Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

IMPORTANCE SAMPLING STRATEGY FOR HEAVY-TAILED SYSTEMS WITH CATASTROPHE PRINCIPLE

Xingyu Wang
Chang-Han Rhee

Department of Industrial Engineering
and Management Sciences
Northwestern University
2145 Sheridan Road
Evanston, IL 60208, USA

ABSTRACT

Large deviations theory has a long history of providing powerful machinery for designing efficient rare-event simulation techniques. However, traditional large deviations theory fails to provide useful bounds in heavy-tailed contexts, and designing efficient rare-event simulation algorithms for heavy-tailed systems has been considered challenging. Recent developments in the theory of heavy-tailed large deviations enable designing a strongly efficient importance sampling scheme that is universally applicable to a wide range of rare events. This tutorial aims to provide an accessible overview of the recent developments in the large deviations theory for heavy-tailed stochastic processes, which is followed by a detailed account of the design principle behind the strongly efficient importance sampling scheme for such processes. The implementations of the general principle are demonstrated through a few specific heavy-tailed rare events that arise in stochastic approximation, finance, and queueing theory contexts.

1 INTRODUCTION

Heavy-tailed phenomena are prevalent in a broad class of stochastic dynamics, ranging from the spread of pandemics (Cohen et al. 2022) and the fluctuations in actuarial and financial assets (Embrechts et al. 2013) to the training of machine learning models (Gurbuzbalaban et al. 2021). Precisely evaluating the risks associated with rare events is crucial in many critical applications. This task typically involves estimating probabilities of the form $p = \mathbf{P}(X \in A)$, where X is a stochastic process with heavy-tailed components, and A is a set of unusual scenarios so that p is close to 0. The crude Monte Carlo estimator $\mathbb{I}\{X \in A\}$ provides a straightforward means to estimate the probability p, but its standard error is of order $\sqrt{p}$, and hence, the number of samples required to attain a given level of relative accuracy is of order $\sqrt{1/p}$. For small p's, this can be prohibitively expensive.

When the underlying uncertainties are light-tailed, the importance sampling strategy has been one of the major success stories in rare-event simulation literature (Bucklew et al. 1990; Boxma et al. 2019; Torrisi 2004; Dupuis et al. 2007). Importance sampling involves generating samples of X from an alternative probability measure $\mathbf{Q}$, i.e., *importance distribution*, instead of the nominal distribution $\mathbf{P}$. Of course, $\mathbb{I}\{X \in A\}$ is a biased estimator of p under $\mathbf{Q}$. To adjust the bias, one calibrates the importance sampling estimator with the likelihood ratio $d\mathbf{P}/d\mathbf{Q}$ between the nominal distribution $\mathbf{P}$ and the importance sampling distribution $\mathbf{Q}$. The resulting importance sampling estimator $\mathbb{I}\{X \in A\}\frac{d\mathbf{P}}{d\mathbf{Q}}$ is valid (i.e., unbiased) in great

generality regardless of the choice of the importance distribution $\mathbf{Q}$:

$$\mathbf{E}_{\mathbf{Q}}\left[\mathbb{1}\{X \in A\}\frac{d\mathbf{P}}{d\mathbf{Q}}\right] = \int \mathbb{1}\{X \in A\}\frac{d\mathbf{P}}{d\mathbf{Q}}d\mathbf{Q} = \int \mathbb{1}\{X \in A\}d\mathbf{P} = \mathbf{P}(X \in A).$$

However, for the importance sampling scheme to be useful, an appropriate choice of the importance distribution $\mathbf{Q}$ is crucial. In principle, $\mathbf{Q}(\cdot) \triangleq \mathbf{P}(\,\cdot\,|X \in A)$ is the optimal choice in the sense that such $\mathbf{Q}$ minimizes the variance of the importance sampling estimator $\mathbb{I}\{X \in A\}\frac{d\mathbf{P}}{d\mathbf{Q}}$. However, this theoretically optimal strategy is not implementable because it requires computation of the exact value of $d\mathbf{P}/d\mathbf{Q} = \mathbf{P}(X \in A)$, which is the target quantity of our original task. Nonetheless, the ideal importance sampling distribution provides a guideline for designing the importance distribution. That is, one wants to pick $\mathbf{Q}$ in such a way that $d\mathbf{P}/d\mathbf{Q}$ is computable, and $\mathbf{Q}(\cdot) \approx \mathbf{P}(\,\cdot\,|X \in A)$ in some sense. On the other hand, it is well known that, without principled approach, seemingly plausible choices of $\mathbf{Q}$ can not only fail to reduce the estimator's variance but also result in infinite variance; see, for instance, Glasserman and Wang (1997) and Glasserman and Kou (1995). Moreover, ill-designed importance samplers can appear to be deceptively robust giving false confidence to wrong answers. In view of these, principled approaches with theoretical guarantee are required in designing importance sampling algorithms. For light-tailed dynamical systems, general principles for constructing provably efficient importance samplers have been established based on large deviations bounds (Dupuis and Wang 2004; Dupuis and Wang 2005; Dupuis and Wang 2009). However, designing provably efficient rare-event simulation algorithms for heavy-tailed rare events has been much more obscure (see, for example, Bassamboo et al. 2007) due to the fundamentally different mechanism through which the system-wide rare events arise and the lack of the heavy-tailed large deviations theory at the sample-path level. Although some importance sampling (e.g., Blanchet and Glynn 2008; Dupuis et al. 2007; Blanchet et al. 2008; Blanchet and Liu 2008; Murthy et al. 2014; Blanchet et al. 2013) and other variance reduction techniques such as conditional Monte Carlo (e.g., Asmussen and Kroese 2006; Hult et al. 2016) and Markov Chain Monte Carlo (e.g., Gudmundsson and Hult 2014) have been designed successfully to address heavy-tailed problems, these works are typically tailored for specific processes and specific rare events, or the generalization of their approaches (such as the Lyapunov inequality technique in Blanchet and Glynn 2008) becomes highly non-trivial beyond relatively simple settings.

Recent developments of heavy-tailed large deviations such as Rhee et al. (2019) and Wang and Rhee (2023) offer critical insights into designing efficient and universal importance sampling schemes for heavy-tailed systems. At the core of this development is the discrete hierarchy of heavy-tailed rare events that is characterized by *catastrophe principle*. Roughly speaking, catastrophe principle dictates that the system-wide rare events in heavy-tailed systems arise due to catastrophic failures of a small number of system components, and the number of such components governs the asymptotic rate at which the associated rare events occur. This creates a discrete hierarchy in heavy-tailed rare events. (Note also that this implies that the most likely scenarios associated with the heavy-tailed rare events are singular to any exponentially tilted measures which are typically the most likely scenarios in light-tailed contexts, thus explaining why the light-tailed approaches fail to provide efficient heavy-tailed importance sampling estimators.) Combining the defensive importance sampling idea with such hierarchy, strongly efficient samplers can be designed for a variety of rare events associated with random walks, compound Poisson processes (Chen et al. 2019), and Lévy processes with infinite activities (Wang and Rhee 2020). The same principle can be applied to more general stochastic processes such as Lévy driven stochastic differential equations and stochastic difference equations.

The goal of this tutorial is to provide an accessible overview of the heavy-tailed large deviations theory (Section 2) and streamlined account of the general principle and intuition behind the universal importance sampling scheme (Section 3). We then illustrate the implementation of the general principle in option pricing, stochastic approximation, fluid queueing networks, and multiple-server queues (Section 4).

2 LARGE DEVIATIONS THEORY FOR HEAVY-TAILED SYSTEMS

This section reviews the sample-path large deviations for heavy-tailed stochastic processes, which is central to the design and analysis of the importance sampling algorithms we will discuss in Section 3. In particular, the theory establishes the *catastrophe principle* that characterizes the most likely cause and the probabilities of the rare events in a variety of heavy-tailed systems.

2.1 Random Walks

We start with the simplest set up and focus on the most fundamental aspect of the catastrophe principle. To facilitate the presentation, we introduce the following notations. Let $\mathbb{Z}$ be the set containing all integers and $\mathbb{Z}_+ = \{0,1,2,\cdots\}$ be the set of non-negative integers. For positive integer k, let $[k] = \{1,2,\cdots,k\}$. For $x \in \mathbb{R}$, let $\lfloor x \rfloor \triangleq \max\{n \in \mathbb{Z} : \ n \le x\}$ and $\lceil x \rceil \triangleq \min\{n \in \mathbb{Z} : \ n \ge x\}$ denote the floor and ceiling of x. Given any $x, y \in \mathbb{R}$, let $x \wedge y \triangleq \min\{x,y\}$ and $x \vee y \triangleq \max\{x,y\}$. Let $(\mathbb{D}_{[0,1],\mathbb{R}}, \boldsymbol{d})$ be the metric space of $\mathbb{D} = \mathbb{D}_{[0,1],\mathbb{R}}$, the space of all real-valued RCLL functions with domain $[0,1]$. The Skorokhod J_1 metric $\boldsymbol{d}$ is defined as $\boldsymbol{d}(x,y) \triangleq \inf_{\lambda\in\Lambda} \sup_{t\in[0,1]} |\lambda(t) - t| \vee |x(\lambda(t)) - y(t)|$ with Λ being the set of all increasing homeomorphisms from $[0,1]$ to itself. For Borel measurable sets $A, B \subset \mathbb{D}$, we say A and B are *bounded away* from each other if $\boldsymbol{d}(A,B) = \inf_{x\in A, y\in B} \boldsymbol{d}(x,y) > 0$. For all $l \ge 0$, let $\mathbb{D}_l$ be the subset of $\mathbb{D}$ containing all the non-decreasing step functions that have exactly l jumps and vanish at the origin. Note that $\mathbb{D}_0 \triangleq \{\mathbf{0}\}$ where $\mathbf{0}(t) \equiv 0$ is the zero function. Set $\mathbb{D}_{<l} \triangleq \bigcup_{j=0}^{l-1} \mathbb{D}_l$.

Next, we review the concept of regular variation, which is by far the most commonly used tool to model the heavy-tailed distributions. A measurable function $\phi : (0,\infty) \to (0,\infty)$ is *regularly varying* (at $+\infty$) with index β if $\lim_{x\to\infty} \phi(tx)/\phi(x) = t^\beta$ for all $t > 0$, and we write $\phi(x) \in \mathbf{RV}_\beta(x)$. If $\phi(x) \in \mathbf{RV}_0(x)$, we say that $\phi(\cdot)$ is slowly varying. It is well known that for any $\phi(\cdot) \in \mathbf{RV}_\beta$, there is some slowly varying $L(\cdot)$ such that $\phi(x) = x^\beta L(x)$. For the purpose of understanding this tutorial, one can consider $L(\cdot)$ more or less as a constant function. See chapter 2 of Resnick (2007) for a standard treatment of this topic.

Now, consider a centered random walk $S_n = Z_1 + \cdots + Z_n$ in $\mathbb{R}$ whose increments Z_i's are heavy-tailed on the positive side. That is, $\mathbf{E}Z_i = 0$, and $\mathbf{P}(Z_1 \ge x) \in \mathbf{RV}_{-\alpha}(x)$ as $x \to \infty$ for some $\alpha > 1$. We assume that Z_i's have a light tail on the negative side so that $\mathbf{P}(-Z_1 \ge x)$ decays at an exponential (or faster) rate as $x \to \infty$. Let $\bar{S}_n(t) = \frac{1}{n} S_{\lfloor nt \rfloor}$ so that $\bar{S}_n = \{\bar{S}_n(t) : \ t \in [0,1]\}$ is a scaled random walk embedded in $\mathbb{D}$. Note that due to the functional law of large numbers, the scaled path $\bar{S}_n$ will converge to a flat straight line $\mathbf{0}$. As an one-sided adaptation of Theorem 4.1 of Rhee et al. (2019), the following result characterizes the probability that $\bar{S}_n$ deviates from its nomial behavior $\mathbf{0}$.

