CWI

Simulation-based computation of the workload correlation function in a Lévy-driven queue

P.W. Glynn, M.R.H. Mandjes

PNA-1004

Centrum Wiskunde & Informatica

Centrum Wiskunde & Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Computing (MAC)

Information Systems (INS)

Copyright © 2010, Centrum Wiskunde & Informatica
P.O. Box 94079, 1090 GB Amsterdam (NL)
Science Park 123, 1098 XG Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

SIMULATION-BASED COMPUTATION OF THE WORKLOAD CORRELATION FUNCTION IN A LÉVY-DRIVEN QUEUE

PETER W. GLYNN AND MICHEL MANDJES

ABSTRACT. In this paper we consider a single-server queue with Lévy input, and in particular its workload process $(Q_t)_{t\geq 0}$, focusing on its correlation structure. With the correlation function defined as $r(t) := \mathbb{C}\mathrm{ov}(Q_0, Q_t)/\mathbb{V}\mathrm{ar}\, Q_0$ (assuming the workload process is in stationarity at time 0), we first study its transform $\int_0^\infty r(t)e^{-\vartheta t}\mathrm{d}t$, both for the case that the Lévy process has positive jumps, and that it has negative jumps. These expressions allow us to prove that $r(\cdot)$ is positive, decreasing, and convex, relying on the machinery of completely monotone functions. For the light-tailed case, we estimate the behavior of $r(t)$ for t large. We then focus on techniques to estimate $r(t)$ by simulation. Naive simulation techniques require roughly $(r(t))^{-2}$ runs to obtain an estimate of a given precision, but we develop a coupling technique that leads to substantial variance reduction (required number of runs being roughly $(r(t))^{-1}$). If this is augmented with importance sampling, it even leads to a logarithmically efficient algorithm.

KEYWORDS. Lévy processes ⋆ reflection ⋆ workload process ⋆ correlation function ⋆ simulation ⋆ coupling ⋆ importance sampling

1. INTRODUCTION

Consider a queueing system, and, more particularly, its workload process $(Q_t)_{t\geq 0}$. Where one usually focuses on the characterization of the (transient or steady-state) workload, another interesting problem relates to the identification of the *workload correlation function* $r(t) := \mathbb{C}\mathrm{ov}(Q_0, Q_t)/\mathbb{V}\mathrm{ar}\, Q_0$, assuming that the workload process is in stationarity at time 0. For several queueing systems this correlation function has been explicitly computed; [17], for instance, analyzes the number of customers in the M/M/1 queue. Often explicit formulae are hard to obtain, but the analysis simplified greatly when looking at the transform

$$\rho(\vartheta) := \int_0^\infty r(t)e^{-\vartheta t}\mathrm{d}t.$$

In his seminal paper [5], Beneš managed to compute $\rho(\cdot)$ for the workload in the M/G/1 queue; relying on the concept of complete monotonicity, [18] elegantly proved that, in this case, $r(\cdot)$ is positive, decreasing and convex. We further mention the survey by [20], and interesting results by [1].

The primary aim of this paper is to explore the workload correlation function for the class of single-server queues fed by *Lévy processes*. Notice that the M/G/1 queue is contained

Date: May 21, 2010.

Part of this work was carried out when MM was at Stanford, and another part when both PG and MM were visiting the Isaac Newton Institute, Cambridge, UK.

in this class; then the Lévy process under consideration is a compound Poisson process with drift. We focus on *spectrally one-sided Lévy input processes*, distinguishing between those with only positive jumps (also referred to as *spectrally positive*), and those with only negative jumps (*spectrally negative*). For the spectrally positive case it was already shown in [12] that $r(\cdot)$ is positive, decreasing, and convex; our first contribution is that we use the results of [10, 19] to show that these properties carry over to the spectrally-negative case. We also estimate the asymptotics of $r(t)$ for t large. These results can be found in Section 2.

A second contribution of the paper (Section 3) considers an intimately related problem: the analysis of the distribution of the residual busy period τ, where the queue starts in stationarity at time 0; the insights developed in this section will be intensively used in Section 4, when setting up schemes to efficiently simulate $r(t)$. For spectrally one-sided input we first derive the Laplace transform of $p(t) := \mathbb{P}(\tau > t)$. Then we use this transform to estimate the tail of $p(t)$ for the case of light-tailed Lévy input, which exhibits (essentially) exponential decay. The fact that $p(t) \to 0$ for $t \to \infty$ implies that estimation through 'naive' simulation may take prohibitively long for large t. We develop a logarithmically efficient importance sampling algorithm; in this scheme the Lévy input (in the interval $(0, t]$) is given a constant exponential twist, but, remarkably, also the workload present at time 0 needs to be sampled from an alternative distribution as well.

The third contribution, presented in Section 4, concerns efficient simulation schemes for estimating $r(t)$; these intensively rely on results that we found for the busy-period distribution $p(t)$. Again, the fact $r(t) \to 0$ (as $t \to \infty$) entails that naive simulation will be extremely time-consuming; we show it takes even roughly $(r(t))^{-2}$ runs to obtain an estimate of a given precision. Then we propose a coupling-based approach yielding substantial variance reduction (so that the number of runs required is just of the order $(r(t))^{-1}$. For the light-tailed case (in which $r(t)$ vanishes essentially exponentially) we propose an importance-sampling based algorithm; if this is applied on top of the coupling technique, then the resulting scheme is asymptotically efficient (i.e., the number of replications needed grows subexponentially in t). To our best knowledge this is the first contribution to variance reduction in the context of the estimation of (small) correlations and covariances. As indicated above, developing simulation-based computation techniques for this is substantially more challenging than for rare-event probabilities.

In Section 5 we present a number of simulation experiments, for the cases of reflected Brownian motion and the M/M/1 queue, showing the substantial speed up achieved by our approach. Section 6 concludes, and discusses a number of open issues.

2. Model and structural results

In this section we find an expression for the transform $\rho(\cdot)$ of the correlation function, which is used to derive a number of structural properties of $r(\cdot)$, as well as asymptotics. We start this section, however, with a formal introduction of our queueing system.

2.1. **Lévy Processes.**

Let $(X_t)_{t\geq 0}$ be a Lévy process, with drift $\mathbb{E}X_1 < 0$. We consider two cases.

(A) $(X_t)_{t\geq 0}$ has no negative jumps. Then the Laplace exponent is given by the function $\varphi(\cdot) : [0,\infty) \mapsto [0,\infty)$, i.e., $\varphi(\alpha) := \log \mathbb{E}e^{-\alpha X_1}$. It is known that $\varphi(\cdot)$ is increasing and convex on $[0,\infty)$, with slope $\varphi'(0) = -\mathbb{E}X_1$ in the origin. Therefore the inverse $\psi(\cdot)$ of $\varphi(\cdot)$ is well-defined on $[0,\infty)$. In the sequel we also require that X_t is not a *subordinator*, i.e., a monotone process; thus X_1 has probability mass on the positive half-line, which implies that $\lim_{\alpha\to-\infty}\varphi(\alpha) = \infty$.

(B) $(X_t)_{t\geq 0}$ has no positive jumps. Now we define $\Phi(\beta) := \log \mathbb{E}e^{\beta X_1}$, which is well-defined for any $\beta \geq 0$. Again ruling out that X_t is a subordinator (and recalling that $\Phi'(0) = \mathbb{E}X_1 < 0$), we see that $\Phi(\beta)$ is no bijection on $[0,\infty)$; we define the *right* inverse through $\Psi(q) := \sup\{\beta \geq 0 : \Phi(\beta) = q\}$. Realize that $\Psi(0) > 0$.

Important examples of such Lévy processes are the following. (1) *Brownian motion with drift*, being actually both spectrally positive and negative. We write $X \in \mathbb{B}\mathrm{m}(\mu,\sigma^2)$ when $\varphi(\alpha) = -\alpha\mu + \frac{1}{2}\alpha^2\sigma^2$. (2) *Compound Poisson with drift*, which is spectrally positive. Non-negative jobs arrive according to a Poisson process of rate λ; the jobs $B_1, B_2, \ldots$ are i.i.d. samples from a distribution with Laplace transform $b(\alpha) := \mathbb{E}e^{-\alpha B}$; the storage system is continuously depleted at a rate 1. We write $X \in \mathbb{C}\mathrm{P}(\lambda, b(\cdot))$; it can be verified that $\varphi(\alpha) = \alpha - \lambda + \lambda b(\alpha)$. Clearly, if the drift would be positive, and the jobs would be i.i.d. samples from a non-positive distribution (that is, the jumps are downward), the process is spectrally negative.

