J. Appl. Prob. **48,** 114–130 (2011)
Printed in England

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

PETER W. GLYNN,* *Stanford University*
MICHEL MANDJES,** *University of Amsterdam, EURANDOM and CWI*

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 $r(t) := \mathrm{cov}(Q_0, Q_t)/\mathrm{var}\, Q_0$ (assuming that the workload process is in stationarity at time 0), we first study its transform $\int_0^\infty r(t)\mathrm{e}^{-\vartheta t}\,\mathrm{d}t$, both for when the Lévy process has positive jumps and when 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 large t. 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 (the 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 process; reflection; workload process; correlation function; simulation; coupling; importance sampling

2010 Mathematics Subject Classification: Primary 60G51; 60K25

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* $r(t) := \mathrm{cov}(Q_0, Q_t)/\mathrm{var}\, Q_0$, assuming that the workload process is in stationarity at time 0. For several queueing systems, this correlation function has been explicitly computed; Morse [17], for instance, analyzed the number of customers in the M/M/1 queue. Often, explicit formulae are hard to obtain, but the analysis simplifies greatly when looking at the transform

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

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, Ott [18] elegantly proved that, in this case, $r(\cdot)$ is positive, decreasing, and convex. We further mention the survey by Reynolds [20], and interesting results by Abate and Whitt [1].

Received 27 May 2010; revision received 10 September 2010.
* Postal address: Department of Management Science & Engineering, Stanford University, Stanford, CA 94305, USA. Email address: glynn@stanford.edu
** Postal address: Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands. Email address: m.r.h.mandjes@uva.nl

The primary aim of this paper is to explore the workload correlation function for the class of single-server queues fed by *Lévy processes*. Note that the M/G/1 queue is contained 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 has already been shown in [12] that $r(\cdot)$ is positive, decreasing, and convex; our first contribution is that we use the results of [10] and [19] to show that these properties carry over to the spectrally negative case. We also estimate the asymptotics of $r(t)$ for large t. 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 $p(t) := \mathrm{P}(\tau > t)$. 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 that $r(t) \to 0$ (as $t \to \infty$) entails that naive simulation will be extremely time consuming; we show that it takes 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 the best of the authors' 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. In Section 6 we conclude, and discuss 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 $\mathrm{E}\, X_1 < 0$. We consider two cases.

Case 1: $(X_t)_{t\geq 0}$ *has no negative jumps.* Then the Laplace exponent is given by the function $\varphi(\cdot)\colon [0,\infty) \mapsto [0,\infty)$, i.e. $\varphi(\alpha) := \log \mathrm{E}\, \mathrm{e}^{-\alpha X_1}$. It is known that $\varphi(\cdot)$ is increasing and

https://doi.org/10.1239/jap/1300198139 Published online by Cambridge University Press

convex on $[0, \infty)$, with slope $\varphi'(0) = -\mathrm{E}\, X_1$ at 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 $\lim_{\alpha \to -\infty} \varphi(\alpha) = \infty$.

Case 2: $(X_t)_{t \geq 0}$ has no positive jumps. Now we define $\Phi(\beta) := \log \mathrm{E}\, \mathrm{e}^{\beta X_1}$, which is well defined for any $\beta \geq 0$. Again, ruling out that X_t is a subordinator (and recalling that $\Phi'(0) = \mathrm{E}\, X_1 < 0$), 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. (i) *Brownian motion with drift*, being actually both spectrally positive and negative. We write $X \in \mathbb{B}\mathrm{m}(\mu, \sigma^2)$ when $\varphi(\alpha) = -\alpha\mu + \frac{1}{2}\alpha^2\sigma^2$. (ii) *Compound Poisson with drift*, which is spectrally positive. Nonnegative jobs arrive according to a Poisson process with rate λ; the jobs $B_1, B_2, \ldots$ are independent and identically distributed (i.i.d.) samples from a distribution with Laplace transform $b(\alpha) := \mathrm{E}\, \mathrm{e}^{-\alpha B}$; 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 is positive, and the jobs are i.i.d. samples from a nonpositive 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, Chapter IX]. Define the decreasing process $(M_t)_{t \geq 0}$ and the resulting reflected process (or workload process or queueing process) $(Q_t)_{t \geq 0}$ by