Theorem 1 There exists a family of measures $\{\mathbf{C}_\alpha^l : \ l \ge 0\}$ with each $\mathbf{C}_\alpha^l$ supported on $\mathbb{D}_l$ such that the following claim holds. Given any measurable $A \subset \mathbb{D}$, let $l^* = l^*(A) \triangleq \min\{l \in \mathbb{Z}_+ : \ \mathbb{D}_l \cap A \neq \emptyset\}$. If $\boldsymbol{d}(A, \mathbb{D}_{<l^*}) > 0$, then

$$\mathbf{C}_\alpha^{l^*}(A^\circ) \le \liminf_{n\to\infty} \frac{\mathbf{P}(\bar{S}_n \in A)}{(n\mathbf{P}(Z_1 \ge n))^{l^*}} \le \limsup_{n\to\infty} \frac{\mathbf{P}(\bar{S}_n \in A)}{(n\mathbf{P}(Z_1 \ge n))^{l^*}} \le \mathbf{C}_\alpha^{l^*}(A^-) < \infty$$

where A°, A^- are the interior and closure of A respectively.

For the specific form of $\mathbf{C}_\alpha^l$, see Theorem 4.1 of Rhee et al. (2019). Note that this is a precise asymptotics as opposed to log-asymptotics as in the traditional (light-tailed) large deviations theory, and l^*, as a function of A, plays the role of the rate function. Due to the precise nature, one can, in fact, prove that the conditional distribution $\mathscr{L}(\bar{S}_n | \bar{S}_n \in A)$ given the rare-event of interest converges to the law of random functions that are piece-wise constant with l^* jumps that are bounded from below; see Corollary 4.1 of Rhee et al. (2019) for more details. This is a crisp characterization of the catastrophe principle. Indeed, the index l^*—the minimum number of jumps that needs to be added to a step function to make it fall into set A—not only determines the rate of decay for $\mathbf{P}(\bar{S}_n \in A)$, but also dictates the way the rare events occur: that is, through exactly l^* of Z_i's that catastrophically deviate from its typical value $0 = \mathbf{E}Z_i$, while

the whole process $\bar{S}_n$ behaves nominally (i.e., resemble $\mathbf{0}$) everywhere else. In particular, for large n, the path of $\bar{S}_n$ closely resembles a step function with exactly l^* upward jumps when conditioned on the event $\{\bar{S}_n \in A\}$.

Theorem 1 takes a more subtle form when there are multiple sources of heavy-tails with different power indices. In such cases, the catastrophe principle minimizes the "cost" of the jumps rather than the number of jumps. For example, suppose that Z_i's have different regular variation indices on negative and positive sides. Specifically, suppose that Z_i is still centered (i.e., $\mathbf{E}Z_i = 0$), but $\mathbf{P}(-Z_1 \geq x) \in \mathbf{RV}_{-\alpha}(x)$ and $\mathbf{P}(Z_1 \geq x) \in \mathbf{RV}_{-\beta}(x)$ as $x \to \infty$ for some $\alpha, \beta > 1$. Analogous to $\mathbb{D}_l$ and $\mathbb{D}_{<l}$ defined previously, we introduce a few notions to describe piece-wise step functions in $\mathbb{D}$ that make both upward and downward jumps. For $i, j \in \mathbb{Z}_+$, let $\mathbb{D}_{i,j}$ be the subset of $\mathbb{D}$ containing all step functions that vanish at the origin and have exactly i downward jumps and j upwards jumps. Let $\mathbb{D}_{<l_-,l_+} = \bigcup_{(i,j)\in\mathbb{I}_{<l_-,l_+}} \mathbb{D}_{i,j}$ where $\mathbb{I}_{<l_-,l_+} = \{(i,j) \in \mathbb{Z}_+^2 \backslash (l_-,l_+) : \ (\alpha-1)\cdot i + (\beta-1)\cdot j \leq (\alpha-1)\cdot l_- + (\beta-1)\cdot l_+\}$, which is the set of indices associated with the combination of jumps with equal or less cost compared to (l_-,l_+).

Below, we present Theorem 4.1 of Rhee et al. (2019). Here, the key is to determine the combination of catastrophes that trigger the target event with the minimum cost. Specifically, by solving for (l_-^*, l_+^*), i.e., the minimizer of the cost function $(\alpha-1)l_- + (\beta-1)l_+$ over all (l_-,l_+) with $A \cap \mathbb{D}_{l_-,l_+} \neq \emptyset$, we gain important insights into the rare events $\{\bar{S}_n \in A\}$: first, $\mathbf{P}(\bar{S}_n \in A)$ is roughly of order $n^{-(\alpha-1)l_-^* - (\beta-1)l_+^*}$ for large n; next, the most likely cause of $\{\bar{X}_n \in A\}$ is dictated by (l_-^*, l_+^*), i.e., through exactly l_-^* large negative jumps and l_+^* positive jumps while $\bar{X}_n$ resembles $\mathbf{0}$ everywhere else. As in Theorem 1, $\mathbf{C}_{\alpha,\beta}^{i,j}$ are explicitly identified in Theorem 4.1 of Rhee et al. (2019).

Theorem 2 Suppose that a measurable set $A \subseteq \mathbb{D}$ is bounded away from $\mathbb{D}_{<l_-^*,l_+^*}$ where $(l_-^*, l_+^*) \triangleq \underset{(l_-,l_+)\in\mathbb{Z}_+^2:\ A\cap\mathbb{D}_{l_-,l_+}\neq\emptyset}{\arg\min} (\alpha-1)l_- + (\beta-1)l_+$, then

$$\begin{aligned}\mathbf{C}_{\alpha,\beta}^{l_-^*,l_+^*}(A^\circ) &\leq \liminf_{n\to\infty} \frac{\mathbf{P}(\bar{S}_n \in A)}{\big(n\mathbf{P}(-Z_1 \geq n)\big)^{l_-^*} \cdot \big(n\mathbf{P}(Z_1 \geq n)\big)^{l_+^*}} \\ &\leq \limsup_{n\to\infty} \frac{\mathbf{P}(\bar{S}_n \in A)}{\big(n\mathbf{P}(-Z_1 \geq n)\big)^{l_-^*} \cdot \big(n\mathbf{P}(Z_1 \geq n)\big)^{l_+^*}} \leq \mathbf{C}_{\alpha,\beta}^{l_-^*,l_+^*}(A^-) < \infty,\end{aligned}$$

where, for each $i, j \in \mathbb{Z}_+$, the measure $\mathbf{C}_{\alpha,\beta}^{i,j}$ is supported on $\mathbb{D}_{i,j}$.

2.2 Stochastic Difference and Differential Equations

Next, we discuss the sample-path large deviations for stochastic difference and differential equations under heavy-tailed perturbations (Wang and Rhee 2023). Consider an iid sequence Z_i satisfying the following assumptions: $\mathbf{P}(|Z_1| > x) \in \mathbf{RV}_{-\alpha}(x)$ as $x \to \infty$ for some $\alpha > 1$; there exist $p^{(+)}, p^{(-)} \in (0,1)$ with $p^{(+)} + p^{(-)} = 1$ such that

$$\lim_{x\to\infty} \mathbf{P}(Z_1 > x)/\mathbf{P}(|Z_1| > x) = p^{(+)}, \qquad \lim_{x\to\infty} \mathbf{P}(-Z_1 > x)/\mathbf{P}(|Z_1| > x) = p^{(-)}. \tag{1}$$

For any $c > 0$, let $\varphi_c(x) = (x \wedge c) \vee (-c)$ be the projection operator from $\mathbb{R}$ to $[-c,c]$. Let $\mathscr{C}^1(\mathbb{R})$ be the set of mappings from $\mathbb{R}$ to $\mathbb{R}$ that have continuous derivatives. Given $a \in \mathscr{C}^1(\mathbb{R})$ and $\sigma \in \mathscr{C}^1(\mathbb{R})$, let $\big(Y^{n|b}(j)\big)_{j\geq 0}$ solves

$$Y^{n|b}(0) = 0, \qquad Y^{n|b}(j) = Y^{n|b}(j-1) + \varphi_b\Big(\frac{1}{n}a\big(Y^{n|b}(j-1)\big) + \frac{1}{n}\sigma\big(Y^{n|b}(j-1)\big)\cdot Z_j\Big)\ \forall j \geq 1. \tag{2}$$

Here, $\big(Y^{n|b}(j)\big)_{j\geq 0}$ can be considered as the truncated counterpart of the stochastic difference equation

$$Y^n(0) = 0, \qquad Y^n(j) = Y^n(j-1) + \frac{1}{n}a\big(Y^n(j-1)\big) + \frac{1}{n}\sigma\big(Y^n(j-1)\big)\cdot Z_j\ \forall j \geq 1 \tag{3}$$

as the distance traveled at each step in $Y^{n|b}(j)$ is truncated under the threshold $b > 0$, and it makes sense to denote $Y^{n|\infty}(j) = Y^n(j)$. Next, for any $A \subseteq R$ and positive integer k, let $A^{k\uparrow} = \{(t_1, \cdots, t_k) \in A^k : \ t_1 < t_2 < \cdots < t_k\}$ be the set of strictly increasing sequence of length k over A. For any $b \in (0, \infty)$ and any positive integer k, define the mapping $h^{(k)|b} : \mathbb{R}^k \times (0,1]^{k\uparrow} \to \mathbb{D}$ as follows. Given $\boldsymbol{w} = (w_1, \cdots, w_k) \in \mathbb{R}^k$ and $\boldsymbol{t} = (t_1, \cdots, t_k) \in (0,1]^{k\uparrow}$, let $\xi = h^{(k)|b}(\boldsymbol{w}, \boldsymbol{t})$ solves

$$d\xi(t)/dt = a\big(\xi(t-)\big) \ \forall t \in [0,1], \ t \notin \{t_1, \cdots, t_k\}, \qquad \xi(t_j) = \xi(t_j-) + \varphi_b\Big(\sigma\big(\xi(t_j-)\big) \cdot w_j\Big) \ \forall j \in [k] \quad (4)$$

under initial value $\xi(0) = 0$. Here, $h^{(k)|b}(\boldsymbol{w}, \boldsymbol{t})$ produces the ODE path with perturbations $w_1, \cdots, w_k$ (with size modulated by the drift coefficient $\sigma(\cdot)$ and truncated under $b > 0$) at times $t_1, \cdots, t_k$, respectively. Let $\mathbb{D}_h^{(k)|b} = h^{(k)|b}\big(\mathbb{R}^k \times (0,1]^{k\uparrow}\big)$ be the set containing all such ODE paths with k (modulated and truncated) perturbations. More generally, we use φ_∞ to denote the identity mapping on $\mathbb{R}$, and we let $h^{(k)} = h^{(k)|\infty}$ and $\mathbb{D}_h^{(k)} = \mathbb{D}_h^{(k)|\infty}$ be the untruncated counterparts of $h^{(k)|b}$ and $\mathbb{D}_h^{(k)|b}$.

Given any measurable $B \subseteq \mathbb{D}$, one can see that $J_b^*(A) \triangleq \min\{k \geq 0 : \ \mathbb{D}_h^{(k)|b} \cap A \neq \emptyset\}$ gives the minimum number of perturbations needed for the ODE path to fall into set A. Let $\bar{Y}^{n|b}(t) = Y_n\big(\lfloor nt \rfloor\big)$ and $\bar{Y}^{n|b} = \{\bar{Y}^{n|b}(t) : \ t \in [0,1]\}$ be the time-scaled version of $Y^{n|b}(j)$. As illustrated in Theorem 3, $J_b^*(A)$ dictates the rate of decay for events $\{\bar{Y}_n \in A\}$.

