n} \log(p^{(n)})^2 \mathbf{E}\big[(\widehat{q}_T^{(n)})^2\big] = 0,$$

Secondly, it is desirable that the variance of the estimator tends to zero as the sample size increases. An estimator $\widehat{q}_T^{(n)}$ is said to be large sample efficient if

$$\lim_{T\to\infty} \mathrm{Var}(\widehat{q}_T^{(n)}) = 0.$$

Two key choices arise that determine the convergence performance of the algorithm. The design of the MCMC sampler and the choice of $V^{(n)}$. The design of the MCMC sampler is essential for the mixing properties of the Markov chain $(X_t^{(n)})_{t\geq 0}$ and thereby

the convergence of the estimator as the sample size increases. The higher the mixing rate of the MCMC sampler, the quicker the Markov chain will converge to its invariant distribution. It has been shown, for instance in [5], that it is sufficient that the Markov chain is geometric ergodic to guarantee large sample efficiency, in the sense that $\mathrm{Var}_{F_A^{(n)}}(\widehat{q}_T^{(n)}) = O(1/T)$.

The choice of $V^{(n)}$ controls the normalised variance of the estimator

$$(p^{(n)})^2 \,\mathrm{Var}_{F_A^{(n)}}(\widehat{q}_T^{(n)}),$$

thus determining the rare-event efficiency of the algorithm. Suppose that the increments Y are light-tailed, meaning that $\mathbf{E}[e^{\theta Y_1}] < \infty$ for some $\theta > 0$. Then it follows by Cramér's theorem that $\{X^{(n)}\}$ satisfies the large deviation principle

$$I(\mathring{A}) \le \liminf_{n\to\infty} -\frac{1}{n}\log p^{(n)} \le \limsup_{n\to\infty} -\frac{1}{n}\log p^{(n)} \le I(\bar{A}), \tag{4.1}$$

where $\mathring{A}$ and $\bar{A}$ are the interior and closure of A, respectively, and $I(A) = \inf_{x\in A} I(x)$ where

$$I(x) = \sup_{\theta\in\mathbb{R}} \big\{\theta x - \Lambda(\theta)\big\},$$

is the rate-function and $\Lambda(\theta) = \log \mathbf{E}[e^{\theta Y_1}]$ is the moment generating function of the increment. The objective is to prove logarithmic efficiency, that is

$$\limsup_{n\to\infty} \frac{1}{n}\log\Big\{(p^{(n)})^2 \mathbf{E}_{F_A^{(n)}}\big[(\widehat{q}_T^{(n)})^2\big]\Big\} = 0.$$

The second moment of the estimator can be rewritten

$$\begin{aligned}\mathbf{E}_{F_A^{(n)}}\big[(\widehat{q}_T^{(n)})^2\big] &= \mathbf{E}_{F_A^{(n)}}\Big[\Big(\frac{1}{T}\sum_{t=0}^{T-1} u(X_t^{(n)})\Big)^2\Big] \\ &= \frac{1}{T}\mathbf{E}_{F_A^{(n)}}\big[u(X^{(n)})^2\big] + \frac{2}{T^2}\sum_{t=0}^{T-1}\sum_{s=t+1}^{T-1} \mathbf{E}_{F_A^{(n)}}\big[u(X_t^{(n)})u(X_s^{(n)})\big]\end{aligned}$$

The crude upper bound $\mathbf{E}_{F_A^{(n)}}\big[u(X_t^{(n)})u(X_s^{(n)})\big] \le \mathbf{E}_{F_A^{(n)}}\big[u(X^{(n)})^2\big]$ implies that

$$\mathbf{E}_{F_A^{(n)}}\big[(\widehat{q}_T^{(n)})^2\big] \le \Big(\frac{2}{T}+1\Big)\mathbf{E}_{F_A^{(n)}}\big(u(X^{(n)})^2\big).$$

Therefore it is clear that the normalised variance of the estimator is controlled by the second moment of u. Now consider,

$$(p^{(n)})^2 \mathbf{E}_{F_A^{(n)}}\big(u(X^{(n)})^2\big) = \mathbf{E}_{V^{(n)}}\Big[\frac{dV^{(n)}}{dF_A^{(n)}}(X^{(n)})\Big],$$

thus motivating taking $V^{(n)}$ as an approximation of $F_A^{(n)}$. Observe that

$$\frac{1}{n}\log \mathbf{E}_{V^{(n)}}\Big[\frac{dV^{(n)}}{dF_A^{(n)}}(X^{(n)})\Big] = \frac{1}{n}\log p^{(n)} + \frac{1}{n}\log \mathbf{E}_{V^{(n)}}\Big[\frac{dV^{(n)}}{dF^{(n)}}(X^{(n)})\Big],$$

so it follows from the large deviation assumption in (4.1) that the objective can be rewritten, so to prove logarithmic efficiency it is sufficient to show that

$$\limsup_{n\to\infty} \frac{1}{n}\log \mathbf{E}_{V^{(n)}}\Big[\frac{dV^{(n)}}{dF^{(n)}}(X^{(n)})\Big] \leq I(A). \tag{4.2}$$

In the next section we consider a specific problem and provide a complete characterisation of logarithmic efficiency.

4.3 Light-tailed random walk

In this section the MCMC simulation technique is developed for computing the probability that a light-tailed random walk exceeds a high threshold. Two cases are considered, firstly, when the distribution of the increments is supported on the whole $\mathbb{R}$ and, secondly, when the distribution of the increments is supported on $\mathbb{R}_+$. To start, a Gibbs sampler is presented which generates a Markov chain with desirable invariant distribution. The Gibbs sampler is common for both cases and is therefore presented first before the section is split where each case is presented separately.

Let $\mathbf{Y}^{(n)} = (Y_1, \ldots, Y_n)$ be a vector of independent and identically distributed random variables with joint probability distribution function $F^{(n)}$. Denote by F_Y the probability distribution for a generic Y and f_Y for its density function with respect to the Lebesgue measure, which is assumed to exist. Consider the problem of computing the probability that the sample mean $S_n = (Y_1 + \cdots + Y_n)/n$ is larger than $\mathbf{E}[Y_1]$,

$$p^{(n)} = \mathbf{P}(S_n > a),$$

for $a > \mathbf{E}[Y_1]$.

The first step is designing a Gibbs sampler that produces a Markov chain $(\mathbf{Y}_t^{(n)})_{t\geq 0}$ having the conditional distribution

$$F_a^{(n)}(\cdot) = \mathbf{P}(\mathbf{Y}^{(n)} \in \cdot \mid S_n > a),$$

as its invariant distribution. In the following algorithm the steps are updated sequentially, such that the sample mean is conditioned on exceeding the threshold after the update. To simplify notation, write $S_{n,-k} = Y_1 + \cdots + Y_{k-1} + Y_{k+1} + \cdots + Y_n$.

Algorithm 4.3.1. Start with an initial state $\mathbf{Y}_0^{(n)} = (Y_{0,1}, \ldots, Y_{0,n})$ where

$$S_{0,n} = \frac{1}{n}\sum_{k=1}^{n} Y_{0,k} > a.$$

Given $\mathbf{Y}_t^{(n)}$ for some $t \geq 0$, the next state $\mathbf{Y}_{t+1}^{(n)}$ is sampled as follows.

1. Draw a randomised ordering $j_1, \ldots, j_n$ of $\{1, \ldots, n\}$ and proceed updating $\mathbf{Y}_t^{(n)}$ in the order thus obtained.

2. For $l = 1, \ldots, n$ set $k = j_l$ and proceed updating Y_k as follows. Sample Y' from the conditional distribution

$$\mathbf{P}(Y' \in \cdot \mid Y' > s),$$

where

$$s = a - S_{t,n,-k}. \tag{4.3}$$

Put $\mathbf{Y}_{t+1}^{(n)} = (Y_{t,1}, \ldots, Y_{t,k-1}, Y', Y_{t,k+1}, \ldots, Y_{t,n})$.

Iterate steps 1 and 2 until the entire Markov chain $(\mathbf{Y}_t^{(n)})_{t=0}^{T-1}$ is constructed.

Remark 4.3.2. In the case when the increments can only take positive values, then the threshold parameter s in (4.3) is $(a - S_{t,n,-k}, 0) \vee 0$.

Proposition 4.3.3. *The Markov chain $(\mathbf{Y}_t^{(n)})_{t\geq 0}$ constructed by Algorithm 4.3.1 has the conditional distribution $F_a^{(n)}$ as its invariant distribution and is uniformly ergodic.*

The proof of this proposition is identical to the proofs of Propositions 3.1 and 3.2 in [5] and is therefore omitted.

This section now splits into two cases, firstly, when the support of the increments Y is $\mathbb{R}$ and, secondly, when the support of the increments is $\mathbb{R}_+$. In the first case, we choose $V^{(n)}$ to be the joint distribution of n normal variables conditioned on the average being larger than the threshold a. The reason for this choice is that the probability of the event conditioned on can easily be computed explicitly. In the second case, we choose $V^{(n)}$ to be the joint distribution of n gamma variables conditioned on the average being larger than a. Again, the reason for this choice is that the probability of the event conditioned on is easily computed explicitly.

4.3.1 Real-valued increments

Assume that $Y_1, Y_2, \ldots$ are independent and identically distributed random variables with density f_Y supported on the whole of $\mathbb{R}$. Let $V^{(n)}$ be the probability distribution given by

$$V^{(n)}(\cdot) = \mathbf{P}\big((Z_1, \ldots, Z_n) \in \cdot \mid (Z_1, \ldots, Z_n) \in R^{(n)}\big),$$

for $R^{(n)} = \{(Z_1 + \cdots + Z_n)/n > a\}$, where the $Z_1, Z_2, \ldots$ are independent and normally distributed variables with mean $\mu = E[Y_1]$ and standard deviation σ. The probability of the event $R^{(n)}$ can be computed explicitly as

$$\begin{aligned} r^{(n)} &= \mathbf{P}\big((Z_1, \ldots, Z_n) \in R^{(n)}\big) \\ &= \mathbf{P}\left(\frac{(Z_1 + \cdots + Z_n)/n - \mu}{\sigma/\sqrt{n}} > \frac{a - \mu}{\sigma/\sqrt{n}}\right) = 1 - \Phi\Big(\frac{a - \mu}{\sigma/\sqrt{n}}\Big), \end{aligned}$$

where Φ is the standard normal probability distribution.

Let $(\mathbf{Y}_t^{(n)})_{t=0}^{T-1}$ be generated via Algorithm 4.3.1 with invariant distribution $F_a^{(n)}$. For the given choice of $V^{(n)}$ it follows that

$$u(\mathbf{Y}^{(n)}) = \frac{1}{r^{(n)}} \prod_{i=1}^{n} \frac{\phi_{\mu,\sigma}(Y_{t,i})}{f_Y(Y_{t,i})} I\Big\{\frac{1}{n}\sum_{i=1}^{n} Y_{t,i} > a\Big\},$$

where $\phi_{\mu,\sigma}$ is the density of a normal distribution with mean μ and standard deviation σ. The MCMC estimator is given by

$$\widehat{q}_T^{(n)} = \frac{1}{r^{(n)}} \frac{1}{T} \sum_{t=0}^{T-1} \prod_{i=1}^{n} \frac{\phi_{\mu,\sigma}(Y_{t,i})}{f_Y(Y_{t,i})} I\Big\{\frac{1}{n}\sum_{i=1}^{n} Y_{t,i} > a\Big\}. \tag{4.4}$$

Observe that the estimator can be viewed as the asymptotic approximation $1/r^{(n)}$ of the true $1/p^{(n)}$ times the stochastic correction factor. The reason for the good efficiency of the MCMC estimator relies on the fact that the asymptotic approximation is good and the stochastic correction factor has small normalised variance.

Recall that the distribution $V^{(n)}$ should be chosen so that $r^{(n)}$ is close to $p^{(n)}$ on a logarithmic scale. The standard deviation σ of the normal variable Z_1 is therefore set so that the large deviation rate of $r^{(n)}$ match the large deviation rate of $p^{(n)}$.

From the assumption that $\Lambda(\theta) = \log \mathbf{E}[e^{\theta Y}] < \infty$ for some $\theta > 0$ it follows from Cramér's theorem, see [2], that

$$\lim_{n\to\infty} -\frac{1}{n} \log p^{(n)} \to I(a),$$

where $I(a) = \sup_{\theta\in\mathbb{R}}\{\theta a - \Lambda(\theta)\}$; the Fenchel-Legendre transform of Λ.

Similarly, the normalising constant of $V^{(n)}$ satisfies $\lim_{n\to\infty} -\frac{1}{n}\log r^{(n)} = \sup_{\theta\in\mathbb{R}}\big\{\theta a - \Lambda(\theta)\big\}$, where $\Lambda(\theta) = \log \mathbf{E}[e^{\theta Z}] = \theta\mu + \frac{(\theta\sigma)^2}{2}$. It follows, by performing the maximisation, that

$$\lim_{n\to\infty} -\frac{1}{n} \log r^{(n)} = \frac{(a-\mu)^2}{2\sigma^2}. \tag{4.5}$$

An appropriate choice of the standard deviation σ is therefore the solution to

$$I(a) = \frac{(a-\mu)^2}{2\sigma^2}.$$

We proceed with the first main result of this paper, the characterisation of our proposed estimator.

Theorem 4.3.4. *Suppose that* $a > \mathbf{E}[Z \log \frac{\phi_{\mu,\sigma}(Z)}{f_Y(Z)}]$ *and*

$$\boldsymbol{H}(p) = \log \mathbf{E}\left[\frac{\phi_{\mu,\sigma}(Z)}{f_Y(Z)} e^{pZ}\right] < \infty, \quad \textit{for some } p > 0. \tag{4.6}$$

Then, the estimator $\widehat{q}_T^{(n)}$ *given by equation* (4.4) *satisfies*

$$\limsup_{n\to\infty} \frac{1}{n} \log \left((p^{(n)})^2 \mathbf{E}_{F_a^{(n)}} [(\widehat{q}_T^{(n)})^2] \right) \leq I(a) - J(a),$$

where $J(a) = \sup_{p>0}\{pa - \boldsymbol{H}(p)\}$.

Theorem 4.3.4 characterises the rare-event efficiency of the algorithm. If $J(a) \geq I(a)$ then $\widehat{q}_T^{(n)}$ is logarithmically efficient. It follows from Jensen's inequality that $I(a) \geq J(a)$, therefore the efficiency is obtain precisely when $J(a) = I(a)$.