2.2. **Reflected Lévy Processes; Queues.** We consider the reflection of $(X_t)_{t\geq 0}$ at 0, which we denote by $(Q_t)_{t\geq 0}$. It is formally introduced as follows, see for instance [3, Ch. IX]. Define the decreasing process $(M_t)_{t\geq 0}$ and the resulting reflected process (or: workload process, queueing process) $(Q_t)_{t\geq 0}$ through

$$M_t = \inf_{0\leq s\leq t} X_s; \quad Q_t := X_t + \max\{-M_t, Q_0\};$$

observe that $Q_t \geq 0$ for all $t \geq 0$. Then the steady-state distribution $Q := \lim_{t\to\infty} Q_t$, which exists due to $\mathbb{E}X_1 < 0$, is known (in terms of its Laplace transform) for both the spectrally positive and spectrally negative case. For spectrally positive input, we have the *generalized Pollaczek-Khinchine formula*, usually attributed to [22]:

$$\kappa(\alpha) := \mathbb{E}e^{-\alpha Q} = \frac{\alpha\varphi'(0)}{\varphi(\alpha)}. \tag{1}$$

This result evidently enable the computation of all moments of the steady-state queue Q (by repeated differentiation and inserting 0). From now one we assume $\mathbb{E}Q^2$ to be finite, so that $v := \mathbb{V}\mathrm{ar}\, Q$ is well-defined.

For spectrally negative input, realize that $\mathbb{E}e^{\beta_0 X_t}$ is a martingale, with $\beta_0 := \Psi(0) > 0$. 'Optional sampling' [21, Ch. A14] thus gives, for any positive x,

$$\mathbb{P}(\exists t \geq 0 : X_t > x)e^{\beta_0 x} = 1,$$

and as Q is distributed as the supremum over $t \geq 0$ of X_t ('Reich's identity'), we obtain Q is exponentially distributed with mean $1/\beta_0$. It follows that $v = 1/\beta_0^2$.

2.3. **Correlation Structure of the Queue.** In this paper we are interested in the correlation structure of the queue process $(Q_t)_{t\geq 0}$. For the spectrally-positive case, structural results were already found in [12]. Relying on the transform of Q_T (where T is exponentially distributed with mean ϑ^{-1}) given that $Q_0 = x$, see e.g. [3, Section IX.3] and [14], it was derived that

$$\rho(\vartheta) := \int_0^\infty r(t)\, e^{-\vartheta t}\mathrm{d}t = \frac{1}{\vartheta} - \frac{\varphi''(0)}{2v\vartheta^2} + \frac{\varphi'(0)}{v\vartheta^2}\left[\frac{1}{\vartheta\psi'(\vartheta)} - \frac{1}{\psi(\vartheta)}\right].$$

Then the machinery of completely monotone functions [6, 18] was used to prove that $r(\cdot)$ is a positive, decreasing, and convex function. We now do the same for the spectrally-negative case.

Following the setup of [15, Chapter 8], we first introduce, for spectrally negative Lévy processes, families of functions $W^{(q)}(\cdot)$ and $Z^{(q)}(\cdot)$ as follows. Let $W^{(q)}(x)$ be a strictly increasing and continuous function whose Laplace transform satisfies

$$\int_0^\infty e^{-\beta x} W^{(q)}(x)\mathrm{d}x = \frac{1}{\Phi(\beta) - q}, \qquad \beta > \Psi(q). \tag{2}$$

In addition,

$$Z^{(q)}(x) := 1 + q\int_0^x W^{(q)}(y)\mathrm{d}y. \tag{3}$$

$W^{(q)}(\cdot)$ and $Z^{(q)}(\cdot)$ are usually referred to as the q-scale functions. Then the results of [19], in conjunction with Exercise 8.5 (both parts (i) and (ii)) of [15] lead, with some abuse of notation, to the following transform (with respect to t) of the density of Q_t, given that $Q_0 = x$:

$$\int_0^\infty e^{-qt}\mathbb{P}_x(Q_t = y)\mathrm{d}t = e^{-\Psi(q)y}\frac{\Psi(q)}{q}Z^{(q)}(x) - W^{(q)}(x-y).$$

It is now a matter of straightforward calculus to show that the previous display leads to, with T denoting an exponential random variable with mean q^{-1},

$$\int_0^\infty e^{-\beta x}\mathbb{E}_x e^{-\alpha Q_T}\mathrm{d}x = I_1 - I_2;$$

where

$$\begin{aligned} I_1 &:= \int_0^\infty\int_0^\infty q e^{-\beta x}e^{-\alpha y}e^{-\Psi(q)y}\frac{\Psi(q)}{q}Z^{(q)}(x)\mathrm{d}x\mathrm{d}y, \\ I_2 &:= \int_0^\infty\int_0^\infty q e^{-\beta x}e^{-\alpha y}W^{(q)}(x-y)\mathrm{d}x\mathrm{d}y. \end{aligned}$$

We now compute $I_1 \equiv I_1(\alpha,\beta,q)$ and $I_2 \equiv I_2(\alpha,\beta,q)$ explicitly. Let us first consider the integral I_1; using (2) and (3), we obtain

$$\begin{aligned} I_1(\alpha,\beta,q) &= \frac{\Psi(q)}{\Psi(q)+\alpha}\int_0^\infty e^{-\beta x}Z^{(q)}(x)\mathrm{d}x \\ &= \frac{\Psi(q)}{\Psi(q)+\alpha}\left(\frac{1}{\beta} + \int_0^\infty\int_y^\infty qW^{(q)}(y)e^{-\beta x}\mathrm{d}x\mathrm{d}y\right) \\ &= \frac{\Psi(q)}{\Psi(q)+\alpha}\frac{1}{\beta}\left(1 + \frac{q}{\Phi(\beta)-q}\right). \end{aligned}$$

Likewise,

$$I_2(\alpha,\beta,q) = \int_0^\infty qe^{-(\alpha+\beta)y}\frac{1}{\Phi(\beta)-q}\mathrm{d}y = \frac{q}{\alpha+\beta}\frac{1}{\Phi(\beta)-q}.$$

Let us perform a few checks; it is readily verified that

- plugging in $\alpha = 0$ in $I_1(\alpha,\beta,q) - I_2(\alpha,\beta,q)$ indeed yields $1/\beta$;
- plugging in $\beta = \beta_0$ into the expression for $\int_0^\infty \beta e^{-\beta x}\mathbb{E}_x e^{-\alpha Q_T}\mathrm{d}x$ indeed yields the steady-state transform $\beta_0/(\beta_0+\alpha)$: when starting in the queue's equilibrium distribution at time 0, the workload is still in stationarity after an exponentially distributed time (irrespectively of q).

Now observe that, recalling that T has an exponential distribution with mean q^{-1},

$$\begin{aligned}\int_0^\infty qe^{-qt}\mathbb{E}(Q_0Q_t)\mathrm{d}t &= \int_0^\infty \beta_0 x e^{-\beta_0 x}\mathbb{E}_x Q_T\,\mathrm{d}x \\ &= \lim_{\alpha\downarrow 0}\frac{\mathrm{d}}{\mathrm{d}\alpha}\left[\beta\cdot\frac{\mathrm{d}}{\mathrm{d}\beta}\int_0^\infty e^{-\beta x}\mathbb{E}_x e^{-\alpha Q_T}\mathrm{d}x\Big|_{\beta=\beta_0}\right].\end{aligned} \tag{4}$$

Upon combining the explicit expression for $I_1(\alpha,\beta,q) - I_2(\alpha,\beta,q)$ with (4), and recalling that $v = 1/\beta_0^2$ (in the spectrally-negative case), we eventually find, after considerable calculus, the following result.

Theorem 2.1. For the spectrally-negative case,

$$\rho(q) := \int_0^\infty r(t)\,e^{-qt}\mathrm{d}t = \frac{1}{q} + \frac{\beta_0^2}{q^2}\Phi'(\beta_0)\left(\frac{1}{\Psi(q)} - \frac{1}{\beta_0}\right).$$

The following corollary follows from applying 'L'Hôpital' twice. It implies that in the spectrally-negative case the workload process is necessarily short-range dependent. Use that $\Psi'(0)\Phi'(\beta_0) = 1$ and $\Phi''(\beta_0) + (\Phi'(\beta_0))^3\Psi''(0) = 0$, which follow from repeated differentiation of the relation $\Phi(\Psi(q)) = q$.