Theorem 3 Suppose that $\sup_{x\in\mathbb{R}} |a(x)| \vee \sigma(x) < \infty$ and $\inf_{x\in\mathbb{R}} \sigma(x) > 0$. Given measurable $A \subseteq \mathbb{D}$ and $b \in (0, \infty]$, if A is bounded away from $\mathbb{D}_h^{(J_b^*(A)-1)|b}$, then

$$\mathbf{C}_h^{(J_b^*(A))|b}(A^\circ) \leq \liminf_{n\to\infty} \frac{\mathbf{P}(\bar{Y}^{n|b} \in A)}{\big(n\mathbf{P}(|Z_1| > n)\big)^{J_b^*(A)}} \leq \limsup_{n\to\infty} \frac{\mathbf{P}(\bar{Y}^{n|b} \in A)}{\big(n\mathbf{P}(|Z_1| > n)\big)^{J_b^*(A)}} \leq \mathbf{C}_h^{(J_b^*(A))|b}(A^-) < \infty$$

where, for each $k \in \mathbb{Z}_+$, $\mathbf{C}_h^{(k)|b}$ is a measure supported on $\mathbb{D}_h^{(k)|b}$.

Lastly, we mention that sample-path large deviation results of the same form can be developed for stochastic differential equations driven by heavy-tailed Lévy processes and the truncated counterparts. To avoid repetitions we omit the details and refer the interested readers to Wang and Rhee (2023).

2.3 Lévy Processes

Lévy processes can be viewed as the continuous-time analog of random walks. Any Lévy process $X(t)$ can decomposed into the sum of Brownian motion and the limit of a sequence of compound Poisson processes with drift; see, e.g., Sato et al. (1999) for details. In particular, each Lévy process $X(t)$ is associated with a Lévy measure ν that indicates the intensity of jumps. Given any open set O, any jump with sizes $\Delta X(t) \in O$ will arrive according to a Poisson process with rate $\nu(O)$, where the size of each jump is iid with law $\nu(\ \cdot \cap O)/\nu(O)$. Therefore, the heavy-tailedness in the increments of $X(t)$ is captured by the tail behavior of its Lévy measure ν.

We first consider the one-dimensional case where a Lévy process $X(t)$ is centered (i.e., $\mathbf{E}X(t) = 0$ for all $t \geq 0$) with Lévy measure ν supported on $(0, \infty)$. In other words, any jump in $X(t)$ will be positive. Suppose that the function $\nu[x, \infty) \in \mathbf{RV}_\alpha(x)$ for some $\alpha > 1$, which captures the heavy-tailedness in the increments of $X(t)$. Let $\bar{X}_n(t) = \frac{1}{n}X(\lfloor nt \rfloor)$ and $\bar{X}_n = \{\bar{X}_n(t) : \ t \in [0,1]\}$. Below, we present Theorem 3.1 of Rhee et al. (2019). Analogous to Theorem 1, the result embodies the catastrophe principle and shows that the key step in characterizing the sharp asymptotics of $\mathbf{P}(\bar{X}_n \in A)$ is to determine l^*, the minimum number of catastrophes required for $\{\bar{X}_n \in A\}$ to occur.

Theorem 4 There exists a family of measures $\{\mathbf{C}_\alpha^l : \ l \geq 0\}$ with each $\mathbf{C}_\alpha^l$ supported on $\mathbb{D}_l$ such that the following claim holds. Given any measurable $A \subset \mathbb{D}$ such that $\boldsymbol{d}(A, \mathbb{D}_{<l^*}) > 0$ where $l^* \triangleq \min\{l \in \mathbb{Z}_+ :$

$\mathbb{D}_l \cap A \neq \emptyset\}$,

$$\mathbf{C}_\alpha^{l^*}(A^\circ) \leq \liminf_{n\to\infty} \frac{\mathbf{P}(\bar{X}_n \in A)}{(n\nu[n,\infty))^{l^*}} \leq \limsup_{n\to\infty} \frac{\mathbf{P}(\bar{X}_n \in A)}{(n\nu[n,\infty))^{l^*}} \leq \mathbf{C}_\alpha^{l^*}(A^-) < \infty.$$

To conclude, we consider $X^{(1)}, \cdots, X^{(d)}$ that are independent and centered Lévy processes in $\mathbb{R}$ with spectrally positive (i.e., restricted on $(0,\infty)$) Lévy measures $\nu^{(1)}, \cdots, \nu^{(d)}$, respectively. For each $i \in [d]$, suppose that $\nu^{(i)}[x,\infty) \in \mathbf{RV}_{-\beta_i}(x)$ fopr some $\beta_i > 1$. Here, we provide an intuitive interpretation of the catastrophe principle in the current setup. The "cost" of observing a large jump (i.e., catastrophe) along the i-th dimension is $\beta_i - 1$, and the most likely cause of a rare event is the one with l_i large jumps along dimension i that minimizes the cost function

$$\mathscr{J}(l_1, \cdots, l_d) \triangleq \sum_{i=1}^{d} l_i \cdot (\beta_i - 1). \tag{5}$$

Let $\mathbb{D}^k$ be the k-fold product space of $\mathbb{D}$. Given any $(l_1^*, \cdots, l_d^*) \in \mathbb{Z}_+^d$, let $\mathbb{D}_{<(l_1^*,\cdots,l_d^*)} \triangleq \bigcup_{(l_1,\cdots,l_d)\in\mathbb{I}_{<(l_1^*,\cdots,l_d^*)}} \prod_{j=1}^d \mathbb{D}_{l_j}$ where $\mathbb{I}_{<(l_1^*,\cdots,l_d^*)} \triangleq \big\{(l_1,\cdots,l_d) \in \mathbb{Z}_+^d \setminus (l_1^*,\cdots,l_d^*) : \ \mathscr{J}(l_1,\cdots,l_d) \leq \mathscr{J}(l_1^*,\cdots,l_d^*)\big\}$. Also, set the maximum metric on the product space $\mathbb{D}^d$ as $\boldsymbol{d}_d\big((x_1,\cdots,x_d),(y_1,\cdots,y_d)\big) = \max_{i\in[d]} \boldsymbol{d}(x_i,y_i)$, and we say that $A \subseteq \mathbb{D}^d$ is bounded away from $B \subseteq \mathbb{D}^d$ if $\boldsymbol{d}_d(A,B) > 0$. We present Theorem 3.6 of Rhee et al. (2017) that embodies the catastrophe principle in the multi-dimensional setting. Let $\bar{X}_n(t) = \big(X^{(1)}(\lfloor nt \rfloor)/n, \cdots, X^{(d)}(\lfloor nt \rfloor)/n\big)$ and $\bar{X}_n = \{\bar{X}_n(t) : \ t \in [0,1]\}$.

Theorem 5 There exists a family of measures $\{\mathbf{C}_{\beta_1,\cdots,\beta_d}^{(l_1,\cdots,l_d)} : \ i_k \geq 0 \ \forall k \in [d]\}$ where each $\mathbf{C}_{\beta_1,\cdots,\beta_d}^{(l_1,\cdots,l_d)}$ is supported on $\prod_{k=1}^d \mathbb{D}_{l_k}$ such that the following claim holds. Suppose that a measurable set $A \subseteq \mathbb{D}^d$ is bounded away from $\mathbb{D}_{<(l_1^*,\cdots,l_d^*)}$ where

$$(l_1^*, \cdots, l_d^*) = \operatorname*{arg\,min}_{(l_1,\cdots,l_d)\in\mathbb{Z}_+^d : \ A\cap\prod_{j=1}^d \mathbb{D}_{l_j} \neq \emptyset} \mathscr{J}(l_1, \cdots, l_d),$$

then (let A°, A^- be the interior and closure of A, respectively)

$$\mathbf{C}_{\beta_1,\cdots,\beta_d}^{(l_1^*,\cdots,l_d^*)}(A^\circ) \leq \liminf_{n\to\infty} \frac{\mathbf{P}(\bar{X}_n \in A)}{\prod_{i=1}^d \big(n\nu^{(i)}[n,\infty)\big)^{l_i^*}} \leq \limsup_{n\to\infty} \frac{\mathbf{P}(\bar{X}_n \in A)}{\prod_{i=1}^d \big(n\nu^{(i)}[n,\infty)\big)^{l_i^*}} \leq \mathbf{C}_{\beta_1,\cdots,\beta_d}^{(l_1^*,\cdots,l_d^*)}(A^-) < \infty.$$

3 ALGORITHM

This section discusses the general principle for designing importance sampling algorithms for heavy-tailed systems based on the sample-path large deviations reviewed in Section 2. The importance sampling algorithm is both readily implementable and universally applicable to a broad class of rare events in heavy-tailed systems. Moreover, this algorithm attains strong efficiency in the following sense. Given two sequences of non-negative real numbers x_n, y_n, we say $x_n = O(y_n)$ if $\limsup_{n\to\infty} x_n/y_n < \infty$, and we say $x_n = o(y_n)$ of $\lim_{n\to\infty} x_n/y_n = 0$. For sequences of events $(A_n)_{n\geq 1}$ and random variables $(L_n)_{n\geq 1}$, we say that the estimators $(L_n)_{n\geq 1}$ are *unbiased and strongly efficient* for $(\mathbf{P}(A_n))_{n\geq 1}$ if

$$\mathbf{E}L_n = \mathbf{P}(A_n) \ \forall n \geq 1; \qquad \mathbf{E}L_n^2 = O\big(\mathbf{P}(A_n)^2\big) \ \text{ as } n \to \infty. \tag{6}$$

It is worth emphasizing that the strongly efficient estimators $(L_n)_{n\geq 1}$ achieve uniformly bounded relative errors for all $n \geq 1$, meaning that the number of samples required to achieve a given level of relative accuracy is uniformly bounded, regardless of how small the target probability $\mathbf{P}(A_n)$ is.

3.1 Importance Sampling Distribution $\mathbf{Q}_n$

For simplicity and clarity of the presentation, we start with the simplest case. Recall $\bar{S}_n(t) = S_{\lfloor nt \rfloor}/n$ defined in Section 2.1. In particular, we assumed that $\mathbf{P}(Z_1 \geq x) \in \mathbf{RV}_{-\alpha}(x)$ as $x \to \infty$ for some $\alpha > 1$, whereas the left tail decayed at least at an exponential rate. Let $A_n \triangleq \{\bar{S}_n \in A\}$ for some $A \subset \mathbb{D}$. This section explains a strongly efficient rare event simulation algorithm for $\mathbf{P}(A_n)$. In Section 3.3, we will discuss how this principle can be generalized to the rare events associated with different classes of heavy-tailed systems.

Recall $l^*(A) = \min\{l \in \mathbb{Z}_+ : \mathbb{D}_l \cap A \neq \emptyset\}$ in Theorem 1. Note that Theorem 1 dictates that $\mathbf{P}(\bar{S}_n \in A)$ is of order $(n\mathbf{P}(Z_1 \geq n))^{l^*(A)}$. Furthermore, any other rare events that require $l^*(A)$ jumps have the same asymptotic rate as A_n. In view of this, a natural approach is to consider an importance distribution of the form $\mathbf{P}(\cdot|\bar{S}_n \in B)$ for some B, which is tractable and satisfies $l^*(B) = l^*(A)$. It turns out that the following defensive importance sampling mixture strikes the right balance:

$$\mathbf{Q}_n(\cdot) \triangleq w\mathbf{P}(\cdot) + (1-w)\mathbf{P}(\ \cdot\ |B_n^\gamma) \tag{7}$$

with some prefixed constant $w \in (0,1)$ and $\gamma \in (0,\infty)$. Here, we set $B_n^\gamma = \{\bar{S}_n \in B^\gamma\}$ with

$$B^\gamma = \Big\{\xi \in \mathbb{D} :\ \#\big\{t \in [0,1] :\ \xi(t) - \xi(t-) \geq \gamma\big\} \geq l^*(A)\Big\}. \tag{8}$$

The choice of γ is crucial and will be discussed in the next subsection. Note that by its construction, we have $l^*(B^\gamma) = l^*(A)$ as desired. If we define

$$L_n \triangleq \mathbb{1}_{A_n} \cdot d\mathbf{P}/d\mathbf{Q}_n \tag{9}$$

where $d\mathbf{P}/d\mathbf{Q}_n$ is the likelihood ratio of $\mathbf{P}$ and $\mathbf{Q}_n$, then L_n is obviously an unbiased estimator for $\mathbf{P}(\bar{S}_n \in A)$ under $\mathbf{Q}_n(\cdot)$. In the next subsection, we will see that L_n is strongly efficient. Here, we just mention that the first term in (7) prevents $d\mathbf{P}/d\mathbf{Q}_n$ from blowing up, whereas the second term makes sure that $\mathbf{Q}_n$ resembles the ideal (but unimplementable) zero-variance importance distribution $\mathbf{P}(\cdot|\bar{S}_n \in A)$.