Proof. The calculations leading up to (4.2) shows that

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n} \log \left((p^{(n)})^2 \mathbf{E}_{F_a^{(n)}} [(\widehat{q}_T^{(n)})^2] \right) \\
&\quad \leq -I(a) + \limsup_{n\to\infty} \frac{1}{n} \log \mathbf{E}_{V^{(n)}}^{(n)} \Big[I\Big\{ \sum_{i=1}^n Y_i > na \Big\} \frac{dV^{(n)}}{dF^{(n)}} (\mathbf{Y}^{(n)}) \Big].
\end{aligned}$$

Since

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{E}_V^{(n)} \Big[I\Big\{ \sum_{i=1}^n Y_i > na \Big\} \frac{dV^{(n)}}{dF^{(n)}} (\mathbf{Y}^{(n)}) \Big] \\
&\quad = \limsup_{n\to\infty} \frac{1}{n} \log \int I\Big\{ \sum_{i=1}^n z_i > na \Big\} \frac{1}{r^{(n)}} \prod_{i=1}^n \frac{\phi_{\mu,\sigma}(z_i)}{f_Y(z_i)} \frac{1}{r^{(n)}} \prod_{i=1}^n \phi_{\mu,\sigma}(z_i)\, dz_1, \dots dz_n \\
&\quad = \limsup_{n\to\infty} -\frac{2}{n} \log r^{(n)} + \limsup_{n\to\infty} \frac{1}{n} \log \mathbf{E} \Big[I\Big\{ \sum_{i=1}^n Z_i > na \Big\} \prod_{i=1}^n \frac{\phi_{\mu,\sigma}(Z_i)}{f_Y(Z_i)} \Big] \\
&\quad = 2I(a) - \liminf_{n\to\infty} -\frac{1}{n} \log \mathbf{E} \Big[I\Big\{ \sum_{i=1}^n Z_i > na \Big\} e^{\sum_{i=1}^n - \log \frac{f_Y(Z_i)}{\phi_{\mu,\sigma}(Z_i)}} \Big],
\end{aligned}$$

it follows that it is sufficient to prove that

$$\liminf_{n\to\infty} -\frac{1}{n} \log \mathbf{E} \left[I\left\{ \sum_{i=1}^n Z_i > na \right\} \prod_{i=1}^n \frac{\phi_{\mu,\sigma}(Z_i)}{f_Y(Z_i)} \right] \geq \sup_p \{pa - \mathbf{H}(p)\}. \tag{4.7}$$

To show (4.7), take $p \geq 0$ such that

$$\mathbf{E} \left[e^{pZ} \frac{\phi_{\mu,\sigma}(Z)}{f_Y(Z)} \right] < \infty.$$

By Chebyshev's inequality

$$
\begin{aligned}
&-\frac{1}{n}\log \mathbf{E}\left[I\left\{\sum_{i=1}^{n} Z_i > na\right\}\prod_{i=1}^{n}\frac{\phi_{\mu,\sigma}(Z_i)}{f_Y(Z_i)}\right]\\
&\geq -\frac{1}{n}\log \mathbf{E}\left[I\left\{\sum_{i=1}^{n} Z_i > na\right\}e^{p(\sum_{i=1}^{n} Z_i - na)}\prod_{i=1}^{n}\frac{\phi_{\mu,\sigma}(Z_i)}{f_Y(Z_i)}\right]\\
&\geq pa - \frac{1}{n}\log \mathbf{E}\Big[e^{p\sum_{i=1}^{n} Z_i}\prod_{i=1}^{n}\frac{\phi_{\mu,\sigma}(Z_i)}{f_Y(Z_i)}\Big]\\
&= pa - \log \mathbf{E}\Big[e^{pZ}\frac{\phi_{\mu,\sigma}(Z)}{f_Y(Z)}\Big]\\
&= pa - \mathbf{H}(p).
\end{aligned}
$$

Taking supremum over $p \geq 0$ yields

$$
-\frac{1}{n}\log \mathbf{E}\left[I\left\{\sum_{i=1}^{n} Z_i > na\right\}\prod_{i=1}^{n}\frac{\phi_{\mu,\sigma}(Z_i)}{f_Y(Z_i)}\right] \geq \sup_{p\geq 0}\{pa - \mathbf{H}(p)\} = J(a).
$$

□

4.3.2 Positive valued increments

Suppose that the $Y_1, Y_2, \ldots$ are independent and identically distributed random variables with density f_Y with support on $\mathbb{R}_+$. Let $V^{(n)}$ be the probability distribution given by

$$
V^{(n)}(\cdot) = \mathbf{P}\big((Z_1, \ldots, Z_n) \in \cdot \mid (Z_1, \ldots, Z_n) \in R^{(n)}\big),
$$

for $R^{(n)} = \{(Z_1 + \cdots + Z_n)/n > a\}$ where $Z_1, Z_2, \ldots$ are independent and gamma distributed variables with shape parameter α and rate β. The sum of n gamma variables with shape α and rate β is itself gamma with shape $n\alpha$ and rate β and therefore the probability of the event $R^{(n)}$ can be computed explicitly.

Let $(\mathbf{Y}_t^{(n)})_{t=0}^{T-1}$ be generated via Algorithm 4.3.1 with invariant distribution $F_a^{(n)}$. For the given choice of $V^{(n)}$ it follows that the MCMC estimator is given by

$$
\widehat{q}_T^{(n)} = \frac{1}{r^{(n)}}\frac{1}{T}\sum_{t=0}^{T-1}\prod_{i=1}^{n}\frac{g_Z(Y_{t,i})}{f_Y(Y_{t,i})} I\Big\{\frac{1}{n}\sum_{i=1}^{n} Y_{t,i} \geq a\Big\}, \tag{4.8}
$$

where g_Z is the gamma density function with shape α and rate β.

Observe as previously, that the estimator can be viewed as the asymptotic approximation $1/r^{(n)}$ of the true probability $1/p^{(n)}$ times the stochastic correction factor. The reason for the good efficiency of the MCMC estimator relies on the fact that this stochastic correction factor has small normalised variance.

The shape α is set such that $\mathbf{E}[Z_1] = \alpha/\beta$ equals $\mathbf{E}[Y_1]$. The rate β is set so that the large deviation rate of $r^{(n)}$ match the large deviation rate of $p^{(n)}$.

Assuming that $\Lambda(\theta) = \log \mathbf{E}[e^{\theta Y}] < \infty$ for some $\theta > 0$ it follows from Cramér's theorem, see [2], that

$$\lim_{n\to\infty} -\frac{1}{n} \log p^{(n)} \to I(a),$$

where $I(a) = \sup_{\theta\in\mathbb{R}}\{\theta a - \Lambda(\theta)\}$; the Fenchel-Legendre transform of Λ. Similarly, by performing the maximisation, then

$$\lim_{n\to\infty} -\frac{1}{n} \log r^{(n)} = \sup_{\theta\in\mathbb{R}}\{\theta a - \Lambda(\theta)\} = a\beta - \alpha + \alpha \log\Big(\frac{\alpha}{a\beta}\Big).$$

A good choice for the standard deviation λ is hence the solution to

$$I(a) = a\beta - \alpha + \alpha \log\Big(\frac{\alpha}{a\lambda}\Big).$$

We proceed with the second main result of this paper, the characterisation of our proposed estimator.

Theorem 4.3.5. *Suppose that* $a > \mathbf{E}[Z \log \frac{g_Z(Z)}{f_Y(Z)}]$ *and*

$$\boldsymbol{H}(p) = \log \mathbf{E}\left[\frac{g_Z(Z)}{f_Y(Z)} e^{pZ}\right] < \infty, \quad \textit{for some } p > 0. \tag{4.9}$$

Then, the estimator $\widehat{q}_T^{(n)}$ *given by equation* (4.8) *satisfies*

$$\limsup_{n\to\infty} \frac{1}{n} \log\Big((p^{(n)})^2 \mathbf{E}_{F_a^{(n)}}[(\widehat{q}_T^{(n)})^2]\Big) \leq I(a) - J(a),$$

where $J(a) = \sup_{p>0}\{pa - \boldsymbol{H}(p)\}$.

The proof is very similar to the proof of Theorem 4.3.4 is therefore skipped.

4.4 Numerical experiments

This paper presents an MCMC estimator $\widehat{q}_T^{(n)}$ of $(p^{(n)})^{-1}$. However for comparison reasons the numerical experiments are based on $\widehat{p}_T^{(n)} = (\widehat{q}_T^{(n)})^{-1}$ as an estimator for $p^{(n)}$. The MCMC algorithm provided in this paper is compared to the strongly efficient importance sampling algorithm provided by Blanchet, Leder and Glynn [1].

Consider the problem of computing the probability that the sample mean exceeds the threshold larger than its mean. The sought probability is

$$\mathbf{P}(S_n > a),$$

where $S_n = (Y_1 + \cdots + Y_n)/n$ for independent and identically distributed Y's and $a > \mathbf{E}[Y_1]$. We provide numerical experiments for both cases presented in this paper. Firstly where the Y's are real-valued and secondly where the Y's only take positive values.

4.4.1 Real-valued increments

Table 4.1: Numerical results for computing $\mathbf{P}(S_n > 1.50)$, where the increments are real-valued. The performance of the MCMC estimator is compared to the IS estimator.

$n = 2$	MCMC	IS	Monte Carlo
Estimate	2.653e-02	2.695e-02	2.694e-02
Std. deviation	2.910e-04	13.501e-04	53.387e-04
Rel. error	0.01097	0.0501	0.1982
Comp. time(s)	0.13	0.29	0.06
$n = 5$	MCMC	IS	Monte Carlo
Estimate	1.434e-03	1.516e-03	1.320e-03
Std. deviation	9.901e-05	11.007e-05	116.237e-05
Rel. error	0.0691	0.0726	0.8806
Comp. time(s)	0.41	0.71	0.29
$n = 10$	MCMC	IS	Monte Carlo
Estimate	1.367e-05	1.524e-05	1.000e-05
Std. deviation	2.583e-06	3.834e-06	100.000e-06
Rel. error	0.1890	0.2515	10.0000
Comp. time(s)	1.10	1.43	1.04
$n = 15$	MCMC	IS	
Estimate	1.486e-07	1.543e-07	
Std. deviation	4.759e-08	1.296e-08	
Rel. error	0.3196	0.0840	
Comp. time(s)	2.10	2.14	
$n = 20$	MCMC	IS	
Estimate	1.773e-09	1.758e-09	
Std. deviation	7.165e-10	3.975e-10	
Rel. error	0.4042	0.2261	
Comp. time(s)	3.53	2.85	
$n = 25$	MCMC	IS	
Estimate	2.152e-11	1.933e-11	
Std. deviation	10.305e-12	1.593e-12	
Rel. error	0.4788	0.0824	
Comp. time(s)	5.28	3.56	

Suppose that the increments Y are independent and identically distributed normal mixtures given by

$$Y_1 = \begin{cases} W_1 & \text{with probability } 0.40, \\ W_2 & \text{with probability } 0.60, \end{cases}$$

where (W_1, W_2) are independent normally distributed with mean 1.20 and 0.80 respectively and standard deviation 0.20 and 0.50 respectively. The parameters are chosen in this way to imitate the behavior of a volatile economic climate, shifting from good to bad and vice versa.

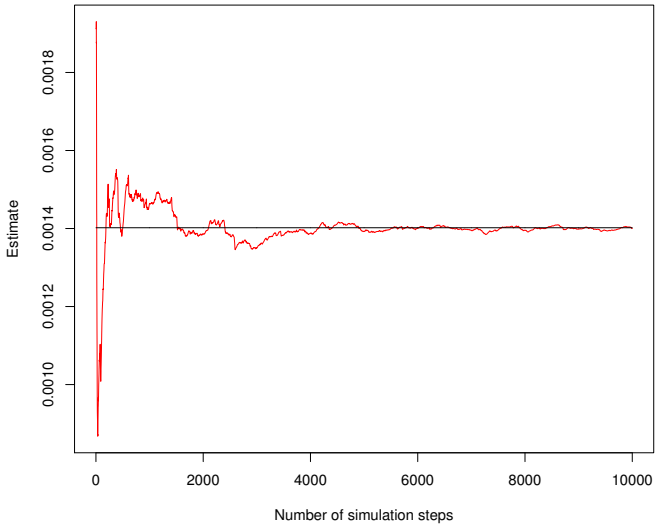
By Theorem 4.3.4 the MCMC estimator is logarithmically efficient if $J(a) = I(a)$. For the normal mixture example we compute numerically

$$J(a) = 0.78, \quad \text{against} \quad I(a) = 1.50.$$

Therefore the estimator is not logarithmically efficient.

Since the MCMC algorithm only generates a single random variable per simulation, while the IS algorithm generates the entire random walk, the number of simulations for the MCMC is scaled up by n so that the computational runtime between the techniques is

Figure 4.1: The point estimate of $\mathbf{P}(S_5 > 1.50)$ as a function of simulations (red line) compared against the true probability (solid line), where the increments are real-valued.

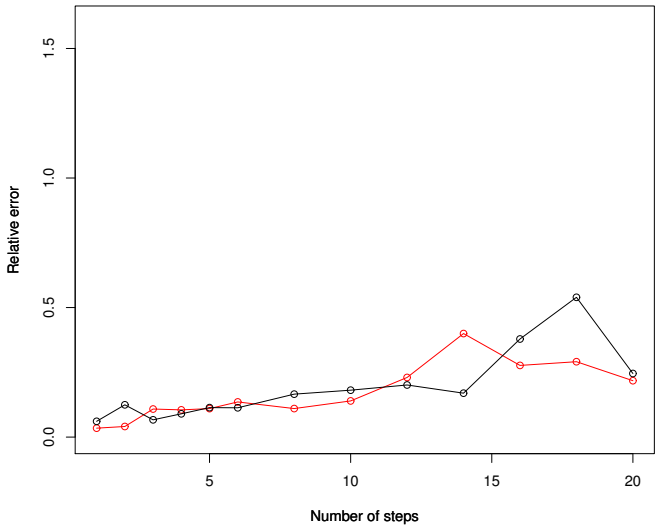


roughly the same. This gives a fairer comparison of the relative error between the algorithms.

Table 4.1 displays the numerical results based on computation using 10^2 batches, each IS batch consisting of 10^3 simulations and MCMC $10^3 n$ simulations. The batch sample mean, sample standard deviation and sample relative error are recorded, as well as the average computational runtime per batch. Figure 4.1 illustrates the convergence of the MCMC algorithm, by recording the point estimate of the probability $\mathbf{P}(S_5 > 1.50)$ as function of number of simulations. The point estimate is compared to the true probability, determined by the standard Monte Carlo estimate using 10^9 simulations. Figure 4.2 illustrates the performance difference between the MCMC algorithm and the IS algorithm, in terms of the relative error as a function of n, the number of increments.

Comparing the numerical results between MCMC and IS shows that the former algorithm outperforms the latter for probabilities in the range from 10^{-2} to 10^{-5} where the IS takes over and is thereafter more efficient. That does not come as a surprise since this particular IS algorithm is strongly efficient against our logarithmically efficient MCMC algorithm.

Figure 4.2: The relative error for estimating $\mathbf{P}(S_n > 1.50)$ as number of increments n increases, where the increments are real-valued. The relative error of the MCMC estimator (red line) compared to the relative error of the IS estimator (solid line).



4.4.2 Positive valued increments

Suppose that the increments Y are independent and identically Weibull distributed with shape $\gamma = 1.5$ and scale $\mu = 0.9$. Since the shape parameter $\gamma > 1$ then it can be shown that $\mathbf{H}$ in Theorem 4.3.5 is not finite for any $p > 0$. Therefore the MCMC estimator is not logarithmically efficient. Nevertheless, we illustrate the performance of the MCMC estimator in this case, which turns out to be comparable with the importance sampling algorithm.

Since the MCMC algorithm only generates a single random variable per simulation, while the IS algorithm generates the entire random walk, the number of simulations for the MCMC is scaled up so that the computational runtime between the techniques is roughly the same. This gives a fairer comparison of the relative error between the algorithms.

Table 4.2 displays the numerical results based on computation using 10 batches, each IS batch consisting of 10^3 simulations and MCMC $10^3 n$ simulations. The batch sample mean, sample standard deviation and sample relative error are recorded, as well as the average computational runtime per batch. Figure 4.3 illustrates the convergence of the MCMC algorithm, by recording the point estimate of the probability $\mathbf{P}(S_{10} > 1.30)$ as function of number of simulations. The point estimate is compared to the true probability, determined by the standard Monte Carlo estimate using 10^9 simulations. Figure 4.2 illustrates the performance difference between the MCMC algorithm and the IS algorithm, in terms of

Table 4.2: Numerical comparison of computing $\mathbf{P}(S_n > 1.30)$, where the increments are positive-valued.

$n = 2$	MCMC	IS	Monte Carlo
Estimate	1.138e-01	1.153e-01	1.135e-01
Std. deviation	2.661e-03	15.658e-03	9.941e-03
Rel. error	0.0234	0.1358	0.0876
Comp. time(s)	0.1030	0.3219	0.0102
$n = 5$	MCMC	IS	Monte Carlo
Estimate	3.556e-02	3.564e-02	3.524e-02
Std. deviation	1.654e-03	5.618e-03	5.523e-03
Rel. error	0.0465	0.1576	0.1567
Comp. time(s)	0.1013	0.7635	0.0106
$n = 10$	MCMC	IS	Monte Carlo
Estimate	6.075e-03	5.226e-03	5.970e-03
Std. deviation	5.650e-04	11.137e-04	24.472e-04
Rel. error	0.0930	0.2131	0.4099
Comp. time(s)	0.1030	1.6570	0.0116
$n = 20$	MCMC	IS	Monte Carlo
Estimate	2.247e-04	1.458e-04	2.300e-04
Std. deviation	3.507e-05	6.522e-05	50.960e-05
Rel. error	0.1561	0.4473	2.2157
Comp. time(s)	0.1033	3.4275	0.0132
$n = 30$	MCMC	IS	Monte Carlo
Estimate	8.745e-06	3.926e-06	1.000e-05
Std. deviation	2.046e-06	1.282e-06	100.000e-06
Rel. error	0.2340	0.3266	10.0000
Comp. time(s)	0.1032	5.2188	0.0146
$n = 40$	MCMC	IS	
Estimate	3.204e-07	1.382e-07	
Std. deviation	9.929e-08	8.834e-08	
Rel. error	0.3099	0.6394	
Comp. time(s)	0.1029	6.9282	

the relative error as a function of n, the number of increments.