Corollary 2.2. For the spectrally-negative case,

$$\rho(0) := \int_0^\infty r(t)\mathrm{d}t = \frac{1}{\beta_0\Phi'(\beta_0)} + \frac{\Phi''(\beta_0)}{2(\Phi'(\beta_0))^3} < \infty.$$

We can now use the transform $\rho(q)$ to establish a number of key structural properties of $r(\cdot)$.

Theorem 2.3. $r(\cdot)$ is positive, decreasing, and convex.

Proof: We mimic the proof that was developed in [12] for the spectrally-positive case. Using integration by parts, we find that

$$\rho^{(1)}(q) := \int_0^\infty r'(t)e^{-qt}\mathrm{d}t = \frac{\beta_0^2}{q}\Phi'(\beta_0)\left(\frac{1}{\Psi(q)} - \frac{1}{\beta_0}\right),$$

which also entails that $r'(0) = -\beta_0\Phi'(\beta_0)$. Analogously,

$$\rho^{(2)}(q) := \int_0^\infty r''(t)e^{-qt}\mathrm{d}t = -r'(0) + \beta_0^2\Phi'(\beta_0)\left(\frac{1}{\Psi(q)} - \frac{1}{\beta_0}\right) = \beta_0^2\frac{\Phi'(\beta_0)}{\Psi(q)}. \tag{5}$$

In the proof of Prop. 3.2 we will show that $\Psi(0)/\Psi(q) \in \mathscr{C}$, where $\mathscr{C}$ is the class of completely monotone functions [6, 13]; completely monotone functions are functions that can, up to some positive multiplicative constant, be considered as Laplace transforms of non-negative random variables. We conclude from (5) that $\rho^{(2)}(q)$ is in $\mathscr{C}$, and hence $r''(\cdot)$ is positive, i.e., $r(\cdot)$ is convex.
We know that $f(q) \in \mathscr{C}$ implies that, with $g(q) := (f(0) - f(q))/q$, also $g(q) \in \mathscr{C}$. Taking $f(q) = \rho^{(2)}(q)$, we obtain that $-\rho^{(1)}(q)$ is in $\mathscr{C}$, and hence $r'(\cdot)$ is negative, i.e., $r(\cdot)$ is decreasing. Applying the same procedure again, we find that $\rho(q)$ is in $\mathscr{C}$, and hence $r(\cdot)$ is positive. □

In [12] the asymptotics of $r(t)$ (for t large) in the spectrally-positive case were addressed. It turned out that the heavy-tailed regime (leading to $r(t)$ decaying essentially polynomially) and the light-tailed regime (leading to $r(t)$ decaying essentially exponentially) had to be treated separately. In the light-tailed regime (where we assume that the equation $\varphi(\alpha) = 0$ has a negative root) it turned out that the exact asymptotics were, up to a multiplicative constant, of the form $t^{-3/2}e^{\vartheta^\star t}$, where $\vartheta^\star < 0$ is the branching point of $\psi(\cdot)$. This means that, with $\zeta < 0$ being the minimizer of $\varphi(\cdot)$, $\varphi(\zeta) = \vartheta^\star$.
Let us now consider the counterpart of these findings for the spectrally-negative case. We will argue that $r(t)$ necessarily decays exponentially, relying on the Heaviside operational principle. Let $\zeta > 0$ denote the minimizer of $\Phi(\cdot)$, so that $\Phi(\zeta) = q^\star < 0$; hence $q^\star < 0$ is the branching point of $\Psi(\cdot)$. Around $q^\star$ we have that $\Psi(q)$ looks like $\zeta + \sqrt{2/v_\Phi} \cdot \sqrt{q - q^\star}$, with $v_\Phi := \Phi''(\zeta) > 0$. After some calculus we obtain that this entails that, for some (irrelevant) constant κ,

$$\rho(q) \sim \kappa + B_\Phi \sqrt{q - q^\star}; \quad B_\Phi := -\frac{\beta_0^2 \Phi'(\beta_0)}{(q^\star)^2 \zeta^2} \sqrt{\frac{2}{v_\Phi}} < 0,$$

so that application of Heaviside heuristics [2] yields, with $f(t) \sim g(t)$ denoting $f(t)/g(t) \to 1$ as $t \to \infty$,

$$r(t) \sim \frac{B_\Phi}{\Gamma(-\frac{1}{2})} \cdot \frac{e^{q^\star t}}{t\sqrt{t}}.$$

3. An intermezzo: efficient estimation of the busy period tail distribution

In this section we address the estimation of the tail distribution of the busy period in a Lévy-driven queue by applying an importance-sampling based simulation procedure. In the next section it will turn out that the insights developed here are useful when setting up an efficient simulation scheme for estimating the workload correlation $r(t)$. We let τ denote the busy-period duration, starting from steady-state at time 0: $\tau := \inf\{t \geq 0 : Q_t = 0\}$, where Q_0 is distributed according to the stationary distribution. Throughout this section we write $p(t) := \mathbb{P}(\tau > t)$. In this section we first derive the Laplace transform of the probability $p(\cdot)$, then we consider the corresponding asymptotics, and finally we set up a logarithmically efficient simulation scheme.

3.1. **Transforms.** Let us start by considering the spectrally-positive case. We have, with $\tau(x) := \inf\{t \geq 0 : X_t = -x\}$

$$\begin{aligned}\int_0^\infty e^{-\vartheta t} p(t)\mathrm{d}t &= \int_0^\infty \left(\int_0^\infty e^{-\vartheta t} \mathbb{P}(\tau(x) > t)\mathrm{d}t \right) \mathrm{d}\mathbb{P}(Q_0 < x) \\ &= \frac{1}{\vartheta} \int_0^\infty \left(1 - e^{-\psi(\vartheta)x}\right) \mathrm{d}\mathbb{P}(Q_0 < x).\end{aligned}$$

Application of 'Pollaczek-Khinchine' now leads to the following result.

Proposition 3.1. In the spectrally-positive case, the Laplace transform of $p(t)$ is given by

$$\int_0^\infty e^{-\vartheta t} p(t)\mathrm{d}t = \frac{1}{\vartheta} - \varphi'(0)\frac{\psi(\vartheta)}{\vartheta^2}.$$

The spectrally-negative case can be dealt with similarly. First recall that

$$\int_0^\infty e^{-qt}\mathbb{P}(\tau > t)\mathrm{d}t = q^{-1}\left(1 - \mathbb{E}e^{-q\tau}\right).$$

Then, using part (ii) of [15, Exercise 6.7], we have

$$\mathbb{E}e^{-q\tau} = \int_0^\infty \beta_0 e^{-\beta_0 x}\mathbb{E}e^{-q\tau(x)}\mathrm{d}x = \beta_0 \cdot \frac{\hat{\kappa}(q,\beta_0) - \hat{\kappa}(q,0)}{\beta_0 \hat{\kappa}(q,\beta_0)};$$

here $\hat{\kappa}(q,\beta)$ relates to the transform of the so-called *descending ladder process*, and is given, in this spectrally-negative case, by $\hat{\kappa}(q,\beta) = (q - \Phi(\beta))/(\Psi(q) - \beta)$. Using that $\Phi(\beta_0) = 0$, we find that $\mathbb{E}e^{-q\tau} = \Psi(0)/\Psi(q)$, and in addition the following result is obtained.

Proposition 3.2. In the spectrally-negative case, the Laplace transform of $p(t)$ is given by

$$\int_0^\infty e^{-qt} p(t)\mathrm{d}t = \frac{1}{q}\left(1 - \frac{\Psi(0)}{\Psi(q)}\right).$$

3.2. **Asymptotics.** . We again use the Heaviside operational principle [2] to (heuristically) estimate the decay of $p(t)$ for t large. We focus on the situation that the Lévy process is (in the upward direction) *light-tailed*; precise definitions follow below. The most important conclusion is that in this light-tailed case $p(t)$ decays to 0 essentially exponentially; up to a multiplicative constant, the exact asymptotics coincide with those of the workload correlation function $r(t)$.