For L_n to be implementable, one should be able to sample from $\mathbf{P}(\ \cdot\ |B_n^\gamma)$ efficiently. To do so, first, let $\mathrm{Binom}(n,p)$ be the count of success trials among n Bernoulli trials with success rate p, and let $p_{n,\gamma} = \mathbf{P}(Z_1 \geq \gamma/n)$. Let k be sampled from the law of $\mathrm{Binom}(n,p_n)$, conditioning on $\mathrm{Binom}(n,p_n) \geq l^*(A)$. Then we uniformly randomly pick indices $1 \leq i_1 < i_2 < \cdots, i_k \leq n$ among $\{1,\cdots,n\}$. For each $i \notin \{i_1,\cdots,i_k\}$, we sample Z_i from $\mathbf{P}(\ \cdot\ |Z_i \leq \gamma/n)$, which can be done via straighforward acceptance-rejection. For each $i \in \{i_1,\cdots,i_k\}$, we instead sample Z_i from $\mathbf{P}(\ \cdot\ |Z_i \geq \gamma/n)$. This can be done through the inverse of $\mathbf{P}(Z_i \geq x)$, i.e., $Q_n^{\leftarrow}(y) \triangleq \inf\{s > 0 : \mathbf{P}(Z_1 \geq s) < y\}$. Specifically, one can sample $\Gamma_1,\cdots,\Gamma_k \overset{\text{iid}}{\sim} \mathrm{Unif}\big(0,\mathbf{P}(Z_1 \geq n\gamma)\big)$ and set $Z_{i_j} = Q_n^{\leftarrow}(\Gamma_j)$ for each $j \in [k]$. See Wang and Rhee (2020) for details.

3.2 Strong Efficiency of L_n and the Choice of γ

This section examines the strong efficiency of L_n and the choice of γ on the performance of the estimator L_n. For a given probability measure μ, let $\mathbf{E}^\mu$ denote the expectation operator under μ. Note that, from the definition of $\mathbf{Q}_n$ in (7), $d\mathbf{P}/d\mathbf{Q}_n \leq 1/w$ on set $(B_n^\gamma)^c$ and $d\mathbf{P}/d\mathbf{Q}_n \leq \mathbf{P}(B_n^\gamma)/(1-w)$ on set B_n^γ. Therefore,

$$\begin{aligned}\mathbf{E}^{\mathbf{Q}_n}[L_n^2] &= \mathbf{E}^{\mathbf{Q}_n}\Big[\mathbb{1}_{A_n} \cdot \frac{d\mathbf{P}}{d\mathbf{Q}_n} \cdot \frac{d\mathbf{P}}{d\mathbf{Q}_n}\Big] = \mathbf{E}\Big[\mathbb{1}_{A_n} \cdot \frac{d\mathbf{P}}{d\mathbf{Q}_n}\Big] = \mathbf{E}\Big[\mathbb{1}_{A_n \cap B_n^\gamma} \cdot \frac{d\mathbf{P}}{d\mathbf{Q}_n}\Big] + \mathbf{E}\Big[\mathbb{1}_{A_n \setminus B_n^\gamma} \cdot \frac{d\mathbf{P}}{d\mathbf{Q}_n}\Big] \\ &\leq \frac{\mathbf{P}(B_n^\gamma)}{1-w} \cdot \mathbf{P}(A_n) + \frac{1}{w} \cdot \mathbf{P}(A_n \setminus B_n^\gamma).\end{aligned} \tag{10}$$

Note that since $l^*(B^\gamma) = l^*(A)$ by design, we have $\mathbf{P}(B_n^\gamma) = O(\mathbf{P}(A_n))$ from Theorem 1, and hence, the first term of (10) is $O\big(\mathbf{P}(A_n)^2\big)$. For the second term, it turns out that we can pick $\gamma > 0$ small enough so that

$(A \setminus B^\gamma) \cap \mathbb{D}_l = \emptyset$ for all $l \le 2l^*(A)$, and hence, from Theorem 1 again, we obtain $l^*(A \setminus B^\gamma) > 2 \cdot l^*(A)$. Therefore,

$$\mathbf{P}(A_n \backslash B_n^\gamma) = O\Big(\big(n\mathbf{P}(Z_1 \ge n)\big)^{l^*(A\setminus B^\gamma)}\Big) = o\Big(\big(n\mathbf{P}(Z_1 \ge n)\big)^{2l^*(A)}\Big) = o\big(\mathbf{P}(A_n)^2\big) \qquad \text{as } n \to \infty \tag{11}$$

under the mild condition that $A \backslash B^\gamma$ is bounded away from $\mathbb{D}_{<l^*(A\setminus B^\gamma)}$. We conclude that the second moment of L_n^2 is $O\big(\mathbf{P}(A_n)^2\big)$, establishing the strong efficiency of L_n. In summary, we arrive at the following unified framework to determine the appropriate importance sampling algorithm that ensures strong efficiency.

1. Solve for $l^* = \min\{l \in \mathbb{Z}_+ :\ \mathbb{D}_l \cap A \ne \emptyset\}$
2. Find $\gamma > 0$ such that $A \cap (\mathbb{D}_l \backslash B^\gamma) = \emptyset$ holds for all $l \le 2l^*$.

3.3 Extension to Other Heavy-tailed Processes

Next, we discuss how to apply the importance sampling strategy in Section 3.1 and 3.2 to the processes beyond the one-sided random walk. By repeating the derivations in (10) and (11) using the sample-path large deviation results stated in Section 2, a framework analogous to the one developed in Section 3.2 can be obtained for other heavy-tailed stochastic processes. Here, we omit the technical details and focus on highlighting the differences therein.

We start from random walks in $\mathbb{R}^1$ with heavy-tailed increments on both sides. Let $S_n = Z_1 + \cdots + Z_n$ be a centered random walk where $\mathbf{P}(-Z_1 \ge x) \in \mathbf{RV}_{-\alpha}(x)$ and $\mathbf{P}(Z_1 \ge x) \in \mathbf{RV}_\beta(x)$ as $x \to \infty$ for some $\alpha, \beta > 1$. To efficiently estimate $\mathbf{P}(A_n)$ with $A_n = \{\bar{S}_n \in A\}$, we adopt the design of the importance sampling estimator $L_n = \mathbb{1}_{A_n} \cdot d\mathbf{P}/d\mathbf{Q}_n$ but with a slightly different choice of $B_n^\gamma = \{\bar{S}_n \in B^\gamma\}$. Specifically, for technical reasons we set $B^\gamma \triangleq \bigcup_{(i,j)\in\partial\mathbb{I}(l_-^*,l_+^*)} B^{\gamma,i,j}$ where $B^{\gamma,i,j} = \{\xi \in \mathbb{D} :\ \#\{t \in [0,1] :\ \xi(t-) - \xi(t) \ge \gamma\} \ge i,\ \#\{t \in [0,1] :\ \xi(t) - \xi(t-) \ge \gamma\} \ge j\}$ and (recall the definition of $\mathbb{I}_{<(l_1,\cdots,l_d)}$ in Section 2.3)

$$\begin{aligned}&\partial\mathbb{I}(j_1,\cdots,j_d)\\ &\triangleq \Big\{(l_1,\cdots,l_d) \in \mathbb{Z}_+^d \backslash \mathbb{I}_{<(j_1,\cdots,j_d)} : (m_1,\cdots,m_d) \prec (l_1,\cdots,l_d) \text{ implies } (m_1,\cdots,m_d) \in \mathbb{I}_{<(j_1,\cdots,j_d)}\Big\}.\end{aligned}$$

Here, we define a partial order on $\mathbb{Z}_+^d$ such that $(l_1,\cdots,l_d) \prec (m_1,\cdots,m_d)$ if and only if $l_i \le m_i\ \forall i \in [d]$ and there exists some $j \in [d]$ such that $l_j < m_j$. The set $\partial\mathbb{I}(l_1^*,\cdots,l_d^*)$ can be viewed as the boundary set or dominating set that "envelopes" the set $\mathbb{I}_{<(l_1^*,\cdots,l_d^*)}$. By repeating the derivation in Section 3.2 with Theorem 2, we arrive at following procedure to determine γ:

1. Solve for (l_-^*, l_+^*), the minimizer of $\min\{(\alpha-1)l_- + (\beta-1)l_+ :\ A \cap \mathbb{D}_{l_-,l_+} \ne \emptyset\}$;
2. Find γ such that $A \cap (\mathbb{D}_{l_-,l_+} \backslash B^\gamma) = \emptyset$ holds for all (l_-, l_+) with $(\alpha-1)l_- + (\beta-1)l_+ \le 2(\alpha-1)l_-^* + 2(\beta-1)l_+^*$.

Analogously, we propose a universal framework for rare event simulation in heavy-tailed stochastic difference/differential equations. We focus on the $\mathbb{R}^1$ case for the simplicity of the presentation, but the method can be easily extended to $\mathbb{R}^d$ settings. Suppose that Z_i are iid RVs such that $\mathbf{E}Z_i = 0$, $\mathbf{P}(|Z_1| > x) \in \mathbf{RV}_{-\alpha}(x)$ for some $\alpha > 1$ and the limits in (1) hold. Given $b \in (0,\infty]$, let $Y^{n|b}(j)$ be defined under the recursion (2). In case that $b = \infty$, the recursion coincides with the one defined in (3) for $Y^n(j)$. Let $\bar{Y}^{n|b}(t) = Y_n\big(\lfloor nt \rfloor\big)$ and $\bar{Y}^{n|b} = \{\bar{Y}^{n|b}(t) :\ t \in [0,1]\}$ be the time-scaled version of $Y^{n|b}(j)$. By repeating the analysis in Sections 3.1 and 3.2 with Theorem 3, we obtain a strongly efficient algorithm for the estimation of $\mathbf{P}(\bar{Y}^{n|b} \in A)$. Specifically, by determining $J_b^*(A) \triangleq \min\{k \ge 0 :\ \mathbb{D}_h^{(k)|b} \cap A \ne \emptyset\}$, we define $B^\gamma = \big\{\xi \in \mathbb{D} :\ \#\{t \in [0,1] :\ |\Delta\xi(t)| \ge \gamma\} \ge J_b^*(A)\big\}$ as the set of RCLL paths with at least $J_b^*(A)$ jumps of size larger than γ. Then in (7), we set $B_n^\gamma = \{\bar{Y}^{n|b} \in B^\gamma\}$. In summary, we obtain the following procedure.

1. Solve for $J_b^*(A) \triangleq \min\{k \ge 0 :\ \mathbb{D}_h^{(k)|b} \cap A \ne \emptyset\}$;

2. Find γ such that $A \cap (\mathbb{D}^{(k)|b} \backslash B^\gamma) = \emptyset$ holds for all $k \le 2J_b^*(A)$.

To avoid repetitions, we omit the details and mention that a strongly efficient importance sampling algorithm can be designed analogously for heavy-tailed stochastic differential equations and the truncated counterparts.