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

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 $\mathrm{E}\, X_1 < 0$, is known (in terms of its Laplace transform) for both the spectrally positive and spectrally negative cases. For spectrally positive input, we have the *generalized Pollaczek–Khinchine formula*, usually attributed to Zolotarev [22]:

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

This result evidently enables the computation of all moments of the steady state queue Q (by repeated differentiation and inserting 0). From now on we assume that $\mathrm{E}\, Q^2$ is finite, so that $v := \mathrm{var}\, Q$ is well defined.

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

$$\mathrm{P}(\text{there exists } t \geq 0 \colon X_t > x)\mathrm{e}^{\beta_0 x} = 1,$$

and, as Q is distributed as the supremum over $t \geq 0$ of X_t ('Reich's identity'), we find that 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 have already been 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

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

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

$$\int_0^\infty \mathrm{e}^{-\beta x} W^{(q)}(x)\,\mathrm{d}x = \frac{1}{\Phi(\beta) - q}, \qquad \beta > \Psi(q). \tag{2.1}$$

In addition,

$$Z^{(q)}(x) := 1 + q\int_0^x W^{(q)}(y)\,\mathrm{d}y. \tag{2.2}$$

The functions $W^{(q)}(\cdot)$ and $Z^{(q)}(\cdot)$ are usually referred to as the q-scale functions. Then [19, Equation (19)] gives, with some abuse of notation, the following transform (with respect to t) of the density of Q_t, given that $Q_0 = x$:

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

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},

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

where

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

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 (2.1) and (2.2), we obtain

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

Likewise,

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

https://doi.org/10.1239/jap/1300198139 Published online by Cambridge University Press

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

- substituting $\alpha = 0$ into $I_1(\alpha, \beta, q) - I_2(\alpha, \beta, q)$ indeed yields $1/\beta$;
- substituting $\beta = \beta_0$ into the expression for $\int_0^\infty \beta e^{-\beta x} \, \mathrm{E}_x \, e^{-\alpha Q_T} \, dx$ 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},

$$\int_0^\infty q e^{-qt} \, \mathrm{E}(Q_0 Q_t) \, dt = \int_0^\infty \beta_0 x e^{-\beta_0 x} \, \mathrm{E}_x \, Q_T \, dx$$
$$= \lim_{\alpha \downarrow 0} \frac{d}{d\alpha} \left(\beta \frac{d}{d\beta} \int_0^\infty e^{-\beta x} \, \mathrm{E}_x \, e^{-\alpha Q_T} \, dx \Big|_{\beta=\beta_0} \right). \tag{2.3}$$

Upon combining the explicit expression for $I_1(\alpha, \beta, q) - I_2(\alpha, \beta, q)$ with (2.3), and recalling that $v = 1/\beta_0^2$ (in the spectrally negative case), we eventually obtain, after considerable calculus, the following result.

Theorem 2.1. *For the spectrally negative case,*

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

Corollary 2.1 below follows from applying l'Hôpital's rule twice. It implies that in the spectrally negative case the workload process is necessarily short-range dependent. Use the fact 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 2.1. *For the spectrally negative case,*

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

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

Theorem 2.2. *It holds that $r(\cdot)$ 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

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

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

$$\rho^{(2)}(q) := \int_0^\infty r''(t) e^{-qt} \, dt = -r'(0) + \beta_0^2 \Phi'(\beta_0) \left(\frac{1}{\Psi(q)} - \frac{1}{\beta_0} \right) = \beta_0^2 \frac{\Phi'(\beta_0)}{\Psi(q)}. \tag{2.4}$$

In the proof of Proposition 3.2 below we will show that $\Psi(0)/\Psi(q) \in \mathcal{C}$, where $\mathcal{C}$ is the class of completely monotone functions [6], [13, pp. 439ff.]; completely monotone functions are functions that can, up to some positive multiplicative constant, be considered as Laplace

https://doi.org/10.1239/jap/1300198139 Published online by Cambridge University Press

transforms of nonnegative random variables. We conclude from (2.4) that $\rho^{(2)}(q)$ is in $\mathcal{C}$, and, hence, $r''(\cdot)$ is positive, i.e. $r(\cdot)$ is convex.