Acknowledgments

Henrik Hult's research was supported by the Swedish Research Council.

4.5 References

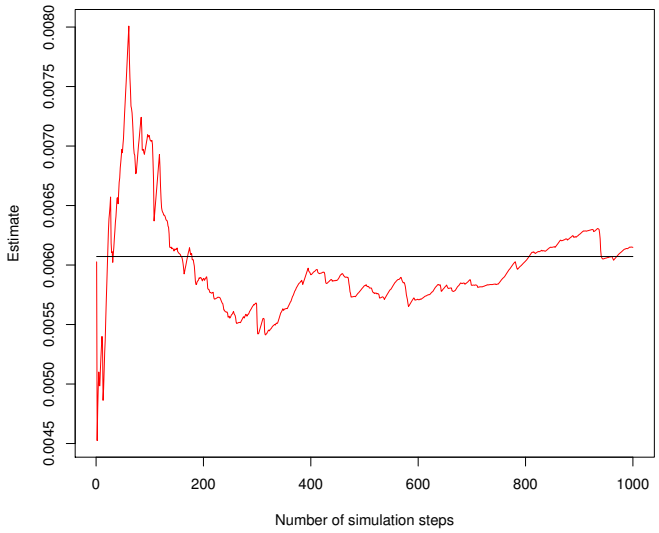
[1] J. Blanchet, K. Leder, and P. W. Glynn. Efficient simulation of light-tailed sums: an old-folk song sung to a faster new tune... *Monte Carlo and Quasi-Monte Carlo Methods 2008*, pages 227–248, 2009.

[2] J.A. Bucklew. *Introduction to Rare Event Simulation*. Springer, New York, 2004.

[3] P. Dupuis and H. Wang. Importance sampling, large deviations, and differential games. *Stoch. Stoch. Rep.*, 76(6):481–508, 2004.

[4] P. Dupuis and H. Wang. Subsolutions of an Isaacs equation and efficient schemes for importance sampling. *Math. Oper. Res.*, 32(3):1–35, 2007.

Figure 4.3: The point estimate of $\mathbf{P}(S_{10} > 1.30)$ as a function of simulations (red line) compared against the true probability (solid line), where the increments are positive-valued.

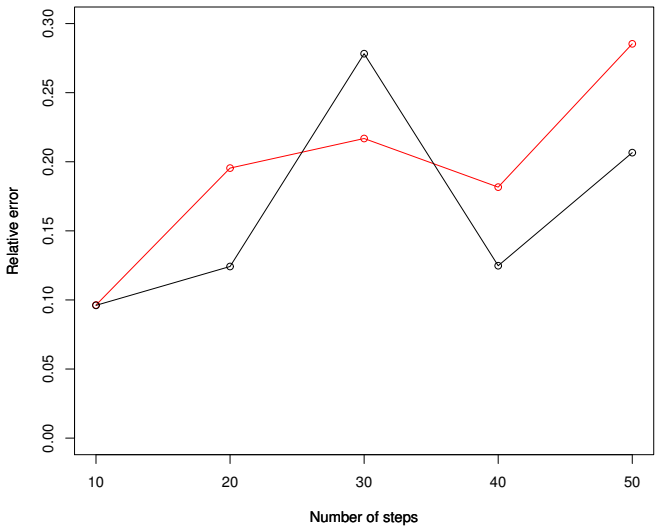


[5] T. Gudmundsson and H. Hult. Markov chain Monte Carlo for computing rare-event probabilities for a heavy-tailed random walk. *J. Appl. Probab.*, 51(2), June 2014.

[6] J. S. Sadowsky. On Monte Carlo estimation of large deviations probabilities. *Ann. Appl. Probab.*, 6(2):399–422, 1996.

[7] D. Siegmund. Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.*, 4(4):673–684, 1976.

Figure 4.4: The relative error for estimating $\mathbf{P}(S_n > 1.30)$ as number of increments n increases, where the increments are positive-valued. The relative error of the MCMC estimator (red line) compared to the relative error of the IS estimator (solid line).



Markov chain Monte Carlo for rare-event simulation for Markov chains

Markov chain Monte Carlo for rare-event simulation for Markov chains

by

Thorbjörn Gudmundsson and Henrik Hult

Abstract

Recently, a method for computing the probability of rare events based on Markov chain Monte Carlo has been developed in [7]. The basic idea is to construct a Markov chain via an MCMC sampler, having as its invariant distribution the conditional distribution given that the event of interest occurs. The probability of the event of interest appears as the normalising constant of that conditional distribution. The normalising constant is estimated from a simulated trajectory of the MCMC sampler.

The purpose of this paper is to investigate the applicability of the MCMC methodology for computing rare-event probabilities in Markov chains that are light-tailed in the sense that they satisfy a logarithmic large deviations principle. Both Markov chains in discrete time and continuous time are treated. The main results gives the exponential rate of decay of the normalised second moment of the estimator. As an illustration the method is applied to a birth-and-death process with state dependent rates.

5.1 Introduction

Efficient computation of probabilities of rare events is a challenging problem that is becoming increasingly important in many areas to quickly and reliably evaluate the risk of disastrous events. When properly designed, stochastic simulation has proven to be a reliable tool to compute such probabilities. The standard Monte Carlo method, where the underlying system is sampled independently and the probability is estimated by the observed frequency of the event, typically fails in the context of rare events. The reason is that when the probability of the event is small a large sample size is needed in order to obtain reliable estimates. To overcome the problems with standard Monte Carlo a number of variance reduction methods have been developed: importance sampling [1], multi-level splitting [2], genealogical particle methods [3], etc. The success of such methods relies on the construction of appropriate changes of measures, related to the large deviations of the

system. In [4] the authors present a importance sampling algorithm based on the construction of subsolutions and illustrate the method for birth-and-death processes.

Recently, a method for computing the probability of rare events that is based on Markov chain Monte Carlo (MCMC) has been developed, see [7]. The basic idea is to construct a Markov chain via an MCMC sampler, having as its invariant distribution the conditional distribution given that the event of interest occurs. The probability of the event of interest appears as the normalising constant of that conditional distribution. The normalising constant is estimated from a trajectory, simulated with the MCMC sampler. The estimator, to be described in detail below, can be viewed as an asymptotic approximation of the sought probability multiplied by a stochastic correction factor. The MCMC methodology has proven to be efficient for computing the probability that a heavy-tailed random walk or random sum exceeds a high threshold. It has also been extended to stochastic recurrence equations with heavy-tailed innovations. In the light-tailed setting it has been developed for computing the probability that a random walks exceeds a high threshold.

The purpose of this paper is to investigate the applicability of the MCMC methodology for computing rare-event probabilities in Markov chains that are light-tailed in the sense that they satisfy a logarithmic large deviations principle. Both Markov chains in discrete time and continuous time are treated. The main results gives the exponential rate of decay of the normalised second moment of the estimator. As an illustration the method is applied to a birth-and-death process with state dependent rates.

The paper is organised as follows. The general MCMC methodology for computing rare-event probabilities and the first steps for proving logarithmic efficiency is presented in Section 5.2. In Section 5.3 discrete-time Markov chains are treated whereas Section 5.4 treats continuous-time Markov chains. An application to a birth-and-death process, including numerical experiments, is presented in Section 5.5.

5.2 Markov chain Monte Carlo in rare-event simulation

In this section the MCMC methodology for rare-event simulation is recaptured and the first steps for proving logarithmic efficiency are explained.

Let $(X^{(n)})_{n\geq 1}$ be a sequence of random elements (e.g. random variables, random vectors, stochastic processes) with $X^{(n)}$ taking values in the state space $\mathscr{E}$ and having probability distribution $F^{(n)}$. Let h_0 be a bounded continuous function from $\mathscr{E}$ to $\mathbb{R}$. Consider the problem of computing the expectation

$$\theta^{(n)} = \mathbf{E}[\exp\{-nh_0(X^{(n)})\}],$$

which is a rare-event problem in the sense that $\theta^{(n)}$ is small and tends to 0 as $n \to \infty$. Denote by $F_{h_0}^{(n)}$ the distribution

$$F_{h_0}^{(n)}(\cdot) = \frac{\mathbf{E}[I\{X^{(n)} \in \cdot\} \exp\{-nh_0(X^{(n)})\}]}{\theta^{(n)}}.$$

Note that $F_{h_0}^{(n)}$ is the zero-variance distribution when computing $\theta^{(n)}$ by importance sampling.

The first step is to design an MCMC sampler which produces an $\mathscr{E}$-valued Markov chain $(X_t^{(n)})_{t\geq 0}$ having $F_{h_0}^{(n)}$ as its invariant distribution. Assuming that such a sampler exists, let $X_0^{(n)}, \ldots, X_{T-1}^{(n)}$ be a sample having $F_{h_0}^{(n)}$ as its invariant distribution.

The second step is to construct an unbiased estimator $\widehat{q}_T^{(n)}$ of $1/\theta^{(n)}$ in the following way. Consider a probability distribution $V^{(n)}$ on $\mathscr{E}$, such that $V^{(n)} \ll F_{h_0}^{(n)}$, and define

$$u^{(n)}(x) = e^{nh_0(x)} \frac{dV^{(n)}}{dF^{(n)}}(x).$$

Taking expectation with respect to $F_{h_0}^{(n)}$ gives

$$\mathbf{E}_{F_{h_0}^{(n)}}[u^{(n)}(X^{(n)})] = \int e^{nh_0(x)} \frac{dV^{(n)}}{dF^{(n)}}(x) F_{h_0}^{(n)}(dx) = \frac{1}{\theta^{(n)}} \int V^{(n)}(dx) = \frac{1}{\theta^{(n)}}.$$

The above calculation motivates the following definition of the MCMC estimator $\widehat{q}_T^{(n)}$ of $1/\theta^{(n)}$

$$\widehat{q}_T^{(n)} = \frac{1}{T} \sum_{t=0}^{T-1} u^{(n)}(X_t^{(n)}).$$

Two key design choices arise which determine the performance of the algorithm. The design of the MCMC sampler and the choice of $V^{(n)}$. The design of the MCMC sampler is essential for the mixing properties of the Markov chain $(X_t^{(n)})_{t\geq 0}$ and thereby the convergence of the estimator as the sample size increases. The higher the mixing rate of the MCMC sampler, the quicker the Markov chain will converge to its invariant distribution. A reasonable criteria is to require that the Markov chain is geometrically ergodic, although it should be noted that geometric ergodicity gives limited information of the finite sample properties of the Markov chain.

The choice of $V^{(n)}$ controls the normalised variance of the estimator:

$$(\theta^{(n)})^2 \operatorname{Var}_{F_{h_0}^{(n)}}(\widehat{q}_T^{(n)}),$$

which determines the rare-event efficiency of the algorithm. Let us assume that $\theta^{(n)}$ decays exponentially in n in such a way that there is a $\gamma \in (0, \infty)$ such that

$$\lim_{n\to\infty} -\frac{1}{n} \log \theta^{(n)} = \gamma. \tag{5.1}$$

The estimator $\widehat{q}_T^{(n)}$ is said to be logarithmically efficient if

$$\limsup_{n\to\infty} \frac{1}{n} \log(\theta^{(n)})^2 \mathbf{E}_{F_{h_0}^{(n)}}[(\widehat{q}_T^{(n)})^2] = 0.$$

The interpretation of logarithmic efficiency is that the exponential rate of growth of the second moment of $\widehat{q}_T^{(n)}$ coincides with that of $(\theta^{(n)})^{-2}$.

For the normalised variance, we have

$$
\begin{aligned}
&(\theta^{(n)})^2 \operatorname{Var}_{F_{h_0}^{(n)}}(\widehat{q}_T^{(n)}) \\
&= (\theta^{(n)})^2 \Big(\frac{1}{T} \operatorname{Var}_{F_{h_0}^{(n)}}(u^{(n)}(X_0^{(n)}) + \frac{2}{T^2} \sum_{t=0}^{T-1} \sum_{s=t+1}^{T-1} \operatorname{Cov}_{F_{h_0}^{(n)}} \left(u^{(n)}(X_s^{(n)}), u^{(n)}(X_t^{(n)}) \right) \Big) \\
&\leq (\theta^{(n)})^2 \Big(\frac{1}{T} + \frac{T(T-1)}{T^2} \Big) \operatorname{Var}_{F_{h_0}^{(n)}}(u^{(n)}(X_0^{(n)})) \\
&= (\theta^{(n)})^2 \operatorname{Var}_{F_{h_0}^{(n)}}(u^{(n)}(X_0^{(n)})),
\end{aligned}
$$

where we have used the crude upper bound

$$
\operatorname{Cov}_{F_{h_0}^{(n)}} \left(u^{(n)}(X_s^{(n)}), u^{(n)}(X_t^{(n)}) \right) \leq \operatorname{Var}_{F_{h_0}^{(n)}}(u^{(n)}(X_0^{(n)})).
$$

Next we see that

$$
\begin{aligned}
&(\theta^{(n)})^2 \operatorname{Var}_{F_{h_0}^{(n)}}(u^{(n)}(X_0^{(n)})) \\
&\quad = (\theta^{(n)})^2 \mathbf{E}_{F_{h_0}^{(n)}} \Big[(u^{(n)}(X_0^{(n)}))^2 \Big] - 1 \\
&\quad = (\theta^{(n)})^2 \int e^{2nh_0(x)} \Big(\frac{dV^{(n)}}{dF^{(n)}}(x) \Big)^2 F_{h_0}^{(n)}(dx) - 1 \\
&\quad = \theta^{(n)} \int e^{nh_0(x)} \frac{dV^{(n)}}{dF^{(n)}}(x) V^{(n)}(dx) - 1 \\
&\quad = \theta^{(n)} \mathbf{E}_{V^{(n)}} \Big[e^{nh_0(X^{(n)})} \frac{dV^{(n)}}{dF^{(n)}}(X^{(n)}) \Big] - 1.
\end{aligned}
$$

Note that, taking $V^{(n)} = F_{h_0}^{(n)}$ implies that

$$
\operatorname{Var}_{F_{h_0}^{(n)}}(u^{(n)}(X_0^{(n)})) = 0,
$$

thus motivating taking $V^{(n)}$ as an approximation of $F_{h_0}^{(n)}$. Observe that

$$
\begin{aligned}
&\frac{1}{n} \log \Big\{ \theta^{(n)} \mathbf{E}_{V^{(n)}} \Big[e^{nh_0(X^{(n)})} \frac{dV^{(n)}}{dF^{(n)}}(X^{(n)}) \Big] \Big\} \\
&= \frac{1}{n} \log \theta^{(n)} + \frac{1}{n} \log \mathbf{E}_{V^{(n)}} \Big[e^{nh_0(X^{(n)})} \frac{dV^{(n)}}{dF^{(n)}}(X^{(n)}) \Big],
\end{aligned}
$$

so it follows from the large deviation assumption in (5.1) that to prove logarithmic efficiency it is sufficient to show that

$$
\limsup_{n \to \infty} \frac{1}{n} \log \mathbf{E}_{V^{(n)}} \Big[e^{nh_0(X^{(n)})} \frac{dV^{(n)}}{dF^{(n)}}(X^{(n)}) \Big] \leq \gamma. \tag{5.2}
$$

5.3 Markov chains in discrete time

In this section the MCMC methodology is developed to compute the probability of rare events for Markov chains in discrete time. We present a Metropolis-Hastings sampler and characterise the rare-event efficiency of the estimator.