Lastly, we describe the importance sampling scheme for (multi-dimensional) Lévy processes with heavy-tailed increments. Recall $\bar{X}_n(t) = \big(X^{(1)}(\lfloor nt \rfloor)/n, \cdots, X^{(d)}(\lfloor nt \rfloor)/n\big)$ and $\bar{X}_n = \{\bar{X}_n(t) : \ t \in [0,1]\}$ defined in Section 2.3. Specifically, for the spectrally positive Lévy measures $\nu^{(1)}, \cdots, \nu^{(d)}$ we have, for each $i \in [d]$, that $\nu^{(i)}[x,\infty) \in \mathbf{RV}_{-\beta_i}(x)$ for some $\beta_i > 1$. Our goal is to estimate $\mathbf{P}(\bar{X}_n \in A)$. In the importance sampling estimator L_n, we set $B_n^\gamma = \{\bar{X}_n \in B^\gamma\}$ with $B^\gamma \triangleq \bigcup_{\boldsymbol{l} \in \partial \mathbb{I}(l_1^*, \cdots, l_d^*)} B^{\gamma;\boldsymbol{l}}$ where $B^{\gamma;(l_1,\cdots,l_d)} \triangleq \Big\{(\xi^{(1)}, \cdots, \xi^{(d)}) \in \mathbb{D}^d : \ \#\{t \in [0,1] : \ \xi^{(i)}(t) - \xi^{(i)}(t-) \ge \gamma\} \ge l_i \ \forall i \in [d]\Big\}$. By repeating the analysis in Sections 3.1 and 3.2 using Theorem 5, we provide the following unified framework to determine the strongly efficient importance sampling scheme. See (5) for the definition of $\mathscr{J}(l_1, \cdots, l_d)$.

1. Solve for $\boldsymbol{l}^* = (l_1^*, \cdots, l_d^*)$, the minimizer of $\min\{\mathscr{J}(l_1, \cdots, l_d) : \ A \cap \prod_i^d \mathbb{D}_{l_i} \neq \emptyset\}$;
2. Find γ such that $A \cap (\prod_i^d \mathbb{D}_{l_i} \backslash B^\gamma) = \emptyset$ holds for all $(l_1, \cdots, l_d)$ with $\mathscr{J}(l_1, \cdots, l_d) \le 2\mathscr{J}(l_1^*, \cdots, l_d^*)$.

To conclude, we note that the sampling of Lévy process $\bar{X}_n$ from $\mathbf{P}(\ \cdot\ | B_n^\gamma)$ has been addressed in both Chen et al. (2019) and Wang and Rhee (2020) and can be understood as a compound Poisson version of the sampling for $\bar{S}_n$ discussed at the end of Section 3.1.

3.4 Dealing with Infinite Activities in Lévy Processes

This section revisits one implicit assumption in the previous sections that the indicator function $\mathbb{1}\{X \in A\}$ for a given $A \subseteq \mathbb{D}$ and some process $X = \{X(t) : \ t \in [0,T]\}$ can be evaluated directly, which may not be the case for all heavy-tailed stochastic processes. In this subsection, we focus on Lévy process with infinite activities. For the rigorous definition of infinite activities and a general approach to rare-event simulation in heavy-tailed Lévy processes with infinite activities, we refer the readers to Wang and Rhee (2020).

Consider the following example. Let $X(t) = B(t) + \sum_{i=1}^{N(t)} (W_i - \mathbf{E}W_i)$ where $B(t)$ is the standard Brownian motion in $\mathbb{R}$, N is a Poisson process with rate $\lambda > 0$, and $(W_i)_{i \ge 1}$ are iid Pareto RVs with law $\mathbf{P}(W_1 > x) = 1/\max\{1,x\}^\alpha$ for some $\alpha > 1$. Let $\bar{X}_n(t) = \frac{1}{n} X(nt)$ and $\bar{X}_n = \{\bar{X}_n(t) : \ t \in [0,1]\}$ be the time-scaled version of $X(t)$. The goal is to estimate $\mathbf{P}(\bar{X}_n \in A)$ where $A = \big\{\xi \in \mathbb{D} : \ \sup_{t\in[0,1]} \xi(t) \ge a, \ \sup_{t\in[0,1]} \xi(t) - \xi(t-) \le b\big\}$ for some $0 < b < a$. Due to the presence of the Brownian motion term, the exact evaluation of $\mathbb{1}\{\bar{X}_n \in A\}$ is computationally challenging under the proposed importance sampling strategy, which prevents us from directly implementing the importance sampling estimator $L_n = \mathbb{1}\{\bar{X}_n \in A\} \cdot d\mathbf{P}/d\mathbf{Q}_n$.

To address this issue, we construct a sequence of approximations for $\mathbb{1}\{\bar{X}_n \in A\}$. First, we introduce a decomposition of the Lévy process $X(t)$. Let $X^{\ge c}(t) = \sum_{s\in[0,t]} \Delta X(t) \mathbb{1}\{\Delta X(t) \ge c\}$, where $\Delta X(t) = X(t) - X(t-)$ denotes the jump at time t. For a chosen value $\gamma \in (0,b)$, we define $J_n(t) = X^{\ge n\gamma}(t)$ and set $X^{<n\gamma}(t) = X(t) - X^{\ge n\gamma}(t)$ as the remaining part of $X(t)$. Now, let $E = \{\xi \in \mathbb{D} : \ \sup_{t\in[0,1]} : \ \xi(t) - \xi(t-) \le b\}$, $A' = \{\xi \in \mathbb{D} : \ \sup_{t\in[0,1]} \xi(t) \ge a\}$ and note that $A = A' \cap E$. Also, define the scaled process $\bar{J}_n = \{\frac{1}{n} J_n(nt) : \ t \in [0,1]\}$ and $\widetilde{X}_n(t) = \{\frac{1}{n} X^{<n\gamma}(nt) : \ t \in [0,1]\}$. We have (due to $\gamma \in (0,b)$)

$$L_n = \mathbb{1}\Big\{\bar{J}_n + \widetilde{X}_n \in A\Big\} \cdot \frac{d\mathbf{P}}{d\mathbf{Q}} = \mathbb{1}\big\{\bar{J}_n + \widetilde{X}_n \in A'\big\} \cdot \mathbb{1}\big\{\bar{J}_n \in E\big\} \cdot \frac{d\mathbf{P}}{d\mathbf{Q}} = \frac{\mathbb{1}\big\{\bar{J}_n + \widetilde{X}_n \in A'\big\} \cdot \mathbb{1}\big\{\bar{J}_n \in E\big\}}{w + \frac{1-w}{\mathbf{P}(B_n^\gamma)} \cdot \mathbb{1}\big\{\bar{J}_n \in B_n^\gamma\big\}}.$$

Due to the infinite activities in $\widetilde{X}_n$, accurately evaluating $\mathbb{1}\big\{\bar{J}_n + \widetilde{X}_n \in A'\big\}$ is computationally challenging. Therefore, we construct an unbiased estimator $\hat{Z}_n(\xi)$ that satisfies $\mathbf{E}\hat{Z}_n(\xi) = \mathbf{P}\big(\widetilde{X}_n + \xi \in A'\big) = \mathbf{P}\big(\sup_{t\in[0,1]} \widetilde{X}_n(t) + \xi(t) \ge a\big)$, and replace $\mathbb{1}\big\{\widetilde{X}_n + \bar{J}_n \in A'\big\}$ with $\hat{Z}_n(\bar{J}_n)$ in our algorithm.

The first key component is the debiasing technique introduced in Rhee and Glynn (2015). Specifically, for $Z_n(\xi) = \mathbb{1}\{\sup_{t\in[0,1]} \widetilde{X}_n(t) + \xi(t) \geq a\}$, suppose that we can construct a sequence of approximators $Z_{n,m}(\xi)$ supported on the same probability space such that the squared error $\mathbf{E}|Z_{n,m}(\xi) - Z_n(\xi)|^2$ quickly approaches 0 as $m \to \infty$. Then the debiasing technique allows us to show that $\hat{Z}_n(\xi) = \sum_{m=1}^{\tau} \big(Z_{n,m}(\xi) - Z_{n,m-1}(\xi)\big)/\mathbf{P}(\tau \geq m)$ satisfies $\mathbf{E}\hat{Z}_n(\xi) = Z_n(\xi)$, where τ is independent of everything else.

To construct such $Z_{n,m}(\xi)$ for some $\xi(t) = \sum_{i=1}^{k} z_i \mathbb{1}_{[u_i,1]}(t)$, we introduce the second key component. First, we partition the timeline $[0,n]$ into $k+1$ disjoint intervals $[0,u_1)$, $[u_1,u_2), \cdots, [u_{k-1},u_k)$, and $[u_k,1]$. We adopt the convention $u_0 = 0, u_{k+1} = 1$ and set $I_i = [u_{i-1},u_i)$ for $i \in [k]$ and $I_{k+1} = [u_k,1]$. Observe that

$$Z_n(\xi) = \mathbb{1}\Big\{\sup_{t\in[0,1]} \widetilde{X}_n(t) + \xi(t) \geq a\Big\} = \mathbb{1}\Big\{\max_{i\in[k+1]} \big[\xi(u_{i-1}) + \widetilde{X}_n(u_{i-1}) + \sup_{t\in I_i} \widetilde{X}_n(t) - \widetilde{X}_n(u_{i-1})\big] \geq a\Big\}.$$

The $Z_{n,m}(\xi)$'s are constructed by approximating $\sup_{t\in I_i} \widetilde{X}_n(t) - \widetilde{X}_n(u_{i-1})$ using the stick-breaking representation in Pitman and Bravo (2012), i.e., an intriguing characterization of the joint law of some Lévy process $X(t)$ with infinite activities and its running supremum $M(t) = \sup_{s\in[0,t]} X(s)$. Here, we fix some $i \in [k+1]$. Let $(U_j^{(i)})_{j\geq 1}$ be iid Unif$(0,1)$. Let $l_0^{(i)} = u_i - u_{i-1}$, and $l_j^{(i)} = U_j^{(i)} \cdot (l_0^{(i)} - l_1^{(i)} - \cdots - l_{j-1}^{(i)})$ for all $j \geq 1$. Conditioning on $(l_j^{(i)})_{j\geq 1}$, we generate $\zeta_j^{(i)}$ as an independent copy of $\widetilde{X}_n(l_j^{(i)})$. Now we set

$$Z_{n,m}(\xi) = \mathbb{1}\Big\{\max_{i\in[k+1]} \big[\xi(u_{i-1}) + \sum_{l=1}^{i-1}\sum_{j\geq 1} \xi_j^{(l)} + \sum_{j=1}^{\lceil \log_2(n^2)\rceil + m} \max\{\zeta_j^{(i)}, 0\}\big] \geq a\Big\}$$

and let $\hat{Z}_n(\xi) = \sum_{m=1}^{\tau} \big(Z_{n,m}(\xi) - Z_{n,m-1}(\xi)\big)/\mathbf{P}(\tau \geq m)$, where $\mathbf{P}(\tau \geq m) = \rho^m$ is the law of a geometric random variable with success rate $\rho \in (0,1)$. Notably, the proposed algorithm achieves unbiasedness and strong efficiency under any $\gamma \in (0,b)$ and any ρ sufficiently close to 1. Besides, thanks to the finite termination threshold τ, infinite sum $\sum_{j>\tau} \xi_j^{(i)}$ can be simply simulated as an independent copy of $X(l_0^{(i)} - \sum_{j\leq\tau} l_j^{(i)})$, so $\hat{Z}_n(\xi)$ can be simulated exactly within finite time. For details on the implementation and efficiency of this algorithm, we refer the readers to Wang and Rhee (2020).