We know that $f(q) \in \mathcal{C}$ implies that, with $g(q) := (f(0) - f(q))/q$, $g(q) \in \mathcal{C}$ also. Taking $f(q) = \rho^{(2)}(q)$, we find that $-\rho^{(1)}(q)$ is in $\mathcal{C}$, and, hence, $r'(\cdot)$ is negative, i.e. $r(\cdot)$ is decreasing. Applying the same procedure again, we find that $\rho(q)$ is in $\mathcal{C}$, and, hence, $r(\cdot)$ is positive. This completes the proof.

In [12] the asymptotics of $r(t)$ (for large t) 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 it was assumed 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 $t^{-3/2}\mathrm{e}^{\vartheta^\star t}$, 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 $q^\star$, $\Psi(q)$ looks like $\zeta + \sqrt{2/v_\Phi}\sqrt{q - q^\star}$, with $v_\Phi := \Phi''(\zeta) > 0$. Let $f(t) \sim g(t)$ (as $t \to A$) denote $f(t)/g(t) \to 1$ as $t \to A$. Then after some calculus we find that, for some (irrelevant) constant κ, as $q \downarrow q^\star$,

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

so that application of Heaviside heuristics [2] yields, as $t \to \infty$,

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

3. An interlude: 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 the steady state at time 0: $\tau := \inf\{t \geq 0\colon Q_t = 0\}$, where Q_0 is distributed according to the stationary distribution. Throughout this section, we write $p(t) := \mathrm{P}(\tau > t)$. 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 $\tau(x) := \inf\{t \geq 0\colon X_t = -x\}$,

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

Application of the Pollaczek–Khinchine formula now leads to the following result.

https://doi.org/10.1239/jap/1300198139 Published online by Cambridge University Press

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

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

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

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

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

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

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 the fact that $\Phi(\beta_0) = 0$, we find that $\mathrm{E}\,\mathrm{e}^{-q\tau} = \Psi(0)/\Psi(q)$, and, in addition, we obtain the following result.

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

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

3.2. Asymptotics

We again use the Heaviside operational principle [2] to (heuristically) estimate the decay of $p(t)$ for large t. 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 Proposition 3.1 then holds for any positive ϑ, but we can consider the analytic continuation up to the branching point $\vartheta^\star < 0$ of $\psi(\cdot)$. In the sequel, let $\zeta < 0$ denote the minimizer of $\varphi(\cdot)$, so that $\varphi(\zeta) = \vartheta^\star < 0$ (note that $v_\varphi := \varphi''(\zeta) > 0$). Then, for $\vartheta \downarrow \vartheta^\star$, we have $\psi(\vartheta) - \zeta \sim \sqrt{2/v_\varphi}\sqrt{\vartheta - \vartheta^\star}$. Hence, around $\vartheta^\star$, we have, for some (irrelevant) constant κ, as $\vartheta \downarrow \vartheta^\star$,

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

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

$$p(t) \sim \frac{A_\varphi}{\Gamma(-1/2)}\frac{\mathrm{e}^{\vartheta^\star t}}{t\sqrt{t}} \quad \text{as } t \to \infty.$$

We now turn to the spectrally negative case. Proposition 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, as $q \downarrow q^\star$,

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

https://doi.org/10.1239/jap/1300198139 Published online by Cambridge University Press

and, hence, the Heaviside operational principle estimates the tail of the busy period distribution to be

$$p(t) \sim \frac{A_\Phi}{\Gamma(-1/2)} \frac{\mathrm{e}^{q^\star t}}{t\sqrt{t}} \quad \text{as } t \to \infty.$$