For each $n \geq 1$, let $(X_i^{(n)}, i \geq 0)$, where $X_0^{(n)} = x_0$ for all $n \geq 1$, be a Markov chain satisfying the updating mechanism

$$X_{i+1}^{(n)} = X_i^{(n)} + \frac{1}{n} \upsilon_i(X_i^{(n)}),$$

where the distribution of the independent and identically distributed random vector fields $\upsilon_i(x)$ are given by the stochastic kernel $\mu(\cdot \mid x)$. The piecewise linear interpolation of $X_1^{(n)}, \ldots, X_n^{(n)}$, is given by $X^{(n)} = (X^{(n)}(s); 0 \leq s \leq 1)$, where $X^{(n)}(0) = x_0$ and

$$X^{(n)}(s) = X_i^{(n)} + \left(s - \frac{i}{n}\right)(X_{i+1}^{(n)} - X_i^{(n)}), \quad \frac{i}{n} \leq s \leq \frac{i+1}{n}.$$

The state space of $X^{(n)}$ is the space of continuous functions $\mathscr{C}([0,1]; \mathbb{R}^d)$. In this paper we will be concerned with the computation of expectations of the form

$$\theta_n = \mathbf{E}[\exp\{-nh_0(X^{(n)})\}],$$

where h_0 is a bounded continuous mapping $\mathscr{C}([0,1]; \mathbb{R}^d) \to \mathbb{R}$.

The first step is designing an MCMC sampler which produces a Markov chain $(X_t^{(n)}(s); 0 \leq s \leq 1)_{t \geq 0}$ where, for each $t \geq 0$, $X_t^{(n)}$ takes values in $\mathscr{C}([0,1]; \mathbb{R}^d)$ and such that $X_t^{(n)}(\cdot)$ has an invariant distribution $F_{h_0}^{(n)}$ given by the conditional distribution

$$F_{h_0}^{(n)}(\cdot) = \frac{\mathbf{E}[I\{X^{(n)} \in \cdot\} \exp\{-nh_0(X^{(n)})\}]}{\theta_n}.$$

5.3.1 Metropolis-Hastings algorithm for sampling from $F_{h_0}^{(n)}$

In this section the MCMC algorithm is presented which generates a $\mathscr{C}([0,1]; \mathbb{R}^d)$-valued Markov chain

$$\left(X_t^{(n)}(s), s \in [0,1]\right)_{t \geq 0},$$

whose invariant distribution is $F_{h_0}^{(n)}$.

Briefly the algorithm is as follows. In each step, a window of random size is selected and the process is updated by a random walk Metropolis-Hastings step within the selected window.

Algorithm 5.3.1. The transition from $X_t^{(n)}(\cdot)$ to $X_{t+1}^{(n)}(\cdot)$, $t \geq 0$, is sampled as follows. Take $h \in (0,1)$, possibly at random.

1. Sample u uniformly on $[0, 1-h]$ and suppose $i-1 \leq nu < i$ and $j \leq n(u+h) < j+1$. Proceed by proposing a new trajectory of $s \mapsto X_t^{(n)}(s)$ in the interval $[i/n, j/n]$. The part of the process which is to be re-sampled is

$$S(t, u, u+h) = \big\{X_t^{(n)}(i/n), X_t^{(n)}((i+1)/n), \ldots, X_t^{(n)}(j/n)\big\}.$$

2. Generate the proposal

$$\hat{S}(t, u, u+h) = \big\{X_t^{(n)}(i/n), \hat{X}_t^{(n)}((i+1)/n), \ldots, \hat{X}_t^{(n)}((j-1)/n), X_t^{(n)}(j/n)\big\},$$

by sampling $\hat{X}_t^{(n)}((i+1)/n), \ldots, \hat{X}_t^{(n)}((j-1)/n)$ from a proposal density $q(\cdot \mid X_t^{(n)}(i/n), X_t^{(n)}(j/n))$. Let $\hat{X}_t^{(n)}$ be the trajectory such that

$$\hat{X}_t^{(n)}(k/n) = X_t^{(n)}(k/n),$$

for $0 \leq k \leq i$ and $j \leq k \leq n$, and determined by $\hat{S}(t, u, u+h)$ for $i < k < j$.

3. The proposed trajectory $\hat{X}_t^{(n)}$ is accepted with probability α given by

$$\begin{aligned}\alpha = 1 \wedge \Bigg(& \frac{e^{-nh_0(\hat{X}_t^{(n)})} \prod_{k=0}^{n-1} d\mu(\hat{\upsilon}_k(\hat{X}_t^{(n)}(k/n)) \mid \hat{X}_t^{(n)}(k/n))}{e^{-nh_0(X_t^{(n)})} \prod_{k=0}^{n-1} d\mu(\upsilon_k(X_t^{(n)}(k/n)) \mid X_t^{(n)}(k/n))} \\ & \times \frac{q(X_t^{(n)}((i+1)/n), \ldots, X_t^{(n)}((j-1)/n) \mid \hat{X}_t^{(n)}(i/n), \hat{X}_t^{(n)}(j/n))}{q(\hat{X}_t^{(n)}((i+1)/n), \ldots, \hat{X}_t^{(n)}((j-1)/n) \mid X_t^{(n)}(i/n), X_t^{(n)}(j/n))} \Bigg) \\ = 1 \wedge \Bigg(& \frac{e^{-nh_0(\hat{X}_t^{(n)})} \prod_{k=i}^{j-1} d\mu(\hat{\upsilon}_k(\hat{X}_t^{(n)}(k/n)) \mid \hat{X}_t^{(n)}(k/n))}{e^{-nh_0(X_t^{(n)})} \prod_{k=i}^{j-1} d\mu(\upsilon_k(X^{(n)}(k/n)) \mid X^{(n)}(k/n))} \\ & \times \frac{q(X_t^{(n)}((i+1)/n), \ldots, X_t^{(n)}((j-1)/n) \mid \hat{X}_t^{(n)}(i/n), \hat{X}_t^{(n)}(j/n))}{q(\hat{X}_t^{(n)}((i+1)/n), \ldots, \hat{X}_t^{(n)}((j-1)/n) \mid X_t^{(n)}(i/n), X_t^{(n)}(j/n))} \Bigg),\end{aligned}$$

where $\hat{\upsilon}_k(\hat{X}_t^{(n)}(k/n)) = n(\hat{X}_t^{(n)}((k+1)/n) - \hat{X}_t^{(n)}(k/n))$. If accepted, put $X_{t+1}^{(n)} = \hat{X}_t^{(n)}$, otherwise put $X_{t+1}^{(n)} = X_t^{(n)}$.

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

5.3.2 Analysis of rare-event efficiency

Given a sample $X_0^{(n)}, \ldots, X_{T-1}^{(n)}$ via Algorithm 5.3.1, let us proceed with the design of $V^{(n)}$ and the corresponding analysis of rare-event efficiency for the estimator $\widehat{q}_T^{(n)}$.

For each $x \in \mathbb{R}^d$, let $H(x, \cdot)$ denote the cumulant generating function of the stochastic kernel $\mu(\cdot \mid x)$, that is,

$$H(x, p) = \log \int_{\mathbb{R}^d} e^{\langle p, y \rangle} \mu(dy \mid x),$$

and $L(x,\cdot)$ denote the Fenchel-Legendre transform of $H(x,\cdot)$,

$$L(x,v) = \sup_{p\in\mathbb{R}^d}\{\langle p,v\rangle - H(x,p)\}.$$

Condition A.

(a) For each $p\in\mathbb{R}^d$, $\sup_{x\in\mathbb{R}^d} H(x,p) < \infty$.

(b) The mapping $x \mapsto \mu(\cdot \mid x) \in \mathscr{P}(\mathbb{R}^d)$ is continuous in the weak topology on $\mathscr{P}(\mathbb{R}^d)$.

Let ri(conv $S_{\mu(\cdot|x)}$) denote the relative interior of the convex hull of the support of $\mu(\cdot \mid x)$.

Condition B.

(a) The sets ri(conv $S_{\mu(\cdot|x)}$) are independent of $x\in\mathbb{R}^d$.

(b) $0 \in$ ri(conv $S_{\mu(\cdot|x)}$).

Under Condition A and Condition B it follows from Theorem 6.3.3 in [5] that the Markov chain $\{X^{(n)}\}$ satisfies the Laplace principle

$$\begin{aligned}&\lim_{n\to\infty} -\frac{1}{n}\log \mathbf{E}[\exp\{-nh(X^{(n)})\}]\\ &\quad = \inf\Big\{\int_0^1 L(\psi(s),\dot\psi(s))ds + h(\psi); \psi\in\mathscr{AC}([0,1];\mathbb{R}^d), \psi(0)=x_0\Big\},\end{aligned}$$

for each bounded continuous $h : \mathscr{C}([0,1];\mathbb{R}^d) \to \mathbb{R}$. Here $\mathscr{AC}([0,1];\mathbb{R}^d)$ denotes the space of absolutely continuous functions $\psi : [0,1] \to \mathbb{R}^d$.

Let $V^{(n)}$ be the probability distribution given by

$$V^{(n)}(\cdot) = \frac{\mathbf{E}\Big[I\big\{(\bar X_1^{(n)},\dots,\bar X_n^{(n)})\in\cdot\big\}\exp\{-nh_1(\bar X^{(n)})\}\Big]}{\bar\theta^{(n)}},$$

where $\bar\theta^{(n)} = \mathbf{E}[\exp\{-nh_1(\bar X^{(n)})\}]$ and the $\bar X_1^{(n)}, \bar X_2^{(n)},\dots$ is a Markov chain of the form

$$\bar X_{i+1}^{(n)} = \bar X_i^{(n)} + \frac{1}{n}\bar v_i(\bar X_i^{(n)}),$$

where the independent and identically distributed random vector fields $\bar v_i(\bar x)$ has distribution $\bar\mu(\cdot \mid \bar x)$ and $h_1 : \mathscr{C}([0,1];\mathbb{R}^d) \to \mathbb{R}$ is bounded and continuous.

We will assume that h_1 and $\bar\mu$ are such that $V^{(n)} \ll F_{h_0}^{(n)}$ and sufficiently simple so that $\bar\theta^{(n)}$ can be computed explicitly. The latter assumption is restrictive and enforces that the Markov chain $\bar X^{(n)}$ is rather simple.

We will assume that the kernel $\bar{\mu}(\cdot \mid \bar{x})$ satisfies Condition A and Condition B so that the Laplace principle holds for $\bar{X}^{(n)}$ as well. In addition we assume that $\bar{\mu}(\cdot \mid \bar{x})$ is taken such that the large deviations rate of $\bar{\theta}^{(n)}$ coincides with that of $\theta^{(n)}$. More precisely, that

$$\begin{aligned}\bar{I}(h_1) &:= \inf\Big\{\int_0^1 \bar{L}(\psi(s),\dot{\psi}(s))ds + h_1(\psi); \psi \in \mathscr{AC}([0,1];\mathbb{R}^d), \psi(0)=x_0\Big\}\\ &= -\frac{1}{n}\log\bar{\theta}^{(n)}\\ &= -\frac{1}{n}\log\theta^{(n)}\\ &= \inf\Big\{\int_0^1 L(\psi(s),\dot{\psi}(s))ds + h_0(\psi); \psi \in \mathcal{AC}[0,1], \psi(0)=x_0\Big\} =: I(h_0),\end{aligned}$$

We proceed with the characterisation of efficiency of the proposed MCMC estimator. For each $\bar{x} \in \mathbb{R}^d$, $p_1 \in \mathbb{R}^d, p_2 \in \mathbb{R}$ let

$$\begin{aligned}\mathbf{H}(\bar{x},p_1,p_2) &= \log\int_{\mathbb{R}^d} e^{\langle p,y\rangle}\Big(\frac{d\bar{\mu}(\cdot \mid \bar{x})}{d\mu(\cdot \mid \bar{x})}(y)\Big)^{p_2}\bar{\mu}(dy \mid \bar{x}),\\ \mathbf{L}(\bar{x},v_1,v_2) &= \sup_{p_1\in\mathbb{R}^d,p_2\in\mathbb{R}}\Big\{\langle p_1,v_1\rangle + p_2v_2 - \mathbf{H}(\bar{x},p_1,p_2)\Big\}.\end{aligned}$$

Theorem 5.3.2. *Suppose that*

$$\sup_{\bar{x}\in\mathbb{R}^d}\mathbf{H}(\bar{x},p_1,p_2) < \infty, \quad \textit{for every } p_1 \in \mathbb{R}^d, p_2 \in \mathbb{R}. \tag{5.3}$$

Then

$$\lim_{n\to\infty}\frac{1}{n}\log\Big((\theta^{(n)})^2\mathbf{E}_{F_{h_0}^{(n)}}[u^{(n)}(X^{(n)})^2]\Big) \le I(h_0) - J(2h_1 - h_0),$$

where, for any bounded continuous $h : \mathscr{C}([0,T];\mathbb{R}^d) \to \mathbb{R}$,

$$\begin{aligned}J(h) = \inf\Big\{\int_0^1 \mathbf{L}(\psi(s),\dot{\psi}_1(s),-\dot{\psi}_2(s))ds + h(\psi_1) + \psi_2(1),\\ \psi \in \mathcal{AC}([0,1];\mathbb{R}^d\times\mathbb{R}), \psi(0)=(x_0,0)\Big\}.\end{aligned}$$

Proof. The calculation leading up to (5.2) shows that

$$\begin{aligned}&\limsup_{n\to\infty}\frac{1}{n}\log\Big((\theta^{(n)})^2\mathbf{E}_{F_{h_0}^{(n)}}[(\widehat{q}_T^{(n)})^2]\Big)\\ &\quad\le -I(h_0) + \limsup_{n\to\infty}\frac{1}{n}\log\mathbf{E}_{V^{(n)}}\Big[\exp\{nh_0(X^{(n)})\}\frac{dV^{(n)}}{dF^{(n)}}(X^{(n)})\Big].\end{aligned}$$

Let $\bar{F}^{(n)}$ denote the distribution of $(\bar{X}_0^{(n)},\dots,\bar{X}_n^{(n)})$ and note that

$$\frac{dV^{(n)}}{d\bar{F}^{(n)}}(x) = \frac{e^{-nh_1(x)}}{\bar{\theta}^{(n)}}.$$

Since

$$\begin{aligned}
&\limsup_{n\to\infty}\frac{1}{n}\log\mathbf{E}_{V^{(n)}}\Big[\exp\{nh_0(X^{(n)})\}\frac{dV^{(n)}}{dF^{(n)}}(X^{(n)})\Big]\\
&\quad=\limsup_{n\to\infty}\frac{1}{n}\log\Big\{\frac{1}{(\bar\theta^{(n)})^2}\mathbf{E}\Big[\exp\big\{n(h_0(\bar X^{(n)})-2h_1(\bar X^{(n)}))\big\}\frac{d\bar F^{(n)}}{dF^{(n)}}(\bar X^{(n)})\Big]\Big\}\\
&\quad\le\limsup_{n\to\infty}-\frac{2}{n}\log\bar\theta^{(n)}\\
&\qquad+\limsup_{n\to\infty}\frac{1}{n}\log\mathbf{E}\left[\exp\{-n(2h_1(\bar X^{(n)})-h_0(\bar X^{(n)}))\}\prod_{i=0}^{n-1}\frac{d\bar\mu(\cdot\mid\bar X_i^{(n)})}{d\mu(\cdot\mid\bar X_i^{(n)})}(\bar v_i(\bar X_i^{(n)}))\right]\\
&\quad=2I(h_0)\\
&\qquad-\liminf_{n\to\infty}-\frac{1}{n}\log\mathbf{E}\left[\exp\{-n(2h_1(\bar X^{(n)})-h_0(\bar X^{(n)}))\}\prod_{i=0}^{n-1}\frac{d\bar\mu(\cdot\mid\bar X_i^{(n)})}{d\mu(\cdot\mid\bar X_i^{(n)})}(\bar v_i(\bar X_i^{(n)}))\right],
\end{aligned}$$

it follows that it is sufficient to prove that

$$\begin{aligned}
&\liminf_{n\to\infty}-\frac{1}{n}\log\mathbf{E}\left[\exp\{-n(2h_1(\bar X^{(n)})-h_0(\bar X^{(n)}))\}\prod_{i=0}^{n-1}\frac{d\bar\mu(\cdot\mid\bar X_i^{(n)})}{d\mu(\cdot\mid\bar X_i^{(n)})}(\bar v_i(\bar X_i^{(n)}))\right]\\
&\qquad\ge J(h_1).
\end{aligned}\tag{5.4}$$