We again start by considering the spectrally-positive case. As before, we assume that the equation $\varphi(\alpha) = 0$ has a negative root. Observe that then Prop. 3.1 holds for any positive ϑ, but we can consider the analytic continuation up to the branching point $\vartheta^\star < 0$ of $\psi(\cdot)$; let in the sequel $\zeta < 0$ denote the minimizer of $\varphi(\cdot)$, so that $\varphi(\zeta) = \vartheta^\star < 0$ (where it is noticed that $v_\varphi := \varphi''(\zeta) > 0$). Then the idea is to write, for $\vartheta \downarrow \vartheta^\star$ we have that $\psi(\vartheta) - \zeta \sim \sqrt{2/v_\varphi} \cdot \sqrt{\vartheta - \vartheta^\star}$. Hence, around $\vartheta^\star$, we have that, for some (irrelevant) constant κ,

$$\int_0^\infty e^{-\vartheta t} p(t)\mathrm{d}t = \frac{1}{\vartheta} - \varphi'(0)\frac{\psi(\vartheta)}{\vartheta^2} \sim \kappa + A_\varphi\sqrt{\vartheta - \vartheta^\star}; \quad A_\varphi := -\frac{\varphi'(0)}{(\vartheta^\star)^2}\sqrt{\frac{2}{v_\varphi}} < 0,$$

and hence, applying 'Heaviside', we estimate the tail distribution of the busy period by

$$p(t) \sim \frac{A_\varphi}{\Gamma(-\frac{1}{2})} \cdot \frac{e^{\vartheta^\star t}}{t\sqrt{t}}. \tag{6}$$

We now turn to the spectrally-negative case. Prop. 3.2 holds for any positive q, but we can consider the analytic continuation up to the branching point $q^\star < 0$ of $\Psi(\cdot)$. Let $\zeta > 0$ denote the minimizer of $\Phi(\cdot)$, so that $\Phi(\zeta) = q^\star < 0$. Similarly to the spectrally-negative case, we obtain, with $v_\Phi := \Phi''(\zeta) > 0$ and κ being some (irrelevant) number,

$$\int_0^\infty e^{-qt} p(t) \mathrm{d}t = \frac{1}{q}\left(1 - \frac{\Psi(0)}{\Psi(q)}\right) \sim \kappa + A_\Phi \sqrt{\vartheta - \vartheta^\star}; \quad A_\Phi := \frac{\Psi(0)}{q^\star \zeta^2}\sqrt{\frac{2}{v_\varphi}} < 0,$$

and hence 'Heaviside' estimates the tail of the busy-period distribution by

$$p(t) \sim \frac{A_\Phi}{\Gamma(-\frac{1}{2})} \cdot \frac{e^{q^\star t}}{t\sqrt{t}}. \tag{7}$$

3.3. **Importance-Sampling Based Simulation.** As $p(t)$ vanishes exponentially fast in the light-tailed case considered above, estimating $\mathbb{P}(\tau > t)$ from naive Monte Carlo simulation would be extremely time consuming. It is known that the number of replications needed (to obtain an estimate of a certain predefined precision) is roughly of the order $(p(t))^{-1}$. This motivates the search for more efficient simulation algorithms. We conclude this section by an algorithm for estimating this probability in an logarithmically efficient way; this algorithm is based on importance sampling, see e.g. [4, pp. 127-128], with an exponential twist of the Lévy process X_t.

We first explain what 'exponentially twisting' means in our Lévy setting; we focus here on the spectrally-positive case, but the spectrally-negative case works analogously. Evidently, the queue is stable under the probability measure $\mathbb{P}$, as we assumed $\mathbb{E}X_1 < 0$. Below we will propose a change of measure, with which we associate $\mathbb{Q}$, under which $\{\tau > t\}$ occurs with substantially higher probability, by application of an exponential twist $-\zeta > 0$ (where ζ was defined in Section 3.2). We have that the Laplace exponent $\varphi(\alpha)$ of X_t reads, with d, $\sigma^2 > 0$ and a measure $\Pi_\varphi(\cdot)$ such that $\int_{(0,\infty)} \min\{1, x^2\}\Pi_\varphi(\mathrm{d}x) < \infty$,

$$\varphi(\alpha) = -\alpha \cdot d + \frac{1}{2}\alpha^2\sigma^2 + \int_{(0,\infty)} (e^{-\alpha x} - 1 + \alpha x 1_{(0,1)})\Pi_\varphi(\mathrm{d}x).$$

It is now a matter of straightforward calculations to show that $\bar\varphi(\alpha) := \varphi(\alpha + \zeta) - \varphi(\zeta)$ is a Laplace exponent as well; let this be Laplace exponent of the Lévy process under $\mathbb{Q}$; it is readily checked that (in self-evident notation) $\mathbb{E}_\mathbb{Q} X_1 = -\bar\varphi'(0) = -\varphi'(\zeta) = 0$, so that the system under the new measure has drift 0. (One can check that under $\mathbb{Q}$ the drift d has increased to $d - \zeta\sigma^2$, the Brownian term remains unchanged, whereas the measure $\Pi_{\bar\varphi}(\mathrm{d}x)$ is given through its exponentially twisted counterpart (with 'twist' $-\zeta$).

In importance sampling one simulates under a different measure than the original one, where unbiasedness is recovered by weighing the simulation output by appropriate likelihood ratios. We propose the following alternative measure.

- Let, in the interval $(0,t]$, the Lévy process be twisted with $-\zeta = -\psi(\vartheta^\star) > 0$, as described above; $\vartheta^\star$ is as defined before.
- We in addition twist the workload at time 0, Q_0; we do so by a factor $\kappa \geq 0$, for which we identify a suitable value later on. This effectively means that we sample Q_0 from a distribution with Laplace transform $\mathbb{E}e^{-(\alpha-\kappa)Q_0}/\mathbb{E}e^{\kappa Q_0}$.

We denote from now on by $\mathbb{Q}_\kappa$ this new measure, consisting of twisting Q_0 (with κ) as well as a twisting $(X_s)_{s\in(0,t]}$ (with ζ).
In each run we simulate the process under $\mathbb{Q}_\kappa$ till time t, so that we can check whether $\tau > t$ or not. In this way, we perform n independent runs. Then the estimator, based on these n runs, reads $n^{-1}\sum_{i=1}^n L_i 1\{\tau_i > t\}$, where L_i is the likelihood ratio of run i. Let us write down this likelihood ratio more explicitly. First there is the contribution due to the twisted queue at time 0; using 'Pollaczek-Khinchine' we obtain

$$L_1 := e^{-\kappa Q_0}\cdot \mathbb{E}e^{\kappa Q_0} = e^{-\kappa Q_0}\cdot \frac{-\kappa\varphi'(0)}{\varphi(-\kappa)}.$$

Then there is the contribution due to the twisted Lévy process between 0 and t:

$$L_2 := e^{\psi(\vartheta^\star)X_t}\cdot \mathbb{E}e^{-\psi(\vartheta^\star)X_t} = e^{\psi(\vartheta^\star)X_t}\cdot e^{\vartheta^\star t}.$$

The 'total likelihood ratio' is thus $L := L_1 \times L_2$. It is standard that the resulting estimator is unbiased as $\mathbb{E}_{\mathbb{Q}_\kappa}$ equals the probability of our interest, i.e., $\mathbb{E}1$.
As $\mathbb{V}\mathrm{ar}_{\mathbb{Q}_\kappa} L1\{\tau > t\} \geq 0$, we see that $\mathbb{E}_{\mathbb{Q}_\kappa}L^2 1\{\tau > t\} \geq (\mathbb{E}_{\mathbb{Q}_\kappa}L1\{\tau > t\})^2$. In this sense, we could call our change of measure logarithmically efficient if

$$\lim_{t\to\infty}\frac{1}{t}\log \mathbb{E}_{\mathbb{Q}_\kappa}L^2 1\{\tau > t\} \leq \lim_{t\to\infty}\frac{1}{t}\log(\mathbb{E}_{\mathbb{Q}_\kappa}L1\{\tau > t\})^2 = 2\vartheta^\star.$$

Logarithmic efficiency essentially means that the number of replications needed to obtain an estimate with a certain fixed precision grows subexponentially in the 'rarity parameter' t, cf. [4, Ch. VI]. We now address the issue of appropriately choosing κ; we do this in three steps.