4 EXAMPLES

4.1 Barrier Option Pricing

Let $S_k = Z_1 + \cdots + Z_k$ be a centered random walk with iid increments Z_k such that $\mathbf{P}(Z_1 \geq x) \in \mathbf{RV}_{-\beta}(x)$ and $\mathbf{P}(-Z_1 \geq x) \in \mathbf{RV}_{-\alpha}(x)$ some $\alpha, \beta > 1$. Let $\bar{S}_n(t) = S_{\lfloor nt \rfloor}/n$ and $\bar{S}_n = \{\bar{S}_n(t) : t \in [0,1]\}$ be the scaled version of S_n. Our goal is to estimate $\mathbf{P}(\bar{S}_n \in A)$ where $A \triangleq \{\xi \in \mathbb{D} : \ \xi(1) \geq b, \ \inf_{t\in[0,1]} \xi(t) + ct \leq -a\}$ for some $a, b, c > 0$. This problem is adapted from Section 5 of Chen et al. (2019) and concerns the chance of exercising a down-in barrier option.

According to the framework outlined in Section 3.3, the first step is to identify the solution (l_-^*, l_+^*) to the minimization problem $\min_{(l_-,l_+)\in\mathbb{Z}_+^2 :\ A\cap\mathbb{D}_{l_-,l_+}\neq\emptyset} (\alpha-1)l_- + (\beta-1)l_+$. First, for some piece-wise step function $\xi \in \mathbb{D}_{i,j}$ to fall into set A, it needs at least one downward jump (otherwise $\xi(t) + ct$ is an increasing function, implying $\inf_{t\in[0,t]} \xi(t) + ct = \xi(0) = 0$) and at least one upward jump (otherwise $\xi(t)$ is a decreasing function so $\xi(1) \leq \xi(0) = 0$). In other words, $\mathbb{D}_{i,j} \cap A \neq \emptyset$ only if $i \geq 1$ and $j \geq 1$. Now, to show that $(l_-^*, l_+^*) = (1,1)$, it suffices to find $\xi \in \mathbb{D}_{1,1}$ such that $\xi \in A$. Indeed, for $\xi(t) = -(a+c)\mathbb{1}_{[0.1,1]}(t) + (a+b+c)\mathbb{1}_{[0.5,1]}(t)$, at $t = 0.1$ we have $\xi(t) + ct = -(a+c) + 0.1c < -a$, and at $t = 1$ we have $\xi(1) = -(a+c) + (a+b+c) = b$. This confirms that $\xi \in A \cap \mathbb{D}_{1,1}$ and hence $(l_-^*, l_+^*) = (1,1)$.

We move onto the second step of the framework developed in Section 3.3 and identify γ such that the claim $A \cap (\mathbb{D}_{l_-,l_+} \setminus B^\gamma) = \emptyset$ holds for all (l_-, l_+) with $(\alpha-1)l_- + (\beta-1)l_+ \leq 2(\alpha-1) + 2(\beta-1)$. Here, recall that in the definition of $B^\gamma = \bigcup_{(i,j)\in\partial\mathbb{I}(1,1)} B^{\gamma,i,j}$, where

$$B^{\gamma,i,j} = \Big\{\xi \in \mathbb{D} : \ \#\{t \in [0,1] : \ \xi(t-) - \xi(t) \geq \gamma\} \geq i, \ \#\{t \in [0,1] : \ \xi(t) - \xi(t-) \geq \gamma\} \geq j\Big\}$$

and the index set $\partial\mathbb{I}(l_-^*, l_+^*)$ contains at least (l_-^*, l_+^*). In other words, $\mathbb{D}_{l_-,l_+}\backslash B^{\gamma} \subseteq \mathbb{D}_{l_-,l_+}\backslash B^{\gamma;1,1}$. Next, for any $\xi \in \mathbb{D}_{i,j}\backslash B^{\gamma;1,1}$, all the i upward jumps and j downward jumps in ξ are bounded by γ, and hence

$$\sup_{t\in[0,1]} \xi(t) < j\cdot\gamma \qquad \text{or} \qquad \inf_{t\in[0,1]} \xi(t) > -i\cdot\gamma. \tag{12}$$

Note that for all (l_-, l_+) satisfying $(\alpha-1)l_- + (\beta-1)l_+ \le 2(\alpha-1)+2(\beta-1)$, we must have $l_- \le i^* \triangleq \lceil \big(2(\alpha-1)+2(\beta-1)\big)/(\alpha-1)\rceil$ and $l_+ \le j^* \triangleq \lceil \big(2(\alpha-1)+2(\beta-1)\big)/(\beta-1)\rceil$. Set γ to be small enough such that $j^*\cdot\gamma < b$ and $i^*\cdot\gamma < a$. Then for any (l_-, l_+) satisfying $(\alpha-1)l_- + (\beta-1)l_+ \le 2(\alpha-1)+2(\beta-1)$ and any $\xi \in \mathbb{D}_{i,j}\backslash B^{\gamma;1,1}$, in light of observation (12) we must have either $\xi(1) \le \sup_{t\in[0,1]} < j^*\cdot\gamma < b$ or $\inf_{t\in[0,1]} \xi(t) + ct \ge \inf_{t\in[0,1]} \xi(t) > -i^*\cdot\gamma > -a$, which confirms that $\xi \notin A$. In summary, by picking $\gamma \in (0, \frac{b}{j^*} \wedge \frac{a}{i^*})$ we obtain a strongly efficient importance sampling algorithm for $\mathbf{P}(\bar{S}_n \in A)$.

4.2 First Exit Time of Stochastic Gradient Descents

Consider the potential function $U : \mathbb{R} \to \mathbb{R}$ and interval $I = [s_{\text{left}}, s_{\text{right}}]$ where $s_{\text{left}} < 0 < s_{\text{right}}$ such that (i) $U'(\cdot)$ is Lipschitz continuous, (ii) $U'(0) = 0$, and (iii) $U'(x) < 0$ for $x \in [s_{\text{left}}, 0)$ and $U'(x) > 0$ for $x \in (0, s_{\text{right}}]$. The origin is the unique stable point of $U(\cdot)$ over I in the sense that the path $\boldsymbol{x}_t(x)$, which solves $d\boldsymbol{x}(t)/dt = -U'(\boldsymbol{x}_t(x))$ under initial condition $\boldsymbol{x}_0(x) = x$, converges to 0 as $t \to \infty$ for all $x \in I$. In contrast, the stochastic gradient descent (SGD) iterates (under initial condition $Y^{n|b}(0) \equiv 0$)

$$Y^{n|b}(k) \triangleq Y^{n|b}(k-1) + \varphi_b\Big(-\frac{1}{n}U'\big(Y^{n|b}(k-1)\big) + \frac{1}{n}Z_k\Big) \ \forall k \ge 1$$

will inevitably exit I, where $\varphi_b(w) \triangleq (w \wedge b) \vee (-b)$ represents the standard gradient clipping technique in SGD, and Z_k are iid RVs such that $\mathbf{P}(|Z_1| > x) \in \mathbf{RV}_{-\alpha}(x)$ as $x \to \infty$ for some $\alpha > 1$ and (1) holds. Denote the first exit time as $\tau_n \triangleq \min\{k \ge 0 :\ Y_n(k) \notin I\}$. We are interested in estimating $\mathbf{P}(\tau_n \le n)$ as it indicates the frequency of transitions between different modes and provides important insights about the local and global stability of SGD. We impose the assumptions that $\frac{|s_{\text{left}}|}{b}, \frac{s_{\text{right}}}{b} \notin \mathbb{Z}$. Also, due to the nature of the first exit time problem, by modifying $U(\cdot)$ outside of the compact set I we can assume without loss of generality that $\sup_{x\in\mathbb{R}} |U'(x)| < \infty$.

Let $\bar{Y}^{n|b}(t) = Y_n\big(\lfloor nt \rfloor\big)$ and $\bar{Y}^{n|b} = \{\bar{Y}^{n|b}(t) :\ t \in [0,1]\}$. Let $A = \{\xi \in \mathbb{D} :\ \xi(t) \notin I \ \text{ for some } t \in [0,1]\}$. Note that $\mathbf{P}(\tau_n \le n) = \mathbf{P}(\bar{Y}^{n|b} \in A)$. As a result, we are at the framework developed in Section 3.3 for stochastic difference equations under the choice of $a(\cdot) = -U'(\cdot)$. The first step is to determine $J_b^*(A) \triangleq \min\{k \ge 0 :\ \mathbb{D}_h^{(k)|b} \cap A \ne \emptyset\}$ with $\mathbb{D}_h^{(k)|b} = h^{(k)|b}\big(\mathbb{R}^k \times (0,1]^{k\uparrow}\big)$ and the perturbed ODE mapping $h^{(k)|b}$ defined in (4). In other words, we need to know the number of perturbations required for the ODE under gradient field $-U'(\cdot)$ and initialized at the origin to exit from I. To this end, we develop the following intuition. To cross the right boundary point (i.e., s_{right}), any leftward jump would send the iterates further from the destination, defeating the purpose. Moreover, due to the constant attraction back to the origin under gradient field $-U'(\cdot)$, to cross s_{right} with rightward jumps bounded by b we need to make at least $l_+(b) = \lceil s_{\text{right}}/b \rceil$ jumps. Similarly, to cross the left boundary point s_{left}, we need at least $l_-(b) = \lceil |s_{\text{left}}|/b \rceil$ leftward jumps. The intuition can be made rigorous with the following bound: Given any $\xi = h^{(k)|b}(\boldsymbol{w}, \boldsymbol{t})$,

$$\sup_{t\in[0,1]} \xi(t) \le \sum_{j=1}^{k} \varphi_b\Big(w_j \cdot \mathbb{1}(w_j > 0)\Big), \qquad \inf_{t\in[0,1]} \xi(t) \ge \sum_{j=1}^{k} \varphi_b\Big(w_j \cdot \mathbb{1}(w_j < 0)\Big). \tag{13}$$

Hence, for any $\xi \in \mathbb{D}^{(k)|b}$ with $k < l_-(b) \wedge l_+(b)$, we have $\sup_{t\in[0,1]} \xi(t) \le \lfloor s_{\text{right}}/b \rfloor \cdot b < b$ and $\inf_{t\in[0,1]} \xi(t) > -\lfloor |s_{\text{left}}|/b \rfloor \cdot b > s_{\text{left}}$, which means $\mathbb{D}^{(k)|b} \cap A = \emptyset$. On the other hand, by setting the arrival times of all rightward (resp., leftward) jumps arbitrarily close to 0 and jumps sizes large enough, one can construct

$\xi \in \mathbb{D}^{(l_+(b))|b}$ (resp., $\mathbb{D}^{(l_-(b))|b}$) such that $\sup_{t\in[0,1]} \xi(t) > s_{\text{right}}$ (resp., $\inf_{t\in[0,1]} \xi(t) < s_{\text{left}}$) and hence $\xi(t) \notin I$ for some $t \in [0,1]$. To conclude, we obtain $J_b^*(A) = l_-(b) \wedge l_+(b)$.