Let us introduce the notation $\bar Z_i^{(n)}=(\bar X_i^{(n)},\bar Y_i^{(n)})$ where

$$\bar Y_i^{(n)}=-\frac{1}{n}\sum_{j=0}^{i-1}\log\Big(\frac{d\bar\mu(\cdot\mid\bar X_i^{(n)})}{d\mu(\cdot\mid\bar X_i^{(n)})}(\bar v_i(\bar X_i^{(n)}))\Big)\ ,\ n\ge1,\ i=0,\dots,n-1.$$

Note that $\bar Z^{(n)}$ is also a Markov chain of the form

$$\bar Z_{i+1}^{(n)}=\bar Z_i^{(n)}+\frac{1}{n}\bar\zeta_i(\bar Z_i^{(n)}),$$

where $\bar\zeta_i(\bar x,\bar y)=\bar\zeta_i(\bar x)=(\bar v_i(\bar x),-\log\frac{d\bar\mu(\cdot|\bar x)}{d\mu(\cdot|\bar x)}(\bar v_i(\bar x)))$ has distribution $\bar\nu(\cdot\mid\bar x)$ given by

$$\int_{\mathbb{R}^d\times\mathbb{R}}f(\bar\zeta)\bar\nu(d\zeta\mid\bar x)=\int_{\mathbb{R}^d}f\Big(z,-\log\frac{d\bar\mu(\cdot\mid\bar x)}{d\mu(\cdot\mid\bar x)}(z)\Big)\bar\mu(dz\mid\bar x),$$

for every bounded measurable function $f:\mathbb{R}^d\times\mathbb{R}\to\mathbb{R}$. Note also that the log moment generating function of $\bar\zeta_i(\bar x)$ is

$$H_{\bar\zeta}(\bar x,p_1,p_2)=\log\int_{\mathbb{R}^d}\exp\{\langle p_1,z\rangle-p_2\log\frac{d\bar\mu(\cdot\mid\bar x)}{d\mu(\cdot\mid\bar x)}(z)\}\bar\mu(dz\mid\bar x)=\mathbf{H}(\bar x,p_1,-p_2).$$

Consequently,

$$L_{\bar{\zeta}}(\bar{x}, v_1, v_2) = \sup_{p_1 \in \mathbb{R}^d, p_2 \in \mathbb{R}} \{\langle p_1, v_1 \rangle + p_2 v_2 - H_{\bar{\zeta}}(\bar{x}, p_1, p_2)\} = \mathbf{L}(\bar{x}, v_1, -v_2).$$

The assumption (5.3) implies that the stochastic kernel $\bar{\nu}(\cdot \mid \bar{x})$ satisfies Condition A and hence, by Proposition 6.2.2 in [5], the Laplace principle upper bound holds for $\bar{Z}^{(n)}$:

$$\begin{aligned}
&\lim_{n\to\infty} -\frac{1}{n} \log \mathbf{E}\left[e^{-n\bar{h}(\bar{Z}^{(n)})}\right] \\
&\quad \geq \inf\Big\{ \int_0^1 L_{\bar{\zeta}}\big(\psi(s), \dot{\psi}_1(s), -\dot{\psi}_2(s)\big) ds + \bar{h}(\psi); \\
&\qquad \psi \in \mathcal{AC}([0,1] \times \mathbb{R}^{d+1}), \psi(0) = (x_0, 0)\Big\} \\
&\quad = \inf\Big\{ \int_0^1 \mathbf{L}\big(\psi(s), \dot{\psi}_1(s), -\dot{\psi}_2(s)\big) ds + \bar{h}(\psi); \psi \in \mathcal{AC}([0,1]; \\
&\qquad \mathbb{R}^{d+1}), \psi(0) = (x_0, 0)\Big\}, \qquad (5.5)
\end{aligned}$$

for each bounded continuous function $\bar{h} : \mathscr{C}([0,1]; \mathbb{R}^d \times \mathbb{R}) \to \mathbb{R}$. We would like to apply the Laplace principle with the function $\bar{h}(x) = 2h_1(x_1) - h_0(x_1) + x_2(1)$, but this function is not bounded. It is, nevertheless, possible to apply the Laplace principle with this function. Indeed, by Theorem 1.3.4 in [5], (5.5) holds for this choice of $\bar{h}$ if

$$\lim_{C\to\infty} \limsup_{n\to\infty} \frac{1}{n} \log \mathbf{E}\left[I\left\{\bar{Y}_n^{(n)} < -C\right\} e^{-n\bar{Y}_n^{(n)}}\right] = -\infty.$$

To see that this is indeed true, write

$$\begin{aligned}
&\lim_{C\to\infty} \limsup_{n\to\infty} \frac{1}{n} \log \mathbf{E}\left[I\left\{\bar{Y}_n^{(n)} < -C\right\} e^{-n\bar{Y}_n^{(n)}}\right] \\
&\quad = \lim_{C\to\infty} \limsup_{n\to\infty} \frac{1}{n} \log \mathbf{E}\Bigg[I\left\{ \prod_{i=0}^{n-1} \frac{d\bar{\mu}(\cdot \mid \bar{X}_i^{(n)})}{d\mu(\cdot \mid \bar{X}_i^{(n)})}(\bar{v}_i(\bar{X}_i^{(n)})) > e^{Cn} \right\} \\
&\qquad \times \prod_{i=0}^{n-1} \frac{d\bar{\mu}(\cdot \mid \bar{X}_i^{(n)})}{d\mu(\cdot \mid \bar{X}_i^{(n)})}(\bar{v}_i(\bar{X}_i^{(n)}))\Bigg].
\end{aligned}$$

For any $p_2 > 0$, by Chebyshev's inequality, the expression in the last display is less than or equal to

$$\begin{aligned}
&\lim_{C\to\infty} \limsup_{n\to\infty} \frac{1}{n} \log \mathbf{E}\left[e^{-p_2 Cn} \Big(\prod_{i=0}^{n-1} \frac{d\bar{\mu}(\cdot \mid \bar{X}_i^{(n)})}{d\mu(\cdot \mid \bar{X}_i^{(n)})}(\bar{v}_i(\bar{X}_i^{(n)})) \Big)^{p_2} \prod_{i=0}^{n-1} \frac{d\bar{\mu}(\cdot \mid \bar{X}_i^{(n)})}{d\mu(\cdot \mid \bar{X}_i^{(n)})}(\bar{v}_i(\bar{X}_i^{(n)})) \right] \\
&\quad = \lim_{C\to\infty} -p_2 C + \limsup_{n\to\infty} \frac{1}{n} \log \mathbf{E}\left[\Big(\prod_{i=0}^{n-1} \frac{d\bar{\mu}(\cdot \mid \bar{X}_i^{(n)})}{d\mu(\cdot \mid \bar{X}_i^{(n)})}(\bar{v}_i(\bar{X}_i^{(n)})) \Big)^{p_2+1} \right].
\end{aligned}$$

For the second term we have the upper bound

$$
\begin{aligned}
&\frac{1}{n}\log \mathbf{E}\left[\Big(\prod_{i=0}^{n-1}\frac{d\bar{\mu}(\cdot \mid \bar{X}_i^{(n)})}{d\mu(\cdot \mid \bar{X}_i^{(n)})}(\bar{v}_i(\bar{X}_i^{(n)}))\Big)^{p_2+1}\right]\\
&\quad=\frac{1}{n}\log \mathbf{E}\left[\Big(\prod_{i=0}^{n-2}\frac{d\bar{\mu}(\cdot \mid \bar{X}_i^{(n)})}{d\mu(\cdot \mid \bar{X}_i^{(n)})}(\bar{v}_i(\bar{X}_i^{(n)}))\Big)^{p_2+1}\right.\\
&\qquad\left.\times\ \mathbf{E}\Big[\Big(\frac{d\bar{\mu}(\cdot \mid \bar{X}_{n-1}^{(n)})}{d\mu(\cdot \mid \bar{X}_{n-1}^{(n)})}(\bar{v}_{n-1}(\bar{X}_{n-1}^{(n)}))\Big)^{p_2+1} \mid \bar{X}_1^{(n)},\ldots,\bar{X}_{n-1}^{(n)}\Big]\right]\\
&\quad\le\frac{1}{n}\log \mathbf{E}\left[\Big(\prod_{i=0}^{n-2}\frac{d\bar{\mu}(\cdot \mid \bar{X}_i^{(n)})}{d\mu(\cdot \mid \bar{X}_i^{(n)})}(\bar{v}_i(\bar{X}_i^{(n)}))\Big)^{p_2+1}\right]\sup_{\bar{x}\in\mathbb{R}^d}\mathbf{E}\Big[\Big(\frac{d\bar{\mu}(\cdot \mid \bar{x})}{d\mu(\cdot \mid \bar{x})}(\bar{v}_{n-1}(\bar{x}))\Big)^{p_2+1}\Big]\\
&\quad\le\cdots\\
&\quad\le\frac{1}{n}\log\Big(\sup_{\bar{x}\in\mathbb{R}^d}\mathbf{E}\Big[\Big(\frac{d\bar{\mu}(\cdot \mid \bar{x})}{d\mu(\cdot \mid \bar{x})}(\bar{v}_0(\bar{x}))\Big)^{p_2+1}\Big]\Big)^n\\
&\quad=\log\Big(\sup_{\bar{x}\in\mathbb{R}^d}\mathbf{E}\Big[\Big(\frac{d\bar{\mu}(\cdot \mid \bar{x})}{d\mu(\cdot \mid \bar{x})}(\bar{v}_0(\bar{x}))\Big)^{p_2+1}\Big]\Big)\\
&\quad=\sup_{\bar{x}\in\mathbb{R}^d}\mathbf{H}(\bar{x},0,p_2+1)<\infty.
\end{aligned}
$$

Combining the last three displays gives

$$
\begin{aligned}
&\lim_{C\to\infty}\limsup_{n\to\infty}\frac{1}{n}\log\mathbf{E}\left[I\left\{\bar{Y}_n^{(n)}<-C\right\}e^{-n\bar{Y}_n^{(n)}}\right]\\
&\qquad\le\lim_{C\to\infty}-p_2C+\sup_{\bar{x}\in\mathbb{R}^d}\mathbf{H}(\bar{x},0,p_2+1)=-\infty,
\end{aligned}
$$

since $p_2>0$. This completes the proof of (5.4) and the theorem. □

5.4 Markov processes in continuous time

In this section the MCMC simulation technique is developed to compute the probability of rare events of pure-jump Markov chains in continuous time.

For every $\epsilon>0$, let $\{X^{(\epsilon)}(t);t\ge 0\}$ be a pure-jump Markov chain in continuous time with generator $\mathscr{A}^\epsilon$ given by

$$
\mathscr{A}^\epsilon f(x)=\frac{1}{\epsilon}\int_{\mathbb{R}^d}f(x+\epsilon y)-f(x)\nu_x(dy),
$$

where $\{\nu_x,x\in\mathbb{R}^d\}$ is a family of measures such that for each Borel set $A\subset\mathbb{R}^d$, $x\mapsto\nu_x(A)$ is a measurable function and for each $x\in\mathbb{R}^d$

$$
\nu_x(\{0\})=0,\quad \nu_x(\mathbb{R}^d)<\infty,\quad \text{and}\ \int_{\mathbb{R}^d}|y|^2\nu_x(dy)<\infty. \tag{5.6}
$$

The stochastic kernel associated with $X^{(\epsilon)}$ is given by

$$\Theta^{(\epsilon)}(dt\,dy \mid x) = \frac{1}{\epsilon} e^{-\nu_x(\mathbb{R}^d)t} dt\,\nu_x(dy), \quad x \in \mathbb{R}^d.$$

In this section we will be concerned with the computation of expectations of the form

$$\theta_\epsilon = \mathbf{E}[\exp\{-\frac{1}{\epsilon} h_0(X^{(\epsilon)})\}],$$

where h_0 is a bounded continuous mapping $\mathscr{D}([0,1];\mathbb{R}^d) \to \mathbb{R}$. Here $\mathscr{D}([0,1];\mathbb{R}^d) \to \mathbb{R}$ denotes the space of càdlàg functions equipped with Skorohod's J_1-metric.

The first step is designing an MCMC sampler which produces a Markov chain $(X_t^{(\epsilon)}(s); 0 \leq s \leq 1)_{t\geq 0}$, having the conditional distribution

$$F_{h_0}^{(\epsilon)}(\cdot) = \frac{\mathbf{E}[I\{X^{(\epsilon)} \in \cdot\} \exp\{-nh_0(X^{(\epsilon)})\}]}{\theta_\epsilon},$$

as its invariant distribution.

5.4.1 Metropolis-Hastings algorithm for sampling from $F_{h_0}^{(n)}$

In this section the MCMC algorithm is presented which generates a $\mathscr{D}([0,1];\mathbb{R}^d)$-valued Markov chain $(X_t^{(\epsilon)}(s), s \in [0,1])_{t\geq 0}$ whose invariant distribution is $F_{h_0}^{(\epsilon)}$.

Briefly the algorithm is as follows. In each step, a window of random length is selected and the process is updated by a random walk Metropolis-Hastings step within the selected window.

Algorithm 5.4.1. The algorithm describes the transition from $X_t^{(\epsilon)}(\cdot)$ to $X_{t+1}^{(\epsilon)}(\cdot)$, $t \geq 0$. Denote by $0 = T_0^{(\epsilon)} < T_1^{(\epsilon)} < T_2^{(\epsilon)} < \cdots < T_m^{(\epsilon)}$ the jump-times of $s \mapsto X_t^{(\epsilon)}(s)$. The trajectory $s \mapsto X_t^{(\epsilon)}(s)$ is described entirely by

$$\big\{(T_0^{(\epsilon)}, x_0), (T_1^{(\epsilon)}, X_t^{(\epsilon)}(T_1^{(\epsilon)})), \ldots, (T_m^{(\epsilon)}, X_t^{(\epsilon)}(T_m)), (T, X_t^{(\epsilon)}(T_m^{(\epsilon)}))\big\}.$$

Take $h \in (0,1)$, possibly at random.

1. Sample u uniformly on $[0, 1-h]$ and proceed by proposing a new trajectory of $s \mapsto X_t^{(\epsilon)}(s)$ in the interval $[u, u+h]$. Suppose that i and j are such that $T_i^{(\epsilon)} < u < T_{i+1}^{(\epsilon)}$ and $T_j^{(\epsilon)} < u+h < T_{j+1}^{(\epsilon)}$. The part of the process which is to be re-sampled is

$$S(t,u,u+h) = \big\{(u, X_t^{(\epsilon)}(u)), (T_{i+1}^{(\epsilon)}, X^{(\epsilon)}(T_{i+1}^{(\epsilon)})), \ldots, (T_j^{(\epsilon)}, X^{(\epsilon)}(T_j^{(\epsilon)})),$$
$$(u+h, X^{(\epsilon)}(T_j^{(\epsilon)}))\big\}.$$

2. Generate the proposal

$$\hat{S}(t,u,u+h) = \big\{(u, X_t^{(\epsilon)}(u)), (\hat{T}_{i+1}^{(\epsilon)}, \hat{X}^{(\epsilon)}(\hat{T}_{i+1}^{(\epsilon)})), \ldots, (\hat{T}_{\hat{j}}^{(\epsilon)}, X_t^{(\epsilon)}(u+h)), \\ (u+h, X_t^{(\epsilon)}(u+h))\big\},$$

by sampling from a proposal density $q(\cdot \mid X_t^{(\epsilon)}(u), X_t^{(\epsilon)}(u+h))$ that depends on $X_t^{(\epsilon)}(\cdot)$ only through $X_t^{(\epsilon)}(u)$ and $X_t^{(\epsilon)}(u+h)$. Let $\hat{X}^{(\epsilon)}$ be the trajectory that is equal to $s \mapsto X_t^{(\epsilon)}(s)$ for $s \in [0,u] \cup [u+h, T]$ and determined by $\hat{S}(t,u,u+h)$ on the interval $[u, u+h]$.