3.3. Importance sampling based simulation

As $p(t)$ vanishes exponentially fast in the light-tailed case considered above, estimating $\mathrm{P}(\tau > t)$ 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 with an algorithm for estimating this probability in a 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 P, as we assumed that $\mathrm{E}\, X_1 < 0$. Below we will propose a change of measure, with which we associate 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 know that the Laplace exponent $\varphi(\alpha)$ of X_t reads, with $d, \sigma^2 > 0$ and a measure $\Pi_\varphi(\cdot)$ such that $\int_{(0,\infty)} \min\{1, x^2\}\Pi_\varphi(\mathrm{d}x) < \infty$,

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

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 the Laplace exponent of the Lévy process under Q. It is readily checked that (using self-evident notation) $\mathrm{E_Q}\, X_1 = -\bar{\varphi}'(0) = -\varphi'(\zeta) = 0$, so that the system under the new measure has drift 0. (We can check that, under Q, the drift d has increased to $d - \zeta\sigma^2$, and 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 we simulate under a measure different to the original measure, 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 above.
- 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 $\mathrm{E}\, \mathrm{e}^{-(\alpha-\kappa)Q_0}/\mathrm{E}\, \mathrm{e}^{\kappa Q_0}$.

From now on we denote this new measure by Q_κ, 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 Q_κ 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 $n^{-1}\sum_{i=1}^n L_i\, \mathbf{1}\{\tau_i > t\}$, 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 the Pollaczek–Khinchine formula, we obtain

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

https://doi.org/10.1239/jap/1300198139 Published online by Cambridge University Press

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

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

The 'total likelihood ratio' is thus $L := L_1 \times L_2$. It is standard that the resulting estimator is unbiased as $\mathrm{E}_{\mathrm{Q}_\kappa} L\,\mathbf{1}\{\tau > t\}$ equals the probability of our interest, i.e. $\mathrm{E}\,\mathbf{1}\{\tau > t\}$.

As $\mathrm{var}_{\mathrm{Q}_\kappa} L\,\mathbf{1}\{\tau > t\} \geq 0$, we see that $\mathrm{E}_{\mathrm{Q}_\kappa} L^2\,\mathbf{1}\{\tau > t\} \geq (\mathrm{E}_{\mathrm{Q}_\kappa} L\,\mathbf{1}\{\tau > t\})^2$. In this sense, we could call our change of measure logarithmically efficient if

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

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, Chapter VI]. We now address the issue of appropriately choosing κ; we do this in three steps.

Step (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 that

$$\mathrm{E}_{\mathrm{Q}_\kappa} L^2\,\mathbf{1}\{\tau > t\} \leq \left(-\frac{\kappa\varphi'(0)}{\varphi(-\kappa)}\right)^2 \mathrm{e}^{2\vartheta^\star t}\,\mathrm{E}_{\mathrm{Q}_\kappa}\,\mathrm{e}^{-2\kappa Q_0}\mathrm{e}^{-2\psi(\vartheta^\star)Q_0}. \tag{3.1}$$

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

Step (ii): $\kappa = -\zeta$ leads to logarithmic efficiency. 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 the fact 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.

Step (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

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

the right-hand side of (3.1) can be rewritten as

$$(\varphi'(0))^2 \frac{-\kappa}{\varphi(-\kappa)}\frac{2\zeta+\kappa}{\varphi(2\zeta+\kappa)}\mathrm{e}^{2\vartheta^\star t}. \tag{3.2}$$

Observe that it consists of two factors that depend on κ, the first of which increases in κ and the second of which 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 (3.2).

4. Simulation-based computation of the correlation function

As noted 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

https://doi.org/10.1239/jap/1300198139 Published online by Cambridge University Press

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 of 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, which 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 $\bar{r}(t) := \mathrm{cov}(Q_0, Q_t)$, as $v = \mathrm{var}\, Q$ is known. The naive estimator of $\bar{r}(t)$ is, using self-evident notation, and recalling that $\mathrm{E}\, Q$ is known,