(i) $\kappa = 0$ *does not necessarily lead to logarithmic efficiency.* A first important observation is that not twisting Q_0 at all (i.e., choosing $\kappa = 0$) does *not* necessarily yield logarithmic efficiency: recalling that a necessary condition for $\{\tau > t\}$ is $\{Q_0 + X_t > 0\}$, we find

$$\mathbb{E}_{\mathbb{Q}_\kappa}L^2 1\{\tau > t\} \leq \left(-\frac{\kappa\varphi'(0)}{\varphi(-\kappa)}\right)^2 e^{2\vartheta^\star t}\mathbb{E}_{\mathbb{Q}_\kappa}e^{-2\kappa Q_0}e^{-2\psi(\vartheta^\star)Q_0}. \tag{8}$$

For logarithmic efficiency we should have that $\limsup_{t\to\infty} t^{-1}\log \mathbb{E}_{\mathbb{Q}_\kappa}L^2 1\{\tau > t\} \leq 2\vartheta^\star$. In other words, when picking $\kappa = 0$ we need to have $\mathbb{E}_{\mathbb{Q}_0}e^{-2\psi(\vartheta^\star)Q_0} < \infty$ for logarithmic efficiency, and this is not *a priori* clear.

(ii) $\kappa = -\zeta$ *leads to logarithmic efficiency.* But let us now check whether with another choice for κ logarithmic efficiency *can* be guaranteed. To this end, note that $\varphi(\psi(\vartheta^\star))$ is finite (to see this, use that ζ is larger than the pole of $\varphi(\cdot)$). Hence, picking $\kappa := -\psi(\vartheta^\star) = -\zeta$ *does* yield logarithmic efficiency! In other words: we have to exponentially twist Q_0 as well to obtain a provably logarithmically efficient procedure, and $\kappa = -\zeta > 0$ is a suitable choice.

(iii) $\kappa = -\zeta$ *is optimal.* The next question is: it is clear that for the $(X_s)_{s\in(0,t]}$-part, a twist by $-\zeta$ is optimal, but for the Q_0-part, can we do better than twisting with $-\zeta$? Interestingly, using

$$\mathbb{E}_{\mathbb{Q}_\kappa} e^{-\alpha Q_0} = \frac{\alpha-\kappa}{\varphi(\alpha-\kappa)} \cdot \frac{\varphi(-\kappa)}{-\kappa},$$

the right-hand side of (8) can be rewritten to

$$(\varphi'(0))^2 \left(\frac{-\kappa}{\varphi(-\kappa)}\right)\left(\frac{2\zeta+\kappa}{\varphi(2\zeta+\kappa)}\right) e^{2\vartheta^\star t}. \tag{9}$$

Observe that it consists of two factors that depend on κ, the first of which increases in κ, the second decreases in κ, so that there is a trade-off. It is a straightforward exercise to show that the minimum is achieved for $\kappa = -\zeta$ (this can be seen by equating the derivative to 0, but it also follows using an elementary symmetry-argument). We conclude that the proposed change of measure is the best possible within the class of exponential twists of Q_0, in the sense that it minimizes (9).

4. Simulation-based computation of the correlation function

As recalled in the previous section, if a probability tends to 0 as some 'rarity parameter' t grows large, then the number of runs needed to estimate the probability by naive simulation, for a given relative precision, is roughly inversely proportional to the probability. At the end of Section 2 we observed that the correlation $r(t)$ also tends to 0 as $t\to\infty$, which raises the question how many runs would be roughly needed to estimate $r(t)$ by naive simulation. We first answer this question, and then propose a coupling-based alternative that performs substantially better. This section concludes with a logarithmically efficient algorithm, that combines the coupling idea with importance sampling. In this section we concentrate on the spectrally-positive case; in the spectrally-negative case, the decay rates $\vartheta^\star$ must be replaced by $q^\star$ (while the proofs are very similar).

4.1. **Naive Simulation.** In the remainder of this section, we concentrate on estimating $\bar r(t) := \mathbb{C}\mathrm{ov}(Q_0, Q_t)$, as $v = \mathbb{V}\mathrm{ar}\, Q$ is known. The naive estimator of $\bar r(t)$ is, in self-evident notation, and recalling that $\mathbb{E}Q$ is known,

$$T_n^{(\mathrm{NS})}(t) := \frac{1}{n}\sum_{i=1}^n Q_0^{(i)}\, Q_t^{(i)} - (\mathbb{E}Q)^2,$$

based on n independent runs. The variance of this estimator reads $(n^{-1})\cdot \mathbb{V}\mathrm{ar}(Q_0Q_t)$. Now note that, as $t\to\infty$,

$$\mathbb{V}\mathrm{ar}(Q_0Q_t) = \mathbb{E}(Q_0^2Q_t^2) - (\mathbb{E}Q_0Q_t)^2 \to (\mathbb{E}Q^2)^2 - (\mathbb{E}Q)^4,$$

which is positive due to the fact that $\mathbb{E}Q^2 > (\mathbb{E}Q)^2$. Suppose our goal is to simulate until our estimate has a certain given relative precision ε (defined as the ratio between the width of the confidence interval and the estimate) and confidence α. The number of runs needed, say $n^{(\mathrm{NS})}(t)$, is roughly equal to the smallest n satisfying

$$2\delta_\alpha \frac{\sqrt{\mathbb{V}\mathrm{ar} T_n^{(\mathrm{NS})}(t)}}{r(t)} < \varepsilon,$$

for an appropriately chosen percentile of the standard Normal distribution δ_α. We obtain the following remarkable result for the naive estimator: it says that the number of runs required blows up exponentially, but it is *quadratically* inversely proportional to $r(t)$, rather than just inversely proportional. This result underscores that efficient (simulation-based) computation of the workload correlation $r(t)$ poses fundamentally new questions, despite the fact that its decay matches that of the busy-period asymptotics $p(t)$.

Proposition 4.1. $\lim_{t\to\infty} t^{-1} \cdot \log n^{(\mathrm{NS})}(t) = -2\vartheta^\star > 0$.

4.2. **A Coupling-based Algorithm.** In this subsection we develop a coupling-based simulation procedure that reduces the number of runs needed from quadratically inversely proportional to $\bar{r}(t)$, to just inversely proportional.
We write

$$\bar{r}(t) = \mathbb{E}(Q_0 \cdot (Q_t - Q_t^\star)),$$

where both Q and $Q^\star$ are stationary versions of the workload, and $Q_t^\star$ is *independent* of Q_0. We construct such a coupling as follows: generate Q_0 and $Q_0^\star$ independently, sampled from the stationary distribution of the workload. Now use exactly the same incoming Lévy process X_t over $(0,t]$ to drive both $(Q_s)_{s\in(0,t]}$ and $(Q_s^\star)_{s\in(0,t]}$ from their two independently generated initial conditions. This makes Q_t and Q_0 correlated but $Q_t^\star$ and Q_0 independent. The new estimator becomes, in self-evident notation,

$$T_n^{(\mathrm{CS})}(t) := \frac{1}{n}\sum_{i=1}^{n} Q_0^{(i)} \left(Q_t^{(i)} - Q_t^{\star\,(i)}\right),$$

based on n independent runs. The key observation is that $|Q_t - Q_t^\star| \le |Q_0 - Q_0^\star|$: the distance between both processes decreases in time. In particular, after the first epoch that *both* queues have been empty, the queueing processes coincide.
We split $\mathbb{E}(Q_0 \cdot (Q_t - Q_t^\star))$ into four terms, as follows. Recall that we defined $M_t := \inf_{0\le s\le t} X_s$. We write $\tau > t$ iff $Q_0 + M_t > 0$ (i.e., busy period has not ended at t) and $\tau^\star > t$ iff $Q_0^\star + M_t > 0$. Then $\bar{r}(t) = r_{++}(t) + r_{+-}(t) + r_{-+}(t) + r_{--}(t)$, where

$$\begin{aligned}
r_{++}(t) &:= \mathbb{E}(Q_0 \cdot (Q_t - Q_t^\star) \cdot 1\{\tau > t, \tau^\star > t\}),\\
r_{+-}(t) &:= \mathbb{E}(Q_0 \cdot (Q_t - Q_t^\star) \cdot 1\{\tau > t, \tau^\star \le t\}),\\
r_{-+}(t) &:= \mathbb{E}(Q_0 \cdot (Q_t - Q_t^\star) \cdot 1\{\tau \le t, \tau^\star > t\}),\\
r_{--}(t) &:= \mathbb{E}(Q_0 \cdot (Q_t - Q_t^\star) \cdot 1\{\tau \le t, \tau^\star \le t\}).
\end{aligned}$$

It is evident that $r_{--}(t) = 0$, as both queues have been empty and are identical from some time s (smaller than t) on. We estimate the other three terms separately. Due to $|Q_t - Q_t^\star| \le |Q_0 - Q_0^\star|$, we thus have that

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(Q_0 \cdot (Q_t - Q_t^\star)) &\le \mathbb{E}Q_0^2 \cdot (Q_t - Q_t^\star)^2\\
&\le \mathbb{E}(Q_0^2 \cdot (Q_0 - Q_0^\star)^2 \cdot 1\{\tau > t, \tau^\star > t\})\\
&\quad + \mathbb{E}(Q_0^2 \cdot (Q_0 - Q_0^\star)^2 \cdot 1\{\tau > t, \tau^\star \le t\})\\
&\quad + \mathbb{E}(Q_0^2 \cdot (Q_0 - Q_0^\star)^2 \cdot 1\{\tau \le t, \tau^\star > t\}).
\end{aligned}$$

With $m_k(t) := \mathbb{E}(Q_0^k 1\{\tau > t\})$, both the first and third term can be bounded from above by $\mathbb{E}(Q_0^4)\mathbb{P}(\tau > t) + \mathbb{E}(Q_0^2)m_2(t)$, whereas the second is majorized by $m_4(t) + \mathbb{E}(Q_0^2)m_2(t)$. The claim of Prop. 4.3 now follows directly from the following lemma (which is proven in the appendix). The number of runs needed, $n^{(\mathrm{CS})}(t)$, is defined analogously to $n^{(\mathrm{NS})}(t)$.