3. The proposed trajectory is accepted with probability α given by

$$\alpha = 1 \wedge \Bigg[\frac{e^{-\frac{1}{\epsilon}h_0(\hat{X}_t^{(\epsilon)})} \exp\{-\nu_{\hat{X}_t^{(\epsilon)}(\hat{T}^{(\epsilon)}_{\hat{N}^{(\epsilon)}(T)})}(\mathbb{R}^d)(T - \hat{T}^{(\epsilon)}_{\hat{N}^{(\epsilon)}(T)})\}}{e^{-\frac{1}{\epsilon}h_0(X_t^{(\epsilon)})} \exp\{-\nu_{X_t^{(\epsilon)}(T^{(\epsilon)}_{N^{(\epsilon)}(T)})}(\mathbb{R}^d)(T - T^{(\epsilon)}_{\hat{N}^{(\epsilon)}(T)})\}} \\ \times \frac{\prod_{l=1}^{\hat{N}^{(\epsilon)}(T)} \exp\{-\nu_{\hat{X}_t^{(\epsilon)}(\hat{T}_l^{(\epsilon)})}(\mathbb{R}^d)(\hat{T}_{l+1}^{(\epsilon)} - \hat{T}_l^{(\epsilon)})\}}{\prod_{l=1}^{N^{(\epsilon)}(T)} \exp\{\nu_{X_t^{(\epsilon)}(T_l^{(\epsilon)})}(\mathbb{R}^d)(T_{l+1}^{(\epsilon)} - T_{l+1}^{(\epsilon)})\}} \\ \times \frac{\prod_{l=1}^{\hat{N}^{(\epsilon)}(T)} d\nu_{\hat{X}_t^{(\epsilon)}(\hat{T}_l^{(\epsilon)})}\Big(\frac{1}{\epsilon}(\hat{X}_t^{(\epsilon)}(\hat{T}_{l+1}^{(\epsilon)}) - \hat{X}_t^{(\epsilon)}(\hat{T}_l^{(\epsilon)}))\Big)}{\prod_{l=1}^{N^{(\epsilon)}(T)} d\nu_{X_t^{(\epsilon)}(T_l^{(\epsilon)})}\Big(\frac{1}{\epsilon}(X_t^{(\epsilon)}(T_{l+1}^{(\epsilon)}) - X_t^{(\epsilon)}(T_l^{(\epsilon)}))\Big)} \\ \times \frac{q(S(t,u,u+h) \mid X_t^{(\epsilon)}(u), X^{(\epsilon)})}{q(\hat{S}(t,u,u+h) \mid X_t^{(\epsilon)}(u), X^{(\epsilon)})} \Bigg],$$

where $\hat{T}_k^{(\epsilon)}$, $k \geq 1$, denotes the jump times of $\hat{X}_t^{(\epsilon)}(\cdot)$ and $\hat{N}^{(\epsilon)}(T) = \sup\{n : \hat{T}_n^{(\epsilon)} \leq T\}$. If accepted, put $X_{t+1}^{(\epsilon)} = \hat{X}_t^{(\epsilon)}$, otherwise, put $X_{t+1}^{(\epsilon)} = X_t^{(\epsilon)}$.

Iterate steps $1-3$ until the entire Markov chain $(X_t^{(\epsilon)})_{t=0}^{T-1}$ is constructed.

5.4.2 Design and rare-event efficiency

The Hamiltonian H is defined by

$$H(x,p) = \int_{\mathbb{R}^d} (e^{\langle p,y\rangle} - 1)\nu_x(dy), \quad x,p \in \mathbb{R}^d.$$

The following condition will be assumed throughout this section.

Condition C.

(i) For each $p \in \mathbb{R}^d$, $\sup_{x\in\mathbb{R}^d} H(x,p) < \infty$.

(ii) The function $(x,p) \mapsto H(x,p)$ is continuous.

For each $x \in \mathbb{R}^d$ let and $L(x,\cdot)$ denote the Fenchel-Legendre transform of $H(x,\cdot)$,

$$L(x,v) = \sup_{p\in\mathbb{R}^d}\{\langle p,v\rangle - H(x,p)\}.$$

Condition D. [c.f. Condition 10.2.4. in Dupuis/Eliis [5]]. Let T_x consist of all points of the form $b_x + y$ where $b_x = -\int z\nu_x(dz)$ and y belongs to the smallest convex cone that contains the support of ν_x. Suppose that

(a) the sets T_x are independent of $x \in \mathbb{R}^d$,

(b) 0 belongs to the interior of T_x.

Under Condition C and Condition D it follows from Theorem 10.2.6 in [5] that the Markov chain $X^{(\epsilon)}$ satisfies the Laplace principle

$$\begin{aligned}&-\epsilon\log\mathbf{E}[\exp\{-\frac{1}{\epsilon}h(X^{(\epsilon)})\}]\\ &\quad=\inf\Big\{\int_0^1 L(\psi(s),\dot\psi(s))ds + h(\psi);\psi\in\mathscr{AC}([0,T];\mathbb{R}^d),\psi(0)=x_0\Big\},\end{aligned}$$

for each bounded continuous $h : \mathscr{C}([0,1];\mathbb{R}^d) \to \mathbb{R}$ and so in particular for h_0.

Let $V^{(\epsilon)}$ be the probability distribution given by

$$V^{(\epsilon)}(\cdot) = \frac{\mathbf{E}\big(I\{\bar X^{(\epsilon)}\in\cdot\}\exp\{-\frac{1}{\epsilon}h_1(\bar X^{(\epsilon)})\}\big)}{\bar\theta^{(\epsilon)}},$$

where $h_1 : \mathscr{D}([0,1];\mathbb{R}^d) \to \mathbb{R}^d$ is bounded and continuous, $\bar\theta^{(\epsilon)} = \mathbf{E}[\exp\{-\frac{1}{\epsilon}h_1(\bar X^{(\epsilon)})\}]$ and the $\{\bar X_t^{(\epsilon)}; t \geq 0\}$ is a Markov chain with generator $\bar{\mathscr{A}}$ given by

$$\bar{\mathscr{A}}f(x) = \frac{1}{\epsilon}\int_{\mathbb{R}^d} f(x+\epsilon y) - f(x)\bar\nu_x(dy),$$

where $\{\bar\nu_x, x \in \mathbb{R}^d\}$ is a family of measures satisfying (5.6). We assume that h_1 and $\bar\nu_x$ are taken such that $V^{(\epsilon)} \ll F_{h_0}^{(\epsilon)}$ and sufficiently simple so that the expectation $\bar\theta^{(\epsilon)}$ can be computed explicitly. This assumption is crucial for the applicability of the MCMC methodology and clearly limits the choice of $V^{(\epsilon)}$ considerably. In fact, in the examples that follow $\bar\nu_x$ will be state-independent in the sense that $\bar\nu_x(dy) = \bar\nu(dy)$.

The corresponding stochastic kernel is given by

$$\bar\Theta(dt\,dy \mid \bar x) = \frac{1}{\epsilon}e^{-\bar\nu_{\bar x}(\mathbb{R}^d)t}dt\,\bar\nu_{\bar x}(dy).$$

We define the associated Hamiltonian $\bar H$ by

$$\bar H(\bar x,p) = \int_{\mathbb{R}^d}(e^{\langle p,y\rangle}-1)\bar\nu_{\bar x}(dy).$$

We will assume that Condition C and Condition D are satisfied for $\bar{H}$ so that the Laplace principle holds for $\bar{X}^{(\epsilon)}$ as well. In addition we assume that h_1 and $\bar{\nu}_{\bar{x}}$ are taken such that the large deviations rate of $\bar{\theta}^{(\epsilon)}$ coincides with that of $\theta^{(\epsilon)}$. More precisely, that

$$\begin{aligned} I(h_0) &:= \inf\Big\{\int_0^1 L(\psi(s),\dot{\psi}(s))ds + h_0(\psi); \psi \in \mathscr{AC}[0,1], \psi(0) = x_0\Big\} \\ &= -\epsilon \log \theta^{(\epsilon)} \\ &= -\epsilon \log \bar{\theta}^{(\epsilon)} \\ &= \inf\Big\{\int_0^1 \bar{L}(\psi(s),\dot{\psi}(s))ds + h_1(\psi); \psi \in \mathcal{AC}[0,1], \psi(0) = x_0\Big\} \\ &=: \bar{I}(h_1). \end{aligned}$$

Let $T_0^{(\epsilon)} = 0 < T_1^{(\epsilon)} < T_2^{(\epsilon)} \dots$ denote the jump times of $X^{(\epsilon)}$and $\tau_k^{(\epsilon)} = T_k^{(\epsilon)} - T_{k-1}^{(\epsilon)}$, $k \geq 1$, the times between the jumps. With $N^{(\epsilon)} = \sup\{n : T_n < T\}$, and $\bar{F}^{(\epsilon)}$ denoting the distribution of $\bar{X}^{(\epsilon)}$ on $\mathscr{D}([0,1];\mathbb{R}^d)$, it follows that the likelihood ratio can be written as

$$\begin{aligned} &\frac{dV^{(\epsilon)}}{dF^{(\epsilon)}}(X^{(\epsilon)}) \\ &\quad = \frac{e^{-\frac{1}{\epsilon}h_1(X^{(\epsilon)})}}{\bar{\theta}^{(\epsilon)}} \frac{d\bar{F}^{(\epsilon)}}{dF^{(\epsilon)}}(X^{(\epsilon)}) \\ &\quad = \frac{e^{-\frac{1}{\epsilon}h_1(x)}}{\bar{\theta}^{(\epsilon)}} \exp\Big\{\int_0^1 \Big(\nu_{X^{(\epsilon)}(t)}(\mathbb{R}^d) - \bar{\nu}_{X^{(\epsilon)}(t)}(\mathbb{R}^d)\Big)dt \\ &\qquad + \sum_{k=1}^{N^{(\epsilon)}} \log\Big(\frac{d\bar{\nu}_{X^{(\epsilon)}(T_{k-1})}}{d\nu_{X^{(\epsilon)}(T_{k-1})}}\Big(\frac{1}{\epsilon}\Delta X_k^{(\epsilon)}\Big)\Big\}, \end{aligned}$$

where $\Delta X_k^{(\epsilon)} = X^{(\epsilon)}(T_k) - X^{(\epsilon)}(T_{k-1})$, $k \geq 1$.

For each $\bar{x} \in \mathbb{R}^d$, $p_1 \in \mathbb{R}^d, p_2 \in \mathbb{R}$ let

$$\mathbf{H}(\bar{x}, p_1, p_2) = \Big(\bar{\nu}_{\bar{x}}(\mathbb{R}^d) - \nu_{\bar{x}}(\mathbb{R}^d)\Big)p_2 + \int_{\mathbb{R}^d}\Big[e^{\langle p_1, z\rangle}\Big(\frac{d\bar{\nu}_{\bar{x}}}{d\nu_{\bar{x}}}(z)\Big)^{p_2} - 1\Big]\bar{\nu}_{\bar{x}}(dz),$$

and define $\mathbf{L}$ as the Fenchel-Legendre transform of $\mathbf{H}$:

$$\mathbf{L}(x, v_1, v_2) = \sup_{p_1 \in \mathbb{R}^d, p_2 \in \mathbb{R}} \{\langle p_1, v_1\rangle + p_2, v_2 - \mathbf{H}(x, p_1, p_2)\}.$$

Theorem 5.4.2. *Suppose that*

$$\sup_{x \in \mathbb{R}^d} \mathbf{H}(x, p_1, p_2) < \infty, \quad \textit{for every } p \in \mathbb{R}^d, p_2 \in \mathbb{R}. \tag{5.7}$$

Then

$$\limsup_{\epsilon \to 0} \epsilon \log\Big((\theta^{(\epsilon)})^2 \mathbf{E}_{F_{h_0}^{(\epsilon)}}[u(X^{(\epsilon)})^2]\Big) \leq I(h_0) - J(2h_1 - h_0),$$

where, for each bounded continuous $h : \mathscr{D}([0,1];\mathbb{R}^d) \to \mathbb{R}$,

$$J(h) = \inf\Big\{ \int_0^1 \mathbf{L}(\psi(s), \dot{\psi}_1(s), -\dot{\psi}_2(s))ds + h(\psi_1) + \psi_2(1),$$
$$\psi \in \mathcal{AC}([0,1];\mathbb{R}^{d+1}), \psi(0) = (x_0, 0)\Big\}.$$

Proof. The calculation leading up to (5.2) shows that

$$\limsup_{\epsilon\to 0} \epsilon \log\Big((\theta^{(\epsilon)})^2 \mathbf{E}_{F_{h_0}^{(\epsilon)}}[(\bar{q}_T^{(\epsilon)})^2]\Big)$$
$$\le -I(h_0) + \limsup_{\epsilon\to 0} \epsilon \log \mathbf{E}_{V^{(\epsilon)}}\Big[\exp\{\frac{1}{\epsilon}h_0(X^{(\epsilon)})\}\frac{dV^{(\epsilon)}}{dF^{(\epsilon)}}(X^{(\epsilon)})\Big].$$

Since

$$\limsup_{\epsilon\to 0} \epsilon \log \mathbf{E}_{V^{(\epsilon)}}\Big[\exp\{\frac{1}{\epsilon}h_0(X^{(\epsilon)})\}\frac{dV^{(\epsilon)}}{dF^{(\epsilon)}}(X^{(\epsilon)})\Big]$$
$$= \limsup_{\epsilon\to 0} \epsilon \log \frac{1}{(\bar{\theta}^{(\epsilon)})^2}\mathbf{E}\Big[\exp\Big\{\frac{1}{\epsilon}\Big(h_0(\bar{X}^{(\epsilon)}) - 2h_1(\bar{X}^{(\epsilon)})\Big)\Big\}\frac{d\bar{F}^{(\epsilon)}}{dF^{(\epsilon)}}(\bar{X}^{(\epsilon)})\Big]$$
$$\le \limsup_{\epsilon\to 0} -2\epsilon \log \bar{\theta}^{(\epsilon)}$$
$$+ \limsup_{\epsilon\to 0} \epsilon \log \mathbf{E}\Bigg[\exp\Big\{-\frac{1}{\epsilon}\Big(2h_1(\bar{X}^{(\epsilon)}) - h_0(\bar{X}^{(\epsilon)})\Big)\Big\}$$
$$\times \exp\Big\{\int_0^1 (\nu_{\bar{X}^{(\epsilon)}(t)}(\mathbb{R}^d) - \bar{\nu}_{\bar{X}^{\epsilon}(t))}(\mathbb{R}^d))dt$$
$$+ \sum_{k=1}^{\bar{N}^{(\epsilon)}} \log\Big(\frac{d\bar{\nu}_{\bar{X}^{(\epsilon)}(T_{k-1})}}{d\nu_{\bar{X}^{(\epsilon)}(T_{k-1})}}\big(\frac{1}{\epsilon}\Delta\bar{X}_k^{(\epsilon)}\big)\Big)\Big\}\Bigg]$$
$$= 2I(h_0) - \liminf_{\epsilon\to 0} -\epsilon \log \mathbf{E}\Bigg[\exp\Big\{-\frac{1}{\epsilon}\Big(2h_1(\bar{X}^{(\epsilon)}) - h_0(\bar{X}^{(\epsilon)})\Big)\Big\}$$
$$\times \exp\Big\{\int_0^1 (\nu_{\bar{X}^{(\epsilon)}(t)}(\mathbb{R}^d) - \bar{\nu}_{\bar{X}^{\epsilon}(t))}(\mathbb{R}^d))dt$$
$$+ \sum_{k=1}^{\bar{N}^{(\epsilon)}} \log\Big(\frac{d\bar{\nu}_{\bar{X}^{(\epsilon)}(T_{k-1})}}{d\nu_{\bar{X}^{(\epsilon)}(T_{k-1})}}\big(\frac{1}{\epsilon}\Delta\bar{X}_k^{(\epsilon)}\big)\Big)\Big\}\Bigg],$$

it follows that it is sufficient to prove that

$$\liminf_{n\to\infty} -\epsilon \log \mathbf{E}\Bigg[\exp\Big\{ -\frac{1}{\epsilon}\Big(2h_1(\bar{X}^{(\epsilon)}) - h_0(\bar{X}^{(\epsilon)})\Big)\Big\} \tag{5.8}$$

$$\times \exp\Big\{ \int_0^1 \Big(\nu_{\bar{X}^{(\epsilon)}(t)}(\mathbb{R}^d) - \bar{\nu}_{\bar{X}^{\epsilon}(t))}(\mathbb{R}^d)\Big)dt$$

$$+ \sum_{k=1}^{\bar{N}^{(\epsilon)}} \log\Big(\frac{d\bar{\nu}_{\bar{X}^{(\epsilon)}(T_{k-1})}}{d\nu_{\bar{X}^{(\epsilon)}(T_{k-1})}}\big(\frac{1}{\epsilon}\Delta\bar{X}_k^{(\epsilon)}\big)\Big)\Big\}\Bigg] \geq J(2h_1 - h_0). \tag{5.9}$$