Under the choice of $B^\gamma = \big\{\xi \in \mathbb{D} : \#\{t \in [0,1] : |\Delta\xi(t)| \geq \gamma\} \geq J_b^*(A)\big\}$, we are now at the second step of the framework in Section 3.3 and need to determine the range of γ such that $A \cap (\mathbb{D}^{(k)|b} \backslash B^\gamma) = \emptyset \; \forall k \leq 2J_b^*(A)$. For any $\xi \in \mathbb{D}^{(k)|b} \backslash B^\gamma$, the count of jumps with size in $[\gamma, b]$ is at most $J_b^*(A) - 1$, and the count of jumps with size $< \gamma$ is at most k. In light of the bound (13), we then get $\sup_{t\in[0,1]} \xi(t) \leq (l_+(b) - 1) \cdot b + k \cdot \gamma$ and $\inf_{t\in[0,1]} \xi(t) \geq -(l_-(b) - 1) \cdot b - k \cdot \gamma$. Therefore, for any $\gamma \in \big(0, \frac{|s_{\text{left}}| - (l_-(b)-1)\cdot b}{2J_b^*(A)} \wedge \frac{s_{\text{right}} - (l_+(b)-1)\cdot b}{2J_b^*(A)}\big)$, we have $\sup_{t\in[0,1]} \xi(t) < s_{\text{right}}$ and $\inf_{t\in[0,1]} \xi(t) > s_{\text{left}}$ and hence $A \cap (\mathbb{D}^{(k)|b} \setminus B^\gamma) = \emptyset$ for all $k \leq 2J_b^*(A)$. Such γ ensures the strong efficiency when estimating $\mathbf{P}(\tau_n \leq n)$.

4.3 Multiple-Server Queues with Heavy-tailed Service Times

Consider a first-come-first-serve GI/GI/d queueing model with d servers where the inter-arrival times of customers are iid copies of some random variable $V > 0$ and the service times are iid copies of some random variable $S > 0$. We assume that V is light-tailed and there exists some $t > 0$ such that $\mathbf{E}\exp(tV) < \infty$. Additionally, we assume that S is heavy-tailed and $\mathbf{P}(S \geq x) \in \mathbf{RV}_{-\alpha}(x)$ for some $\alpha > 1$. Without loss of generality, we set $\mathbf{E}S = 1$ and denote the arrival rate by $\lambda = 1/\mathbf{E}V$. Let $Q(t)$ be the length of the queue at time t with initial condition $Q(0) = 0$. We are interested in estimating the probability of observing an extreme queue length at time n, i.e., $\mathbf{P}(Q(n) > n\theta)$, for some $\theta > 0$. This estimation is carried out under the stability condition $\lambda < d$. We impose the mild condition that $\lambda - \theta \notin \mathbb{Z}$ and focus on the case where $\theta < \lambda$: otherwise, for event $\{Q(n) > n\theta\}$ to occur there needs to be more than $n\theta$ jobs arriving by time n even though the arrival rate of jobs is $\lambda < \theta$; therefore, observing such an event would require exponentially rare behavior in the light-tailed arrival process of jobs, which is beyond the scope of this tutorial.

Henceforth in Section 4.3, we focus on providing the intuition behind the typical behavior of the queueing system, and we note that the arguments can be made rigorous by adapting the technical tools in Bazhba et al. (2019) to the regularly varying setting at hand.

The first step is to understand the rate of decay for $\mathbf{P}(Q(n) > n\theta)$ as $n \to \infty$. Let $l^* = \lceil d - (\lambda - \theta) \rceil$, and suppose that l^* out of the d servers are completely blocked over the period $[0, n]$. In other words, each of these l^* servers is busy serving some job with extremely high workload and is unable to serve any other jobs. Then in the long run, the jobs arrive at rate λ while each of the remaining $d - l^*$ servers completes jobs at rate 1. As a result, $Q(n)$ should roughly increase at rate $\lambda - (d - l^*) > \lambda - [d - d + (\lambda - \theta)] = \theta$, leading to $Q(n) > n\theta$. In summary, the occurrence of $Q(n) > n\theta$ requires l^* catastrophes (i.e., jobs with extremely high workload), and we expect $\mathbf{P}(Q(n) > n\theta)$ to be of order $(n\mathbf{P}(S \geq n))^{l^*}$.

Next, we determine the choice of event B_n^γ in the importance sampling estimator. Considering the calculations in (10), it only remains to find B_n^γ such that $\mathbf{P}(B_n^\gamma) = O\big((n\mathbf{P}(S \geq n))^{l^*}\big)$ and $\mathbf{P}(\{Q(n) > n\theta\} \backslash B_n^\gamma) = o\big((n\mathbf{P}(S \geq n))^{2l^*}\big)$ as $n \to \infty$. Let $S_j^{(i)}$ be iid copies of S, representing the service time of the j-th customer arrived at the i-th server. Let $S^{(i)}(t) = S_1^{(i)} + \cdots + S_{\lfloor t \rfloor}^{(i)}$. Let $\bar{S}_n^{(i)}(t) = S^{(i)}(nt)/n$ and $\bar{S}_n^{(i)} = \{S_n^{(i)}(t) : \; t \in [0,1]\}$ be the scaled version of $S^{(i)}(t)$. Let $B_n^\gamma \triangleq \{(\bar{S}_n^{(1)}, \cdots, \bar{S}_n^{(d)}) \in B^\gamma\}$ with $B^\gamma = \big\{(\xi^{(1)}, \cdots, \xi^{(d)}) \in \mathbb{D}^d : \sum_{i=1}^d \#\{t \in [0,1] : \xi(t) - \xi(t-) \geq \gamma\} \geq l^*\big\}$. That is, at least l^* of the service times $S_j^{(i)} \; \forall i \in [d], \; j \in [n]$ are larger than $n\gamma$. Obviously, we have $\mathbf{P}(B_n^\gamma) = O\big((n\mathbf{P}(S \geq n))^{l^*}\big)$ due to the l^*-jump nature of the set B^γ. Now, our goal is to find γ such that $\mathbf{P}(\{Q(n) > n\theta\} \backslash B_n^\gamma) = o\big((n\mathbf{P}(S \geq n))^{2l^*}\big)$.

To proceed, we provide the intuition of the typical behavior of the scaled queue $\bar{Q}_n(t) = Q(nt)/n$ on event $(B_n^\gamma)^c$. On event $(B_n^\gamma)^c$ there are at most $l^* - 1$ jobs with service time longer than $n\gamma$, and the most extreme case is that these jobs completely block $l^* - 1$ servers. Suppose that during the time period $[0, n]$ there are also j jobs with workload bounded by $n\gamma$ handled by the remaining $d - l^* + 1$ servers. Then under the scaling of $\bar{Q}_n(t)$, each of these j jobs amounts can block one server by time γ at most. In summary, on event $(B_n^\gamma)^c$, with $l^* - 1 + j$ catastrophes (i.e., jobs with workload of scale $O(n)$), we expect $\bar{Q}_n(1)$ to be upper bounded

by $q(\gamma, j) = \lambda - \big[d - (l^* - 1) - j\gamma\big]$. To ensure $\mathbf{P}(\{Q(n) > n\theta\} \backslash B_n^\gamma) = o\big((n\mathbf{P}(S \geq n))^{2l^*}\big)$, it suffices to find γ small enough such that even with $2l^*$ catastrophes (that is, with $j = l^* + 1$) we still have $q(\gamma, l^* + 1) < \theta$. Now observe that $\lambda - (d - l^* + 1) = \lambda - d - 1 + \lceil d - (\lambda - \theta) \rceil < \lambda - d - 1 + d - (\lambda - \theta) + 1 < \theta$. By setting $\Delta = \theta - [\lambda - (d - l^* + 1)] > 0$ and picking $\gamma \in (0, \frac{\Delta}{l^*+1})$, we identify the construction of B^γ that ensures the strong efficiency of the importance sampling algorithm for the multiple-server queue.

4.4 Fluid Queueing Networks

Consider a fluid queueing network with d stations where jobs arrive independently to the i-th station according to a Poisson process (denoted by $N^{(i)}(t)$) with unit rate. For the i-th station, let $W^{(i)}(k)$ be the workload of the k-th job and suppose the law of $(W^{(i)}(k))_{k \geq 1}$ are iid and $\mathbf{P}(W^{(i)}(k) \geq x) \in \mathbf{RV}_{-\beta_i}(x)$ for some $\beta_i > 1$ (as $x \to \infty$). The total amount of external work arrived at station i by time t is given by $J^{(i)}(t) = \sum_{j=1}^{N^{(i)}(t)} W^{(i)}(j)$. Let $J(t) = (J^{(1)}(t), \cdots, J^{(d)}(t))$, and we use vector $\rho \triangleq \mathbf{E}J(1)$ to denote the expected amount of work of a job at different stations. Each station processes the workload as fluid with rate r_i, and a proportion $Q_{i,j} \in [0,1]$ of the processed fluid will be routed to the j-th station. Let $Z^{(i)}(t)$ be the remaining workload at the i-th station with initial condition $Z^{(i)}(0) = 0 \ \forall i \in [d]$. We are interested in estimating the probability that an extreme amount of workload accumulates in certain parts of the network. To be specific, we set $d = 3$, $\rho = (0.8, 0.8, 1)$, $r = (1, 1, 2.5)$, and

$$Q = \begin{bmatrix} 0 & 0.1 & 0.8 \\ 0.1 & 0 & 0.8 \\ 0 & 0 & 0 \end{bmatrix}$$

and suppose that $\beta_1 + \beta_2 - 2 < \beta_3 - 1$. Our goal is to estimate $\mathbf{P}(Z^{(3)}(n) > na)$ with $a = 0.05$. Henceforth in Section 4.4 we focus on the insights in this example. We refer the readers to Section 6 of Chen et al. (2019) for a thorough and rigorous analysis in the more general setting.

We start by analyzing $\mathbf{P}(Z^{(3)}(n) > an)$. The most obvious cause of $Z^{(3)}(n) > na$ is that a job with high workload arrived at station 3 during the $[0, n]$ period. Using Theorem 5, the probability of this occurrence is approximately of order $n^{-(\beta_3 - 1)}$ as $n \to \infty$. However, note that fluid processed by the first and second stations is partially routed to the third station, which may also lead to the accumulation of fluid at station 3. Therefore, determining the probability of this case is the key step in analyzing $\mathbf{P}(Z^{(3)}(n) > na)$.

Let $\bar{Z}_n^{(i)}(t) = Z^{(i)}(nt)/n \ \forall t \geq 0$ and $\bar{Z}_n(t) = (\bar{Z}_n^{(1)}(t), \bar{Z}_n^{(2)}(t), \bar{Z}_n^{(3)}(t))$ be the scaled version of the queue length process. Typically, the scaled process $\bar{Z}_n(t)$ resembles a fluid network where external fluid arrives constantly at each station with rate ρ_i. By considering the recursive routing $Q_{1,2} = 10\%$ and $Q_{2,1} = 10\%$ between stations 1 and 2, the essential arrival rate of fluid at these two stations should be multiplied by $10/9$ and is equal to $0.8 \cdot 10/9 = 8/9$. Given the routing ratio $Q_{1,3} = Q_{2,3} = 80\%$ and the external arrival rate $\rho_3 = 1$, fluid enters station 3 with rate $1 + 2 \cdot 0.8 \cdot 8/9 = 2.42$. Since the (maximal) service rate is $r_3 = 2.5$, fluid should not accumulate at the third station.

However, if a large job with workload nc arrives at station 1 while station 2 funcitons normally, then under the scaling of $\bar{Z}_n$ station 1 will remain busy and process fluid at rate $r_1 = 1$ for at least time c. A similar calculation can show that, during such periods, the fluid enters station 3 at a rate of $1 + 0.8 \cdot 1 + 0.8 \cdot 0.9 = 2.52$. Since the (maximal) service rate at station 3 is $r_3 = 2.5$, the fluid can only accumulate with rate $2.52 - 2.5 = 0.02$. The same conclusion applies if a large job occupies station 2 while station 1 functions normally, due to the symmetry between stations 1 and 2. In comparison, if both station 1 and station 2 are occupied by jobs with extreme workload, then fluid enters station 3 at a rate of $\rho_3 + r_1 \cdot Q_{1,3} + r_2 \cdot Q_{2,3} = 2.6$ under the scaling of $\bar{Z}_n$. During this busy period, the fluid would typically accumulate with rate $2.6 - 2.5 = 0.1$ at station 3, which eventually leads to $\bar{Z}_n^{(3)}(1) \approx 0.1 > a = 0.05$. Applying Theorem 5 and recalling the assumption $\beta_1 + \beta_2 - 2 < \beta_3 - 1$, we conclude that $\mathbf{P}(\bar{Z}_n^{(3)}(1) > a)$ is of order $\big(n \cdot \mathbf{P}(W^{(1)}(1) \geq n)\big) \cdot \big(n \cdot \mathbf{P}(W^{(2)}(1) \geq n)\big) \in \mathbf{RV}_{-(\beta_1 + \beta_2 - 2)}(n)$.