Lemma 4.2. For any $k \geq 0$, we have that $\limsup_{t\to\infty} t^{-1} \log m_k(t) \leq \vartheta^\star$.

Proposition 4.3. $\limsup_{t\to\infty} t^{-1} \cdot \log n^{(\mathrm{CS})}(t) \leq -\vartheta^\star$.

4.3. **Importance-Sampling Based Algorithm.** We now apply importance sampling on top of the coupling idea presented in the previous subsection. As we are dealing with the light-tailed case, an importance sampling measure $\mathbb{Q}$ is logarithmically efficient if

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbb{E}_{\mathbb{Q}}(L^2 Q_0^2 (Q_t - Q_t^\star)^2) \leq 2\vartheta^\star.$$

We again consider four scenarios by comparing τ and $\tau^\star$ with t; the idea is to estimate $r_{++}(t)$, $r_{+-}(t)$, and $r_{-+}(t)$ separately (recall that $r_{--}(t) = 0$).

- First focus on $r_{++}(t)$. We define

$$T_{n,++}^{(\mathrm{IS})}(t) := \frac{1}{n} \sum_{i=1}^{n} L_i^2 Q_0^{(i)} \left(Q_t^{(i)} - Q_t^{\star\,(i)}\right) 1\{\tau_i > t, \tau_i^\star > t\},$$

as an (unbiased) estimator of $r_{++}(t)$. Notice that in this case $Q_t - Q_t^\star = Q_0 - Q_0^\star$. Let, as in Section 3.3, the Lévy process on $(0,t]$ be twisted with $-\zeta = -\psi(\vartheta^\star) > 0$, with $\vartheta^\star$ as defined before. Also Q_0 is twisted by a factor κ and $Q_0^\star$ by a factor $\kappa^\star$, for which we identify suitable values below. In each run we simulate the process till time t. Let us write down the likelihood ratio at time t; we call the new measure $\mathbb{Q}_{\vec{\kappa}}$, with $\vec{\kappa}$ denoting the vector $(\kappa, \kappa^\star)$. We find that the likelihood equals

$$L = \left(e^{-\kappa Q_0} \cdot \frac{-\kappa\varphi'(0)}{\varphi(-\kappa)}\right) \times \left(e^{-\kappa^\star Q_0^\star} \cdot \frac{-\kappa^\star\varphi'(0)}{\varphi(-\kappa^\star)}\right) \times \left(e^{\zeta X_t} \cdot e^{\vartheta^\star t}\right).$$

We conclude that the second moment of the estimator reads

$$\mathbb{E}_{\mathbb{Q}_{\vec{\kappa}}} \left(L^2 Q_0^2 (Q_0 - Q_0^\star)^2 \cdot 1\{\tau > t, \tau^\star > t\}\right).$$

It is clear that $1\{\tau > t, \tau^\star > t\} \leq 1\{\tau > t\}$, and on $\{\tau > t\}$ we have that $-X_t < Q_0$. We thus find the upper bound

$$\begin{aligned}
&\mathbb{E}_{\mathbb{Q}_{\vec{\kappa}}} \left(\left(e^{-\kappa Q_0} \cdot \frac{-\kappa\varphi'(0)}{\varphi(-\kappa)}\right)^2 \left(e^{-\kappa^\star Q_0^\star} \cdot \frac{-\kappa^\star\varphi'(0)}{\varphi(-\kappa^\star)}\right)^2 \left(e^{-\zeta Q_0} \cdot e^{\vartheta^\star t}\right)^2 Q_0^2 (Q_0 - Q_0^\star)^2\right) \\
&\quad \leq \left(\frac{-\kappa\varphi'(0)}{\varphi(-\kappa)}\right)^2 \left(\frac{-\kappa^\star\varphi'(0)}{\varphi(-\kappa^\star)}\right)^2 e^{2\vartheta^\star t} \times \\
&\qquad \left(\mathbb{E}_{\mathbb{Q}_{\vec{\kappa}}} \left(Q_0^4 e^{-2(\kappa+\zeta)Q_0}\right) \mathbb{E}_{\mathbb{Q}_{\vec{\kappa}}} \left(e^{-2\kappa^\star Q_0}\right) + \right. \\
&\qquad \left. \mathbb{E}_{\mathbb{Q}_{\vec{\kappa}}} \left(Q_0^2 e^{-2(\kappa+\zeta)Q_0}\right) \mathbb{E}_{\mathbb{Q}_{\vec{\kappa}}} \left((Q_0^\star)^2 e^{-2\kappa^\star Q_0^\star}\right)\right).
\end{aligned}$$

Now we use our findings from Section 3.3. It is readily seen that the choice $\kappa = -\zeta$ and $\kappa^\star = 0$ yields logarithmic efficiency, as the above display reduces to a finite number

multiplied with $e^{2\vartheta^\star t}$. We here use, in the same way as in Section 3, that ζ is larger than the pole of $\varphi(\cdot)$, so that twisting with $-\zeta$ keeps all means finite, that is, $\mathbb{E}_{\mathbb{Q}_{\vec{\kappa}}} Q_0^4 < \infty$, $\mathbb{E}_{\mathbb{Q}_{\vec{\kappa}}} Q_0^2 < \infty$, and $\mathbb{E}_{\mathbb{Q}_{\vec{\kappa}}}((Q_0^\star)^2) = \mathbb{E} Q_0^2 < \infty$

- Now consider the second term: $r_{+-}(t)$. The estimator $T^{(\mathrm{IS})}_{n,+-}(t)$ is defined as $T^{(\mathrm{IS})}_{n,++}(t)$. Apparently $Q_0 > Q_0^\star$, and therefore also $Q_t \geq Q_t^\star$ for all $t \geq 0$. We also have $Q_t - Q_t^\star \leq Q_0 - Q_0^\star$ for all $t \geq 0$. With $1\{\tau > t, \tau^\star > t\} \leq 1\{\tau > t\}$, we can use the bounds above. We again obtain that $\kappa = -\zeta$ and $\kappa^\star = 0$ yields logarithmic efficiency.

- Finally, the case $r_{-+}(t)$ is essentially identical, but now we should pick $\kappa^\star = -\zeta$ and $\kappa = 0$.

As we can now estimate $r_{++}(t)$, $r_{+-}(t)$, and $r_{-+}(t)$ logarithmically efficiently, we arrive at the following result. Here $n^{(\mathrm{IS})}(t)$ denotes the number of runs needed to estimate $r(t)$ with a predefined precision, for a given confidence. The result states that the number of runs needed increases only subexponentially fast in the 'rarity parameter' t, and hence we have achieved a huge improvement over the naive scheme, and a still quite substantial improvement over the coupling-based algorithm (without importance sampling).

Theorem 4.4. $\lim_{t\to\infty} t^{-1} \cdot \log n^{(\mathrm{IS})}(t) = 0$.

5. Experimental results

In this section we discuss a number of implementation issues, and demonstrate the efficiency gain. We do this by considering two important special cases: reflected Brownian motion and the M/M/1 queue.

5.1. **Reflected Brownian motion.**

We consider standard Brownian motion with drift -1, such that $\varphi(\alpha) = \alpha + \frac{1}{2}\alpha^2$. We now provide some details regarding the implementation of the three simulation schemes.