Let us introduce the notation $\bar{Z}^{(\epsilon)}(t) = (\bar{X}^{(\epsilon)}(t), \bar{Y}^{(\epsilon)}(t))$ where

$$\bar{Y}^{(\epsilon)}(t) = -\epsilon \int_0^1 \Big(\nu_{\bar{X}^{(\epsilon)}(t)}(\mathbb{R}^d) - \bar{\nu}_{\bar{X}^{\epsilon}(t))}(\mathbb{R}^d)\Big)dt$$

$$- \epsilon \sum_{k=1}^{\bar{N}^{(\epsilon)}} \log\Big(\frac{d\bar{\nu}_{\bar{X}^{(\epsilon)}(T_{k-1})}}{d\nu_{\bar{X}^{(\epsilon)}(T_{k-1})}}\big(\frac{1}{\epsilon}\Delta\bar{X}_k^{(\epsilon)}\big)\Big).$$

Note that $\bar{Z}^{(\epsilon)}$ is a continuous time Markov chain with generator

$$\mathscr{A}_{\bar{Z}^{(\epsilon)}} f(x,y) = (\bar{\nu}_x(\mathbb{R}^d) - \nu_x(\mathbb{R}^d))\frac{\partial f}{\partial y}(x,y)$$

$$+ \frac{1}{\epsilon}\int_{\mathbb{R}^d} \Big(f(x+\epsilon z, y - \epsilon \log\frac{d\bar{\nu}_x}{d\nu_x}(z)) - f(x,y)\Big)\bar{\nu}_x(dz).$$

The Hamiltonian associated with $\bar{Z}^{(\epsilon)}$ is given by $H_{\bar{Z}}(\bar{x}, p_1, p_2) = \mathbf{H}(\bar{x}, p_1, -p_2)$.

The assumption (5.7) implies that $H_{\bar{Z}}$ satisfies Condition C and hence, by Theorem 10.2.6 in [5] the Laplace principle upper bound holds for $\bar{Z}^{(\epsilon)}$ (it may be observed that only Condition C is used in the proof of the Laplace principle upper bound):

$$\lim_{\epsilon\to 0} -\epsilon \log \mathbf{E}\Big[e^{-n\frac{1}{\epsilon}\bar{h}(\bar{Z}^{(\epsilon)})}\Big]$$

$$\geq \inf\Big\{ \int_0^1 L_{\bar{Z}}\big(\psi(s), \dot{\psi}_1(s), \dot{\psi}_2(s)\big)ds + \bar{h}(\psi); \psi \in \mathcal{AC}[0,1], \psi(0) = (x_0, 0)\Big\}$$

$$= \inf\Big\{ \int_0^1 \mathbf{L}\big(\psi(s), \dot{\psi}_1(s), -\dot{\psi}_2(s)\big)ds + \bar{h}(\psi); \psi \in \mathcal{AC}[0,1], \psi(0) = (x_0, 0)\Big\}, \tag{5.10}$$

for each bounded continuous function $\bar{h} : \mathscr{D}([0,1]; \mathbb{R}^d \times \mathbb{R}) \to \mathbb{R}$. We would like to apply the Laplace principle with the function $\bar{h}(x,y) = 2h_1(x) - h_0(x) + y(1)$, but this function is not bounded. It is, nevertheless, possible to apply the Laplace principle with this function. Indeed, by Theorem 1.3.4 in [5] (5.5) holds for this choice of $\bar{h}$ if

$$\lim_{C\to\infty} \limsup_{\epsilon\to 0} \epsilon \log \mathbf{E}\Big[I\Big\{\bar{Y}^{(\epsilon)}(1) < -C\Big\} e^{-\frac{1}{\epsilon}\bar{Y}^{(\epsilon)}(1)}\Big] = -\infty. \tag{5.11}$$

To see that this is indeed true, write

$$\lim_{C\to\infty} \limsup_{\epsilon\to 0} \epsilon \log \mathbf{E}\left[I\left\{ \bar{Y}^{(\epsilon)}(1) < -C \right\} e^{-\frac{1}{\epsilon}\bar{Y}^{(\epsilon)}(1)} \right]$$
$$= \lim_{C\to\infty} \limsup_{\epsilon\to 0} \epsilon \log \mathbf{E}\left[I\left\{ -\frac{1}{\epsilon}\bar{Y}^{(\epsilon)}(1) > \frac{C}{\epsilon} \right\} e^{-\frac{1}{\epsilon}\bar{Y}^{(\epsilon)}(1)} \right].$$

For any $p > 0$, by Chebyshev's inequality, the expression in the last display is less than or equal to

$$\lim_{C\to\infty} \limsup_{\epsilon\to 0} \epsilon \log \mathbf{E}\left[e^{-\frac{pC}{\epsilon}} e^{-\frac{p+1}{\epsilon}\bar{Y}^{(\epsilon)}(1)} \right]$$
$$= \lim_{C\to\infty} -pC + \limsup_{\epsilon\to 0} \epsilon \log \mathbf{E}\left[e^{-\frac{p+1}{\epsilon}\bar{Y}^{(\epsilon)}(1)} \right].$$

Condition C implies that the latter term is finite, the argument follows shortly, and we conclude that the limit is $-\infty$. This proves (5.11) and completes the proof.

To show that $\limsup_{\epsilon\to 0} \epsilon \log \mathbf{E}\left[e^{-\frac{p+1}{\epsilon}\bar{Y}^{(\epsilon)}(1)} \right] < \infty$, let us introduce the notation

$$\theta_1 = (p+1) \sup_{\bar{x}} \{ \bar{\nu}_{\bar{x}}(\mathbb{R}^d) - \nu_{\bar{x}}(\mathbb{R}^d) \} < \infty,$$
$$\theta_2 = \sup_{\bar{x}} \int \left(\frac{d\bar{\nu}_{\bar{x}}}{d\nu_{\bar{x}}}(z) \right)^{p+1} \bar{\nu}_{\bar{x}}(dz).$$

Both quantities are finite by Condition C. It follows that

$$
\begin{aligned}
&\limsup_{\epsilon\to 0}\epsilon\log\mathbf{E}\left[e^{-\frac{p+1}{\epsilon}\bar{Y}^{(\epsilon)}(1)}\right]\\
&\quad=\limsup_{\epsilon\to 0}\epsilon\log\mathbf{E}\left[\exp\Big\{(p+1)\int_0^1\Big(\nu_{\bar{X}^{(\epsilon)}(t)}(\mathbb{R}^d)-\bar{\nu}_{\bar{X}^{\epsilon}(t))}(\mathbb{R}^d)\Big)dt\Big\}\right.\\
&\qquad\left.\times\Big(\prod_{k=1}^{\bar{N}^{(\epsilon)}}\Big(\frac{d\bar{\nu}_{\bar{X}^{(\epsilon)}(T_{k-1})}}{d\nu_{\bar{X}^{(\epsilon)}(T_{k-1})}}\Big(\frac{1}{\epsilon}\Delta\bar{X}_k^{(\epsilon)}\Big)\Big)^{p+1}\right]\\
&\quad\le\limsup_{\epsilon\to 0}\epsilon\log\mathbf{E}\left[e^{\theta_1}\Big(\prod_{k=1}^{\bar{N}^{(\epsilon)}}\Big(\frac{d\bar{\nu}_{\bar{X}^{(\epsilon)}(T_{k-1})}}{d\nu_{\bar{X}^{(\epsilon)}(T_{k-1})}}\Big(\frac{1}{\epsilon}\Delta\bar{X}_k^{(\epsilon)}\Big)\Big)^{p+1}\right]\\
&\quad\le\limsup_{\epsilon\to 0}\epsilon\log\mathbf{E}\Big[\Big(\prod_{k=1}^{\bar{N}^{(\epsilon)}(1)-1}\Big(\frac{d\bar{\nu}_{\bar{X}^{(\epsilon)}(T_{k-1})}}{d\nu_{\bar{X}^{(\epsilon)}(T_{k-1})}}\Big(\frac{1}{\epsilon}\Delta\bar{X}_k^{(\epsilon)}\Big)\Big)^{p+1}\\
&\qquad\times\mathbf{E}\Big[\Big(\frac{d\bar{\nu}_{\bar{X}^{(\epsilon)}(T_{\bar{N}^{(\epsilon)}(1)-1})}}{d\nu_{\bar{X}^{(\epsilon)}(T_{\bar{N}^{(\epsilon)}(1)-1})}}\Big(\frac{1}{\epsilon}\Delta\bar{X}^{(\epsilon)}_{\bar{N}^{(\epsilon)}(1)-1}\Big)\Big)^{p+1}\mid\bar{X}^{(\epsilon)}(T_{N^{(\epsilon)}(1)t})\Big]\Big]\\
&\quad\le\limsup_{\epsilon\to 0}\epsilon\log\mathbf{E}\Big[\Big(\prod_{k=1}^{\bar{N}^{(\epsilon)}(1)-1}\Big(\frac{d\bar{\nu}_{\bar{X}^{(\epsilon)}(T_{k-1})}}{d\nu_{\bar{X}^{(\epsilon)}(T_{k-1})}}\Big(\frac{1}{\epsilon}\Delta\bar{X}_k^{(\epsilon)}\Big)\Big)^{p+1}\theta_2\Big]\\
&\quad\le\cdots\\
&\quad\le\limsup_{\epsilon\to 0}\epsilon\log\mathbf{E}[\theta_2^{N^{(\epsilon)}(1)}].
\end{aligned}
$$

The number of jumps, $N^{(\epsilon)}(T)$, of the Markov chain is stochastically bounded above by a Poisson distributed random variable N^* with parameter $\frac{1}{\epsilon}\sup_{\bar{x}}\bar{\nu}_{\bar{x}}(\mathbb{R}^d)$. Therefore, the expression in the last display is bounded above by

$$
\limsup_{\epsilon\to 0}\epsilon\log\mathbf{E}[\theta_2^{N^*}]=\limsup_{\epsilon\to 0}\epsilon\frac{1}{\epsilon}\sup_{\bar{x}}\bar{\nu}_{\bar{x}}(\mathbb{R}^d)(\theta_2-1)<\infty.
$$

This completes the proof. □

5.5 An application to a birth-and-death process

Let $\{Q^{(n)}(t);t\ge 0\}$, $Q^{(n)}(0)=nx_0\in\{1,2,\dots,n\}$, denote a one-dimensional birth-and-death process, taking values in $\{0,\dots,n\}$ and let $0=T_0^{(n)}<T_1^{(n)}<\dots$ denote the jump times of $Q^{(n)}$. That is, $Q^n(T_{k+1})=Q^n(T_k)\pm 1$, for every $k\ge 0$. The birth intensity is given by $n\lambda(x)$ and the death intensity by $n\mu(x)$ where λ and μ are bounded Lipschitz continuous function on $[0,1]$, with $\mu(0)=\lambda(0)=0$ so that 0 is an absorbing state. Let $\epsilon_n=n^{-1}$ and put $X^{(\epsilon_n)}(t)=Q^{(n)}(t)/n$. Then $X^{(\epsilon_n)}=(X^{(\epsilon_n)}(t);t\ge 0)$,

$X^{(\epsilon_n)}(0) = x_0$, is a continuous time Markov chain with values in $[0,1]$ with generator

$$\mathscr{A}^{(\epsilon_n)} f(x) = \frac{1}{\epsilon_n} \int_{\mathbb{R}} f(x + \epsilon_n y) - f(x) \nu_x(dy),$$

where $\nu_x(dy) = \lambda(x)\delta_1(dy) + \mu(x)\delta_{-1}(dy)$ and $\delta_y(A) = 1$ if $y \in A$ and 0 otherwise. The stochastic kernel associated with $X^{(\epsilon_n)}$ is given by

$$\Theta^{(\epsilon_n)}(dy, dy \mid x) = \frac{1}{\epsilon_n} \Big(\lambda(x)\delta_1(dy) + \mu(x)\delta_{-1}(dy) \Big) e^{-(\mu(x)+\lambda(x))t} dt.$$

Let $a > x_0$ and $\tau_a^{(\epsilon_n)} = \inf\{t > 0 : X^{(\epsilon_n)}(t) \geq a\}$ be the first time the process exceeds a. We will be interested in computing a rare-event probability of the form

$$p^{(\epsilon_n)} = \mathbf{P}(\tau_a^{(\epsilon_n)} \leq 1).$$

Define the Hamiltonian H by

$$H(x,p) = \lambda(x)(e^p - 1) + \mu(x)(e^{-p} - 1),$$

and L as the Fenchel-Legendre transform of H:

$$L(x,v) = \sup_{p \in \mathbb{R}^d} \Big[\langle p, v \rangle - H(x,p) \Big].$$

Since λ and μ are assumed to be bounded and Lipschitz continuous Condition C is satisfied. Since the smallest convex cone containing $-1, 1$ is $\mathbb{R}$ Condition D is also satisfied and we conclude that the Laplace principle holds for $X^{(\epsilon_n)}$.

The sought probability $p^{(\epsilon_n)}$ can be written as an expectation of the form

$$\mathbf{E}_{x_0}[\exp\{-\frac{1}{\epsilon_n} h_0(X^{(\epsilon_n)})\}],$$

where

$$h_0(x) = \begin{cases} 0, & \text{if } x(t) \geq a, \text{ for some } t \in [0,1], \\ \infty, & \text{otherwise.} \end{cases}$$

In Section 5.4 we treated expectations of this form for bounded and continuous h. By approximating h_0 by a sequence of bounded continuous functions standard arguments show that the results in Section 5.4 can be extended to cover this case, see e.g. [5], proof of Theorem 1.2.3, pp. 10-11.