Next, we determine the choice of event B_n^γ in the importance sampling estimator. Recall that the process $J^{(i)}(t) = \sum_{j=1}^{N^{(i)}(t)} W^{(i)}(j)$ denotes the amount of external work arrived at station i by time t, and define $\bar{J}_n^{(i)}(t) = J^{(i)}(nt)/n \ \forall t \in [0,1]$. In the importance sampling distribution we set $B_n^\gamma = \{(\bar{J}_n^{(1)}, \bar{J}_n^{(2)}, \bar{J}_n^{(3)}) \in B^\gamma\}$ where (let $\mathscr{D}_\gamma(\xi) = \#\{t \in [0,1] : \ \Delta\xi(t) \geq \gamma\}$ for $\xi \in \mathbb{D}$)

$$B^\gamma = \big\{(\xi^{(1)}, \xi^{(2)}, \xi^{(3)}) \in \mathbb{D}^3 : \ \mathscr{D}_\gamma(\xi^{(3)}) \geq 1 \text{ or } \min\{\mathscr{D}_\gamma(\xi^{(1)}), \mathscr{D}_\gamma(\xi^{(2)})\} \geq 1\big\}.$$

Note that B_n^γ captures the two different causes of $\{Z^{(3)}(n) > na\}$ identified earlier. Using Theorem 5, we see that $\mathbf{P}(B_n^\gamma) = O(\mathbf{P}(Z^{(3)}(n) > na))$ as $n \to \infty$. Considering the calculations in (10), we reduce the problem to finding γ small enough such that $\mathbf{P}(\{Z^{(3)}(n) > na\} \backslash B_n^\gamma) = o(\mathbf{P}^2(Z^{(3)}(n) > na))$. Through an in-depth analysis of the corresponding Skorokhod problem, one can show that it suffices to pick $\gamma > 0$ such that

$$\min\Big\{\lceil \frac{1}{20}/\gamma \rceil (\beta_3 - 1), \ \lceil \frac{3}{20}/\gamma \rceil (\beta_1 - 1) + (\beta_2 - 1), \ \lceil \frac{3}{20}/\gamma \rceil (\beta_2 - 1) + (\beta_1 - 1)\Big\} > 2(\beta_1 + \beta_2 - 1).$$

Here, we briefly describe the intuition and refer the interested readers to Section 6.3 of Chen et al. (2019) for details. Let l_i be the number of large jobs (i.e., with workload of scale $O(n)$) that arrived at station i by time n. On event $(B_n^\gamma)^c$, station 3 received no job with workload $\geq n\gamma$ by time n, and at least one of stations 1 and 2 received no job with workload $\geq n\gamma$ by time n. Consider an extreme case where station 1 is completely occupied due to some job (with workload $\geq n\gamma$) during period $[0,n]$. Then stations 2 and 3 receive l_2 and l_3 large jobs respectively by time n, with workload bounded by $n\gamma$. First, the contribution of the l_3 jobs at station 3 to $\bar{Z}_n^{(3)}(1)$ is bounded by $l_3 \cdot \gamma$. For $\gamma > 0$ small enough, we have $a - l_3 \cdot \gamma > 0.02$. Next, to cover the remaining $a - l_3 \cdot \gamma$ gap, we haven shown that $\bar{Z}_n^{(3)}(1)$ increases at rate $0.02 < a - l_3 \cdot \gamma$ when station 1 is blocked and station 2 works normally, and it increases at rate 0.1 when both station 1 and station 2 are blocked. The latter only happens when station 2 is also processing the l_2 large jobs. In particular, the smaller γ is, the shorter such busy periods would be. By picking $\gamma > 0$ small enough we ensure that $\bar{Z}_n^{(3)}(1) < a$ for all $\mathscr{J}(l_1, l_2, l_3) \leq 2(\beta_1 + \beta_2 - 2)$ with the cost function $\mathscr{J}(\cdot)$ in (5), thus implying $\mathbf{P}(\{Z^{(3)}(n) > na\} \backslash B_n^\gamma) = o(\mathbf{P}^2(Z^{(3)}(n) > na))$.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the National Science Foundation [Grant CMMI-2146530].

REFERENCES

Asmussen, S., and D. P. Kroese. 2006. "Improved Algorithms for Rare Event Simulation with Heavy Tails". *Advances in Applied Probability* 38(2):545–558.

Bassamboo, A., S. Juneja, and A. Zeevi. 2007. "On the Inefficiency of State-Independent Importance Sampling in the Presence of Heavy Tails". *Operations Research Letters* 35(2):251–260.

Bazhba, M., J. Blanchet, C.-H. Rhee, and B. Zwart. 2019. "Queue Length Asymptotics for the Multiple-Server Queue with Heavy-Tailed Weibull Service Times". *Queueing Systems* 93:195–226.

Blanchet, J., and P. Glynn. 2008. "Efficient Rare-Event Simulation for the Maximum of Heavy-Tailed Random Walks". *The Annals of Applied Probability* 18(4):1351 – 1378.

Blanchet, J., P. Glynn, and J. Liu. 2008. "Efficient Rare Event Simulation for Heavy-Tailed Multiserver Queues". Technical report, Department of Statistics, Columbia University.

Blanchet, J., H. Hult, and K. Leder. 2013, dec. "Rare-Event Simulation for Stochastic Recurrence Equations with Heavy-Tailed Innovations". *ACM Trans. Model. Comput. Simul.* 23(4).

Blanchet, J. H., and J. Liu. 2008. "State-Dependent Importance Sampling for Regularly Varying Random Walks". *Advances in Applied Probability* 40(4):1104–1128.

Boxma, O. J., E. J. Cahen, D. Koops, and M. Mandjes. 2019. "Linear Stochastic Fluid Networks: Rare-Event Simulation and Markov Modulation". *Methodology and Computing in Applied Probability* 21(1):125–153.

Bucklew, J. A., P. Ney, and J. S. Sadowsky. 1990. "Monte Carlo Simulation and Large Deviations Theory for Uniformly Recurrent Markov Chains". *Journal of Applied Probability* 27(1):44–59.

Chen, B., J. Blanchet, C.-H. Rhee, and B. Zwart. 2019. "Efficient Rare-Event Simulation for Multiple Jump Events in Regularly Varying Random Walks and Compound Poisson Processes". *Mathematics of Operations Research* 44(3):919–942.

Cohen, J. E., R. A. Davis, and G. Samorodnitsky. 2022. "COVID-19 Cases and Deaths in the United States Follow Taylor's Law for Heavy-Tailed Distributions with Infinite Variance". *Proceedings of the National Academy of Sciences* 119(38):e2209234119.

Dupuis, P., K. Leder, and H. Wang. 2007. "Importance Sampling for Sums of Random Variables with Regularly Varying Tails". *ACM Trans. Model. Comput. Simul.* 17(3):Article 14.

Dupuis, P., A. D. Sezer, and H. Wang. 2007. "Dynamic Importance Sampling for Queueing Networks". *The Annals of Applied Probability* 17(4):1306 – 1346.

Dupuis, P., and H. Wang. 2004. "Importance Sampling, Large Deviations, and Differential Games". *Stochastics and Stochastic Reports* 76(6):481–508.

Dupuis, P., and H. Wang. 2005. "On the Convergence from Discrete to Continuous Time in an Optimal Stopping Problem". *The Annals of Applied Probability* 15(2):1339 – 1366.

Dupuis, P., and H. Wang. 2009. "Importance Sampling for Jackson Networks". *Queueing Systems* 62(1-2):113–157.

Embrechts, P., C. Klüppelberg, and T. Mikosch. 2013. *Modelling Extremal Events: for Insurance and Finance*, Volume 33. Springer Science & Business Media.

Glasserman, P., and S.-G. Kou. 1995. "Analysis of an Importance Sampling Estimator for Tandem Queues". *ACM Trans. Model. Comput. Simul.* 5(1):22–42.

Glasserman, P., and Y. Wang. 1997. "Counterexamples in Importance Sampling for Large Deviations Probabilities". *The Annals of Applied Probability* 7(3):731 – 746.

Gudmundsson, T., and H. Hult. 2014. "Markov Chain Monte Carlo for Computing Rare-Event Probabilities for a Heavy-Tailed Random Walk". *Journal of Applied Probability* 51(2):359–376.

Gurbuzbalaban, M., U. Simsekli, and L. Zhu. 2021. "The Heavy-Tail Phenomenon in SGD". In *Proceedings of the 38th International Conference on Machine Learning*, edited by M. Meila and T. Zhang, Volume 139 of *Proceedings of Machine Learning Research*, 3964–3975: PMLR.

Hult, H., S. Juneja, and K. Murthy. 2016. "Exact and Efficient Simulation of Tail Probabilities of Heavy-Tailed Infinite Series". arXiv:1609.01807[v1] (math.PR).

Murthy, K. R. A., S. Juneja, and J. Blanchet. 2014. "State-Independent Importance Sampling for Random Walks with Regularly Varying Increments". *Stochastic Systems* 4(2):321–374.

Pitman, J., and G. U. Bravo. 2012. "The Convex Minorant of a Lévy Process". *The Annals of Probability* 40(4):1636 – 1674.

Resnick, S. I. 2007. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer Science & Business Media.

Rhee, C.-H., J. Blanchet, and B. Zwart. 2017. "Sample Path Large Deviations for Lévy Processes and Random Walks with Regularly Varying Increments". arXiv:1606.02795[v2] (math.PR).

Rhee, C.-H., J. Blanchet, and B. Zwart. 2019. "Sample Path Large Deviations for Lévy Processes and Random Walks with Regularly Varying increments". *The Annals of Probability* 47(6):3551–3605.

Rhee, C.-H., and P. W. Glynn. 2015. "Unbiased Estimation with Square Root Convergence for SDE Models". *Operations Research* 63(5):1026–1043.

Sato, K.-i., S. Ken-Iti, and A. Katok. 1999. *Lévy Processes and Infinitely Divisible Distributions*. Cambridge university press.

Torrisi, G. 2004. "Simulating the ruin probability of risk processes with delay in claim settlement". *Stochastic Processes and their Applications* 112(2):225–244.

Wang, X., and C.-H. Rhee. 2020. "Rare-Event Simulation for Multiple Jump Events in Heavy-Tailed Lévy Processes with Infinite Activities". In *Proceedings of the 2020 Winter Simulation Conference*, edited by K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, 409–420. Piscataway, New Jersey: IEEE.

Wang, X., and C.-H. Rhee. 2023. "Large Deviations and Metastability Analysis for Heavy-Tailed Dynamical Systems". arXiv:2307.03479[v1] (math.PR).

AUTHOR BIOGRAPHIES

XINGYU WANG is is a PhD candidate in the Department of Industrial Engineering and Management Sciences at Northwestern University. His research interests include stochastic simulation and applied probability. His e-mail address is xingyuwang2017@u.northwestern.edu, and his website is https://joshwang0322.github.io.

CHANG-HAN RHEE is an Assistant Professor in Industrial Engineering and Management Sciences at Northwestern University. He received his Ph.D. in Computational and Mathematical Engineering from Stanford University. His research interests include stochastic simulation, applied probability, and machine learning. His e-mail address is chang-han.rhee@northwestern.edu, and his website is https://chrhee.github.io.