- *Naive simulation.* It is readily checked that $\zeta = -1$. Remember that

$$Q_t = X_t + \max\left\{ -\inf_{0\leq s\leq t} X_s, Q_0 \right\}.$$

  It is a matter of straightforward verification that Q_0 is $\exp(2)$-distributed, i.e., has an exponential distrbution with mean $\frac{1}{2}$. Then we sample X_t from a normal distribution with mean $-t$ and variance t; say it has value z. Using known results for the Brownian Bridge, it is immediate that

$$\mathbb{P}\left(-\inf_{0\leq s\leq t} X_s \leq x \;\middle|\; X_t = z \right) = \exp\left(-2\frac{x}{t}(x+z) \right).$$

  Then it can be verified that

$$Y_z := \left(-\inf_{0\leq s\leq t} X_s \;\middle|\; X_t = z \right) \stackrel{\mathrm{d}}{=} -\frac{z}{2} + \frac{1}{2}\sqrt{z^2 - 2t\log U},$$

		Naive	*Coupling*	*IS*
$t = 10$	$7.91 \cdot 10^{-4}$	35%	0.85%	0.038%
$t = 12$	$2.21 \cdot 10^{-4}$	75%	1.50%	0.042%
$t = 14$	$6.75 \cdot 10^{-5}$	133%	2.82%	0.045%
$t = 16$	$2.17 \cdot 10^{-5}$	151%	4.99%	0.049%
$t = 18$	$6.83 \cdot 10^{-6}$	160%	8.4%	0.054%
$t = 20$	$2.27 \cdot 10^{-6}$	188%	11.9%	0.057%

TABLE 1. Numerical results, reflected Brownian motion.

where U has a uniform distribution over $(0, 1]$. The above observations enable easy simulation of Q_t, requiring just three random numbers, which can be sampled in a standard manner.

- *Coupling-based algorithm.* In this variant we sample Q_0 and $Q_0^\star$ independently of each other, both from an $\exp(2)$-distribution. In each simulation run, we simulate Q_t and $Q_t^\star$ by using the *same* samples for X_t and U.
- *Importance Sampling.* In the importance sampling variant, we let when simulating $r_{++}(t)$ and $r_{+-}(t)$ the initial workload $Q_0^\star$ be sampled from $\exp(2)$, and Q_0 from $\exp(1)$, leading to the likelihood ratio $L_1 := 2e^{-Q_0}$; when simulating $r_{-+}(t)$ we do this vice versa, resulting in $L_1 := 2e^{-Q_0^\star}$. Then we simulate X_t from a normal distribution with mean 0 and variance t. Supposing X_t has value z, we sample Y_z as explained above. This yields likelihood ratio

$$L_2 := e^{-X_t - t/2}.$$

  Then in each run the simulation output $Q_0(Q_t - Q_t^\star)$ needs to be multiplied with $L_1 L_2$.

Table 1 (in which per experiment 10^8 runs were performed) convincingly shows the enormous efficiency gain achieved, both when comparing the naive approach with the coupling approach, and when comparing the coupling approach with importance sampling. The second column of the table gives, for various values of t, the estimate of $r(t)$, obtained by the most efficient of the three methods, viz. importance sampling. Then the table gives, for the three methods, the *relative error*, i.e., the ratio of the width of the confidence interval (at a confidence level of 95%) and the estimate. Strikingly, under importance sampling the relative error is more or less constant, underscoring the superior performance of this method.

5.2. **M/M/1 queue.** We now take

$$\varphi(\alpha) = \alpha - \lambda + \frac{\lambda\mu}{\mu + \alpha},$$

i.e., arrivals occur according to a Poisson process with rate λ, and service times are $\exp(\mu)$. It is readily checked that $\zeta = -\mu + \sqrt{\lambda\mu}$. From

$$\mathbb{E}e^{\alpha Q_0} = (1 - \varrho) \Big/ \left(1 - \frac{\varrho\mu}{\mu - \alpha}\right) = (1 - \varrho) \sum_{n=0}^{\infty} \varrho^n \left(\frac{\mu}{\mu - \alpha}\right)^n,$$

		Naive	*Coupling*	*IS*
$t = 50$	$6.25 \cdot 10^{-3}$	18%	7.0%	0.53%
$t = 60$	$2.26 \cdot 10^{-3}$	41%	12.6%	0.52%
$t = 70$	$8.20 \cdot 10^{-4}$	65%	18.7%	0.54%
$t = 80$	$3.01 \cdot 10^{-4}$	76%	31.8%	0.59%
$t = 90$	$1.15 \cdot 10^{-4}$	87%	46.4%	0.61%
$t = 100$	$4.20 \cdot 10^{-5}$	101%	69.1%	0.62%

TABLE 2. Numerical results, M/M/1.

we retrieve the known fact that Q_0 is distributed as a Geometric number (with parameter $1 - \varrho$) of i.i.d. $\exp(\mu)$ random variables. Likewise,

$$\frac{\mathbb{E}e^{(\alpha-\zeta)Q_0}}{\mathbb{E}e^{-\zeta Q_0}} = (1-\sqrt{\varrho})\sum_{n=0}^{\infty}\sqrt{\varrho}^n\left(\frac{\sqrt{\lambda\mu}}{\sqrt{\lambda\mu}-\alpha}\right)^n.$$

We conclude that, in order to estimate $r_{++}(t)$ and $r_{+-}(t)$, Q_0 is, under the importance sampling measure, distributed as a Geometric number (with parameter $1 - \sqrt{\varrho}$) of i.i.d. $\exp(\sqrt{\lambda\mu})$ random variables; in order to estimate $r_{-+}(t)$, we let $Q_0^\star$ have this distribution. In this importance sampling, during the interval $(0, t]$ jobs arrive according to a Poisson process with rate $\sqrt{\lambda\mu}$, whereas their service times are i.i.d. samples from an $\exp(\sqrt{\lambda\mu})$ distribution.
In our experiments we chose $\mu = 1$, and $\lambda = \varrho = \frac{1}{2}$. Table 2 should be read as Table 1; the number of runs per experiment is now 10^7. The conclusions are very much in line with those of the Brownian case.

6. PRACTICAL ASPECTS AND DISCUSSION

Application of the simulation algorithms proposed in the previous sections, requires the ability to sample Lévy processes. Guidelines on this issue are presented in [4, Ch. XII].
In addition, one should be able to draw variates from exponentially twisted versions of the stationary workloads. In the spectrally-negative case this is straightforward, as Q_0 has an exponential distribution. In the spectrally-positive case, the Laplace transform of Q_0 is known (by 'Pollaczek-Khinchine'), and one could use methods as those described in [9] to generate samples. An alternative, only useful in the case of compound Poisson input, is to recognize that then the steady state workload is distributed as a geometric sum of residual job sizes, and hence so is its exponentially twisted version; in this situation one could also use the exact sampling technique proposed in [11].
Observe, however, that spectrally-positive light-tailed Lévy inputs are always just the sum of (i) Brownian motions, (ii) compound Poisson processes with light-tailed jobs, (iii) a negative drift. Restricting ourselves to *phase-type* jobs, it is readily seen from the generalized Pollaczek-Khinchine formula that also the steady-state workload is phase-type as well, and hence easy to generate variates from. In addition, the phase-type property is closed under exponential twisting, so it is straightforward to sample from this exponentially twisted workload.

In this paper we presented efficient algorithms for estimating the tail of the busy period $p(t)$ and the workload correlation function $r(t)$. In the spectrally one-sided cases Laplace transforms are known in closed-form, so the obvious alternative to simulation is to perform numerical inversion of these transforms. It should be noted, however, that the importance-sampling based simulation method can also be applied (and has good variance properties) if the driving Lévy process has both positive and negative jumps.
Potential subjects for future research are the following. (i) One could try to apply the coupling idea in settings in which the queue's input process does *not* have stationary independent increments. Can we for instance develop an algorithm of this kind for a queue fed by on-off sources with generally distributed on- and off times, or for queues with Gaussian input [16]? (ii) Is it possible to develop a simulation scheme with bounded relative error [4, p. 159]. Is it, perhaps for special cases such as reflected Brownian motion, possible to compute a zero-variance change of measure?