5.5.1 The design of $V^{(\epsilon_n)}$

Let us use the methodology in Section 5.4 and specify the MCMC estimator by proposing a probability distribution $V^{(\epsilon_n)}$ as follows. Let $\{\bar{X}^{(\epsilon_n)}(t); t \geq 0\}$, $\bar{X}^{(\epsilon_n)}(0) = x_0$, be a Markov chain with generator $\bar{\mathscr{A}}$ given by

$$\bar{\mathscr{A}} f(x) = \frac{1}{\epsilon_n} \int_{\mathbb{R}^d} f(x + \epsilon_n y) - f(x) \bar{\nu}_x(dy),$$

and $\bar{\nu}_x(dy) = \bar{\lambda}\delta_1(dy) + \bar{\mu}\delta_{-1}(dy)$. That is, $\bar{X}^{(\epsilon_n)}$ is a continuous time random walk with intensity $n\bar{\lambda}$ for upwards jumps and $n\bar{\mu}$ for downwards jumps. Put $\bar{\tau}_a^{(\epsilon_n)} = \inf\{t > 0 : \bar{X}^{(\epsilon_n)}(t) \geq a\}$ and let

$$V^{(\epsilon_n)}(\cdot) = \mathbf{P}\big(\bar{X}^{(\epsilon_n)} \in \cdot \mid \bar{\tau}_a^{(\epsilon_n)} \leq 1\big).$$

The corresponding stochastic kernel is given by

$$\bar{\Theta}(dt\, dy) = \frac{1}{\epsilon_n}(\bar{\lambda}\delta_1(dy) + \bar{\mu}\delta_{-1}(dy))e^{-(\bar{\lambda}+\bar{\mu})t}dt,$$

the associated Hamiltonian $\bar{H}$ is given by

$$\bar{H}(p) = \bar{\lambda}(e^p - 1) + \bar{\mu}(e^{-p} - 1).$$

Given a trajectory $t \mapsto X^{(\epsilon_n)}(t)$, $0 \leq t \leq 1$, which is described entirely by

$$\big\{(T_0^{(\epsilon_n)}, x_0), (T_1^{(\epsilon_n)}, X^{(\epsilon_n)}(T_1^{(\epsilon_n)}), \ldots, (T_m^{(\epsilon_n)}, X^{(\epsilon_n)}(T_m)), (1, X^{(\epsilon_n)}(T_m^{(\epsilon_n)}))\big\},$$

the likelihood ratio is given by

$$\begin{aligned}\frac{dV^{(\epsilon_n)}}{dF^{(\epsilon_n)}}(X^{(\epsilon_n)}) = \frac{1}{\bar{p}^{(\epsilon_n)}} \prod_{k=1}^{m} \exp\Big\{ - \Big(\bar{\nu}_{X^{(\epsilon_n)}(T_{k-1}^{(\epsilon_n)})}(\mathbb{R}) - \nu_{X^{(\epsilon_n)}(T_{k-1}^{(\epsilon_n)})}(\mathbb{R})\Big)\Big(T_k^{(\epsilon_n)} - T_{k-1}^{(\epsilon_n)}\Big)\Big\} \\ \times \prod_{k=1}^{m} \frac{d\bar{\nu}_{X^{(\epsilon_n)}(T_{k-1}^{(\epsilon_n)})}}{d\nu_{X^{(\epsilon_n)}(T_{k-1}^{(\epsilon_n)})}}\Big(X^{(\epsilon_n)}(T_k^{(\epsilon_n)}) - X^{(\epsilon_n)}(T_{k-1}^{(\epsilon_n)})\Big) \\ \times \exp\Big\{ - \Big(\bar{\nu}_{X^{(\epsilon_n)}(T_m^{(\epsilon_n)})}(\mathbb{R}) - \nu_{X^{(\epsilon_n)}(T_m^{(\epsilon_n)})}(\mathbb{R})\Big)\Big(1 - T_m^{(\epsilon_n)}\Big)\Big\},\end{aligned}$$

where

$$\bar{p}^{(\epsilon_n)} = \mathbf{P}(\bar{\tau}_a^{(\epsilon_n)} \leq 1).$$

Denote by $u_{z,n}$ the probability that a simple random walk, starting at z, with absorption in 0 and a, ends up at 0 after exactly n increments. The explicit formula for $u_{z,n}$, as derived by Feller [6][Ch. XIV, Sec. 5, p. 353], is given by

$$u_{z,n} = a^{-1}2^n p^{\frac{1}{2}(n-z)} q^{\frac{1}{2}(n+z)} \sum_{j=1}^{a-1} \cos^{n-1}\Big(\frac{\pi j}{a}\Big) \sin\Big(\frac{\pi j}{a}\Big) \sin\Big(\frac{\pi z j}{a}\Big),$$

where p is the probability to jump up and q is the probability to jump down. The explicit formula for the probability of ruin at the nth trial, $u_{z,n}$, goes back to Lagrange and can be found in the literature from the 19th century.

In our continuous time setting, consider a birth-and-death process with absorption in 0 and a, starting at x_0, with birth intensity $\bar{\lambda}$ and death intensity $\bar{\mu}$. Let N be the number of jumps taken until it ends up at a after exactly N increments. Then,

$$\mathbf{P}(N = n) = u_{a-x_0,n},$$

with $p = \bar{\mu}/(\bar{\lambda}+\bar{\mu})$ and $q = \bar{\lambda}/(\bar{\lambda}+\bar{\mu})$. Note that because we want to end up at a we have to flip the "Feller random walk" upside down.

Letting $Z_1, Z_2, \ldots$ be iid Exp$(\lambda+\mu)$ we get that the time $\bar{\tau}_a^{(\epsilon_n)}$ until exit at a satisfies

$$\begin{aligned}\bar{p}^{(\epsilon_n)} = \mathbf{P}(\bar{\tau}_a^{(\epsilon_n)} \le 1) &= \sum_{n=1}^{\infty} \mathbf{P}\Big(\sum_{k=1}^{n} Z_k \le 1\Big)\mathbf{P}(N=n) \\ &= \sum_{n=1}^{\infty} \mathbf{P}(\gamma_{n,\bar{\lambda}+\bar{\mu}} \le 1) u_{a-x_0,n},\end{aligned}$$

where $\gamma_{\alpha,\beta}$ denotes a random variable with gamma(α,β)-distribution.

5.5.2 The MCMC algorithm

In this section the MCMC algorithm is presented which generates a Markov chain $(X_t^{(\epsilon_n)}(s), 0 \le s \le 1)_{t\ge 0}$ whose invariant distribution is $F_{h_0}^{(\epsilon_n)}(\cdot) = \mathbf{P}(X^{(\epsilon_n)} \in \cdot \mid \tau_a^{(\epsilon_n)} \le 1)$.

Algorithm 5.5.1. The transition from $X_t^{(\epsilon_n)}(\cdot)$ to $X_{t+1}^{(\epsilon_n)}(\cdot)$, $t \ge 0$. Denote by $0 = T_0^{(\epsilon)} < T_1^{(\epsilon)} < T_2^{(\epsilon)} < \cdots < T_m^{(\epsilon)}$ the jump-times of $s \mapsto X_t^{(\epsilon_n)}(s)$. The trajectory $s \mapsto X_t^{(\epsilon_n)}(s)$ is described entirely by

$$\big\{(T_0^{(\epsilon_n)}, x_0), (T_1^{(\epsilon_n)}, X_t^{(\epsilon_n)}(T_1^{(\epsilon_n)}), \ldots, (T_m^{(\epsilon_n)}, X_t^{(\epsilon_n)}(T_m)), (1, X_t^{(\epsilon_n)}(T_m^{(\epsilon_n)}))\big\}.$$

Take $h \in (0,1)$, possibly at random.

1. Sample u uniformly on $[0, 1-h]$ and proceed by proposing a new trajectory of $s \mapsto X_t^{(\epsilon_n)}(s)$ in the interval $[u, u+h]$. Suppose that $T_i^{(\epsilon_n)} < u < T_{i+1}^{(\epsilon_n)}$ and $T_j^{(\epsilon_n)} < u+h < T_{j+1}^{(\epsilon_n)}$. The part of the process which is to be re-sampled is

   $$S(t,u,u+h) = \big\{(u, X_t^{(\epsilon_n)}(T_i^{(\epsilon_n)}), (T_{i+1}^{(\epsilon_n)}, X^{(\epsilon_n)}(T_{i+1}^{(\epsilon_n)})), \ldots, (T_j^{(\epsilon_n)}, X^{(\epsilon_n)}(T_j^{(\epsilon_n)}), (u+h, X^{(\epsilon_n)}(T_j^{(\epsilon)})\big\}.$$

2. Generate the proposal

   $$\hat{S}(t,u,u+h) = \big\{(u, X_t^{(\epsilon_n)}(u), (\hat{T}_1^{(\epsilon_n)}, \hat{X}^{(\epsilon_n)}(\hat{T}_1^{(\epsilon_n)})), \ldots, (\hat{T}_k^{(\epsilon_n)}, X_t^{(\epsilon_n)}(u+h)), (u+h, X_t^{(\epsilon_n)}(u+h))\big\},$$

   as follows.

   a Let $x_i = X_t^{(\epsilon_n)}(u)$, $x_j = X_t^{(\epsilon_n)}(u+h)$ and $k = \epsilon_n^{-1}(x_j - x_i)$. Let B and D be the proposed number of births and deaths. Let B and D be independent Poisson(λ_B) and Poisson(μ_D), respectively, where $\lambda_B = \lambda([x_i + x_j]/2)$ and

$\mu_D = \mu([x_i + x_j]/2)$. Sample B and D conditioned on $B - D = k$. That is, if $k \geq 0$, sample B from

$$\mathbf{P}(B = l \mid B - D = k) = \frac{e^{-\lambda_B} \frac{\lambda_B^l}{l!} e^{-\mu_D} \frac{\mu_D^{(l-k)}}{(l-k)!}}{\sum_{l \geq k} e^{-\lambda_B} \frac{\lambda_B^l}{l!} e^{-\mu_D} \frac{\mu_D^{(l-k)}}{(l-k)!}},$$

and set $D = k - B$. Otherwise, if $k < 0$, sample D from

$$\mathbf{P}(D = l \mid D - B = -k) = \frac{e^{-\mu_D} \frac{\mu_D^l}{l!} e^{-\lambda_B} \frac{\lambda_B^{(l+k)}}{(l+k)!}}{\sum_{l \geq k} e^{-\mu_D} \frac{\mu_D^l}{l!} e^{-\lambda_B} \frac{\lambda_B^{(l+k)}}{(l+k)!}},$$

and set $B = k - D$.

b Given the number of births B and deaths D, sample the corresponding jump times independently from the uniform distribution on $(u, u + h)$.

c Let $\hat{S}(t, u, u + h)$ be the trajectory that is determined by (u, x_i), $(u + h, x_j)$, B, D, and the associated jump times on the interval $[u, u + h]$.

3. Let $s \mapsto \hat{X}_t^{\epsilon_n}(s)$, $0 \leq s \leq 1$, be the trajectory that is equal to $s \mapsto X_t^{(\epsilon_n)}(s)$ for $s \in [0, u] \cup [u + h, 1]$ and determined by $\hat{S}(t, u, u + h)$ on the interval $[u, u + h]$ and put $\hat{\tau}_a^{(\epsilon_n)} = \inf\{s > 0 : \hat{X}_t^{(\epsilon_n)}(s) \geq a\}$. Denote by g the proposal density, as described in Step 2 above. The proposed trajectory is accepted with probability α given by

$$\begin{aligned}\alpha = {} & \frac{\prod_{l=1}^k r\big(\hat{X}_t^{(\epsilon)}(\hat{T}_{l-1}^{(\epsilon)}); \hat{X}_t^{(\epsilon)}(\hat{T}_l^{(\epsilon)}) - \hat{X}_t^{(\epsilon)}(\hat{T}_{l-1}^{(\epsilon)})\big)}{\prod_{l=i}^j r\big(X_t^{(\epsilon)}(T_{l-1}^{(\epsilon)}); X_t^{(\epsilon)}(T_l^{(\epsilon)}) - X_t^{(\epsilon)}(T_{l-1}^{(\epsilon)})\big)} \\ & \times \frac{\prod_{l=1}^k e^{-R\left(\hat{X}_t^{(\epsilon)}(\hat{T}_{l-1}^{(\epsilon)})\right)(\hat{T}_l^{(\epsilon)} - \hat{T}_{l-1}^{(\epsilon)})}}{\prod_{l=i}^j e^{-R\left(X_t^{(\epsilon)}(T_{l-1}^{(\epsilon)})\right)(T_l^{(\epsilon)} - T_{l-1}^{(\epsilon)})}} \\ & \times \frac{e^{-R\left(\hat{X}_t^{(\epsilon)}(\hat{T}_k^{(\epsilon)})\right)(1 - \bar{T}_k^{(\epsilon)})}}{e^{-R\left(X_t^{(\epsilon)}(T_j^{(\epsilon)})\right)(1 - T_j^{(\epsilon)})}} \\ & \times \frac{g(S(t, u, u + h) \mid X_t^{(\epsilon)}(u), X_t^{(\epsilon)}(u + h))}{g(\hat{S}(t, u, u + h) \mid X_t^{(\epsilon)}(u), X_t^{(\epsilon)}(u + h))} I\{\hat{\tau}_a^{(\epsilon_n)} \leq 1\} \wedge 1\end{aligned}$$

where $T_0^{(\epsilon)} = u$ and

$$r(x; 1) = \lambda(x), \quad r(x; -1) = \mu(x), \quad R(x) = \lambda(x) + \mu(x).$$

If accepted, we put $X_{t+1}^{(\epsilon_n)} = \hat{X}_t^{(\epsilon_n)}$. Otherwise, $X_{t+1}^{(\epsilon_n)} = X_t^{(\epsilon_n)}$.

Iterate steps $1 - 3$ until the entire Markov chain $(X_t^{(\epsilon_n)})_{t=0}^{T-1}$ is constructed.

5.5.3 Numerical experiments

This section illustrates numerical experiments for the first passage problem

$$p^{(\epsilon_n)} = \mathbf{P}(\tau^{(\epsilon_n)} \leq 10),$$

where $\tau^{(\epsilon_n)} = \inf\{t > 0 : X^{(\epsilon_n)} \geq 0.75\}$ where $\{X^{(\epsilon_n)}\}$ is a birth-death process with intensities $\lambda(x) = \rho x(1-x)$ and $\mu(x) = x$, and starts at $X^{(\epsilon_n)}(0) = 0.67$. The size of the population is denoted by n.

We set $\bar{\lambda} = \int_{0.67}^{0.75} \lambda(x)dx \approx 0.5625813$ and compute $\bar{\mu} = 0.648999$.

Table 5.1 shows the simulation estimates for computing $p^{(\epsilon_n)}$ using the MCMC estimator. The estimates are recorded based on 10 batches each consisting of 10^4 simulations.

Table 5.1: Batch estimates, standard deviation, relative error and computer runtime for computing $p^{(\epsilon_n)}$.

n	Estimate	Std. deviation	Rel. error	Comp. time(s)
50	2.111e-01	2.539e-03	0.0120	8.08
100	1.106e-02	6.449e-05	0.0058	9.25
150	1.925e-04	1.477e-06	0.0076	11.77
200	1.684e-06	1.067e-08	0.0063	13.32
250	9.496e-09	5.490e-11	0.0057	16.15

Table 5.2 shows the comparison between the simulation estimates for computing $p^{(\epsilon_n)}$ using the MCMC estimator against the standard Monte Carlo estimator. The estimates are recorded based on 10 batches each consisting of 10^4 simulations.

Table 5.2: Numerical comparison of computing $p^{(\epsilon_n)}$ between the MCMC and standard Monte Carlo.

$n = 110$	MCMC	Monte Carlo
Estimate	4.200e-03	3.700e-03
Std. deviation	3.666509e-05	266.875e-05
Rel. error	0.0087	0.7213
Comp. time(s)	0.9706	0.3396
$n = 140$	MCMC	Monte Carlo
Estimate	2.261e-04	8.000e-04
Std. deviation	7.062e-06	918.937e-06
Rel. error	0.0312	1.1487
Comp. time(s)	1.0757	0.3390

Acknowledgments

Henrik Hult's research was supported by the Swedish Research Council.

5.6 References

[1] J. Blanchet, K. Leder, and P. W. Glynn. Efficient simulation of light-tailed sums: an old-folk song sung to a faster new tune... *Monte Carlo and Quasi-Monte Carlo Methods 2008*, pages 227–248, 2009.

[2] Thomas Dean and Paul Dupuis. Splitting for rare event simulation: A large deviation approach to design and analysis. *Stochastic processes and their applications*, 119(2):562–587, 2009.

[3] P. Del Moral and J. Garnier. Genealogical particle analysis of rare events. *Ann. Appl. Probab.*, 15(4):2496–2534, 2005.

[4] B. Djehiche, H. Hult, and P. Nyquist. Min-max representation of viscosity solutions of Hamilton-Jacobi equations and applications in rare-event simulation. Royal Institute of Technology, 2014.

[5] P. Dupuis and R. S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley, New York, 1997.

[6] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley Series in Probability and Mathematical Statistics, 3 edition, 1950.

[7] T. Gudmundsson and H. Hult. Markov chain monte carlo for computing rare-event probabilities for a heavy-tailed random walk. *Journal of Applied Probability*, 51(2), June 2014.