APPENDIX A. APPENDIX

We here present the proof of Lemma 4.2. Take $\varepsilon > 0$ arbitrary. Let m denote $-\mathbb{E}X_1 > 0$, and $m_\varepsilon := \lfloor m/\varepsilon \rfloor$. By splitting the interval $[0,\infty)$ into intervals of the form $[i\varepsilon t,(i+1)\varepsilon t)$, for $i = 0, 1, \ldots$, we obtain, using that $\mathbb{P}(\tau(x) > t)$ increases monotonically in x,

$$\begin{aligned} m_k(t) &= \int_0^\infty x^k \mathbb{P}(\tau(x) > t) \mathrm{d}\mathbb{P}(Q_0 \le x) \\ &\le \sum_{i=0}^{\infty} ((i+1)\varepsilon t)^k \mathbb{P}(\tau((i+1)\varepsilon t) > t) \mathbb{P}(Q_0 > i\varepsilon t) \\ &\le \sum_{i=0}^{m_\varepsilon} ((i+1)\varepsilon t)^k \mathbb{P}(\tau((i+1)\varepsilon t) > t) \mathbb{P}(Q_0 > i\varepsilon t) \\ &\quad + \sum_{i=m_\varepsilon+1}^{\infty} ((i+1)\varepsilon t)^k \mathbb{P}(Q_0 > i\varepsilon t). \end{aligned}$$

With $I(a) := \sup_\theta(\theta a - \log \mathbb{E}\exp(\theta X_1))$, the Chernoff bound immediately gives

$$\mathbb{P}(\tau(x) > t) \le \mathbb{P}(X(t) > -x) \le e^{-tI(-x/t)}$$

for all $x < mt$. In addition, [7, Remark 5.3] yields that $\mathbb{P}(Q_0 > x) \le \exp(-\xi x)$, where $\xi := \inf_{x>0} I(x)/x$. Hence, $m_k(t)$ is bounded from above by

$$\sum_{i=0}^{m_\varepsilon} h_i(t) + g(t),$$

where

$$h_i(t) := ((i+1)\varepsilon t)^k e^{-tI(-(i+1)\varepsilon)} e^{-\xi i\varepsilon t}, \;\; g(t) := \sum_{i=m_\varepsilon+1}^{\infty} ((i+1)\varepsilon t)^k e^{-\xi i\varepsilon t}.$$

It is readily checked that $\lim_{t\to\infty} t^{-1} \log h_i(t) = -I(-(i+1)\varepsilon) - \xi i\varepsilon$. Also

$$\int_a^\infty x^k e^{-xt} \mathrm{d}x \sim s(t) e^{-at},$$

for some subexponential function $s(\cdot)$ (as $t \to \infty$), which leads to

$$\lim_{t\to\infty} t^{-1} \log g(t) \leq \xi\varepsilon - (m_\varepsilon + 1)\xi\varepsilon.$$

Now [8, Lemma 1.2.15], stating that the decay rate of a finite sum equals the maximum of the decay rates, yields that

$$\limsup_{t\to\infty} \frac{1}{t} \log m_k(t) \leq \max\left\{ \max_{i=0,\ldots,m_\varepsilon} \{-I(-(i+1)\varepsilon) - \xi i\varepsilon\}, \xi\varepsilon - (m_\varepsilon+1)\xi\varepsilon \right\}.$$

Note that $k_i := -I(-(i+1)\varepsilon) - \xi i\varepsilon$ is concave in i, and hence $k_0 > k_1$ would imply that $\max_{i\in\{0,1,\ldots\}} k_i = k_0$. It is seen that $k_0 > k_1$ is equivalent to

$$\varepsilon^{-1} \cdot (I(-\varepsilon) - I(-2\varepsilon)) < \xi.$$

Observing that the convexity of $I(\cdot)$ implies that

$$\xi := \inf_{x>0} \frac{I(x)}{x} \geq \inf_{x>0} \frac{I(0) + xI'(0)}{x} > I'(0),$$

we have that for ε sufficiently small it indeed holds that $k_0 > k_1$, and hence

$$\limsup_{t\to\infty} t^{-1} \cdot \log m_k(t) \leq k_0 = -I(-\varepsilon).$$

Now letting $\varepsilon \to 0$, and realizing that $I(0) = -\vartheta^\star$, we have shown the stated. □.

REFERENCES

[1] J. Abate and W. Whitt. Transient behavior of the M/G/1 workload process. *Oper. Res.*, 42:750–764, 1994.

[2] J. Abate and W. Whitt. Asymptotics for M/G/1 low-priority waiting-time tail probabilities. *Queueing Syst.*, 25:173–233, 1997.

[3] S. Asmussen. *Applied Probability and Queues, 2nd ed.* Springer, New York, NY, USA, 2003.

[4] S. Asmussen and P. Glynn. *Stochastic Simulation – Algorithms and Analysis*. Springer, New York, NY, USA, 2007.

[5] V. Beneš. On queues with Poisson arrivals. *Ann. Math. Statist.*, 28:670–677, 1957.

[6] S.N. Bernstein. Sur les fonctions absolument monotones. *Acta Math.*, 52:1–66, 1929.

[7] K. Dębicki, A. Es-Saghouani, and M. Mandjes. Transient asymptotics of Lévy-driven queues. *J. Appl. Probab.*, 47:109–129, 2010.

[8] A. Dembo and O. Zeitouni. *Large deviations techniques and applications, 2nd edition*. Springer, New York, NY, USA, 1998.

[9] L. Devroye. *Non-uniform Random Variate Generation*. Springer, Berlin, Germany, 1986.

[10] R. Doney. Some excursion calculations for spectrally one-sided Lévy processes. *Séminaire de Probabilités XXXVIII, Lecture Notes in Math.*, 1857:5–15, 2005.

[11] K. Ensor and P. Glynn. Simulating the maximum of a random walk. *J. Stat. Plan. Inf.*, 85:127–135, 2000.

[12] A. Es-Saghouani and M. Mandjes. On the correlation structure of Lévy-driven queues. *J. Appl. Prob.*, 45:940–952, 2008.

[13] W. Feller. *An Introduction to Probability Theory and its Applications, 2nd ed.* Wiley, New York, NY, USA, 1971.

[14] O. Kella, O.J. Boxma, and M. Mandjes. A Lévy process reflected at a Poisson age process. *J. Appl. Prob.*, 43:221–230, 2006.

[15] A. Kyprianou. *Introductory Lectures on Fluctuations of Lévy Processes with Applications*. Springer, Berlin, Germany, 2006.

[16] M. Mandjes. *Large Deviations for Gaussian Queues*. Wiley, Chichester, UK, 2007.

[17] P. Morse. Stochastic properties of waiting lines. *Oper. Res.*, 3:255–262, 1955.

[18] T. Ott. The covariance function of the virtual waiting-time process in an M/G/1 queue. *Adv. Appl. Prob.*, 9:158–168, 1977.
[19] M. Pistorius. On exit and ergodicity of the completely asymmetric Lévy process reflected at its infimum. *J. Th. Prob.*, 17:183–220, 2004.
[20] J.F. Reynolds. The covariance structure of queues and related processes – a survey of recent work. *Adv. Appl. Prob.*, 7:383–415, 1975.
[21] D. Williams. *Probability with Martingales*. Cambridge University Press, Cambridge, UK, 1991.
[22] V. Zolotarev. The first passage time of a level and the behaviour at infinity for a class of processes with independent increments. *Th. Prob. Appl.*, 9:653–661, 1964.

DEPARTMENT OF MANAGEMENT SCIENCE & ENGINEERING, STANFORD UNIVERSITY, STANFORD, CA 94305, USA.
E-mail address: glynn@stanford.edu

KORTEWEG-DE VRIES INSTITUTE FOR MATHEMATICS, UNIVERSITY OF AMSTERDAM, SCIENCE PARK 904, 1098 XH AMSTERDAM, THE NETHERLANDS — EURANDOM, EINDHOVEN, THE NETHERLANDS — CWI, AMSTERDAM, THE NETHERLANDS.
E-mail address: m.r.h.mandjes@uva.nl

Centrum Wiskunde & Informatica (CWI) is the national research institute for mathematics and computer science in the Netherlands. The institute's strategy is to concentrate research on four broad, societally relevant themes: earth and life sciences, the data explosion, societal logistics and software as service.

Centrum Wiskunde & Informatica (CWI) is het nationale onderzoeksinstituut op het gebied van wiskunde en informatica. De strategie van het instituut concentreert zich op vier maatschappelijk relevante onderzoeksthema's: aard- en levenswetenschappen, de data-explosie, maatschappelijke logistiek en software als service.

Bezoekadres:
Science Park 123
Amsterdam

Postadres:
Postbus 94079, 1090 GB Amsterdam
Telefoon 020 592 93 33
Fax 020 592 41 99
info@cwi.nl
www.cwi.nl

CWI

Centrum Wiskunde & Informatica