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

based on n independent runs. The variance of this estimator reads $n^{-1}\, \mathrm{var}(Q_0 Q_t)$. Now note that, as $t \to \infty$,

$$\mathrm{var}(Q_0 Q_t) = \mathrm{E}(Q_0^2 Q_t^2) - (\mathrm{E}\, Q_0 Q_t)^2 \to (\mathrm{E}\, Q^2)^2 - (\mathrm{E}\, Q)^4,$$

which is positive due to the fact that $\mathrm{E}\, Q^2 > (\mathrm{E}\, Q)^2$. Suppose that 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

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

for an appropriately chosen percentile of the standard normal distribution δ_α. We obtain the following remarkable result for the naive estimator; it states 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. *It holds that* $\lim_{t\to\infty} t^{-1} \log n^{(\mathrm{NS})}(t) = -2\vartheta^\star > 0$.

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

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

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

https://doi.org/10.1239/jap/1300198139 Published online by Cambridge University Press

initial conditions. This makes Q_t and Q_0 correlated but $Q_t^\star$ and Q_0 independent. The new estimator becomes, using self-evident notation,

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

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 $\mathrm{E}(Q_0(Q_t - Q_t^\star))$ into four terms, as follows. Recall that we defined $M_t := \inf_{0\le s\le t} X_s$. We write $\tau > t$ if and only if $Q_0 + M_t > 0$ (i.e. the busy period has not ended at t) and $\tau^\star > t$ if and only if $Q_0^\star + M_t > 0$. Then $\bar{r}(t) = r_{++}(t) + r_{+-}(t) + r_{-+}(t) + r_{--}(t)$, where

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

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

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

With $m_k(t) := \mathrm{E}(Q_0^k\,\mathbf{1}\{\tau > t\})$, both the first and third terms can be bounded from above by $\mathrm{E}(Q_0^4)\,\mathrm{P}(\tau > t) + \mathrm{E}(Q_0^2)m_2(t)$, whereas the second term is majorized by $m_4(t) + \mathrm{E}(Q_0^2)m_2(t)$. The claim of Proposition 4.2 now follows directly from the following lemma (whose proof can be found in Appendix A). The number of runs needed, $n^{(\mathrm{CS})}(t)$, is defined analogously to $n^{(\mathrm{NS})}(t)$.

Lemma 4.1. *For any $k \ge 0$, we have* $\limsup_{t\to\infty} t^{-1}\log m_k(t) \le \vartheta^\star$.

Proposition 4.2. *It holds that* $\limsup_{t\to\infty} t^{-1}\cdot\log n^{(\mathrm{CS})}(t) \le -\vartheta^\star$.

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 Q is logarithmically efficient if

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

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 $r_{--}(t) = 0$).

https://doi.org/10.1239/jap/1300198139 Published online by Cambridge University Press

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

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

as an (unbiased) estimator of $r_{++}(t)$. Note 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 $-\zeta = -\psi(\vartheta^\star) > 0$, where $\vartheta^\star$ is 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 $\mathrm{Q}_{\vec{\kappa}}$, with $\vec{\kappa}$ denoting the vector $(\kappa,\kappa^\star)$. We find that the likelihood equals

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

We conclude that the second moment of the estimator reads

$$\mathrm{E}_{\mathrm{Q}_{\vec{\kappa}}}(L^2 Q_0^2(Q_0 - Q_0^\star)^2\,\mathbf{1}\{\tau > t,\ \tau^\star > t\}).$$

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

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

Now we use our findings from Section 3.3. It is readily seen that the choices $\kappa = -\zeta$ and $\kappa^\star = 0$ yield logarithmic efficiency, as the above display reduces to a finite number multiplied by $\mathrm{e}^{2\vartheta^\star t}$. Here we use, in the same way as in Section 3, the fact that ζ is larger than the pole of $\varphi(\cdot)$, so that twisting with $-\zeta$ keeps all means finite, that is, $\mathrm{E}_{\mathrm{Q}_{\vec{\kappa}}}\,Q_0^4 < \infty$, $\mathrm{E}_{\mathrm{Q}_{\vec{\kappa}}}\,Q_0^2 < \infty$, and $\mathrm{E}_{\mathrm{Q}_{\vec{\kappa}}}((Q_0^\star)^2) = \mathrm{E}\,Q_0^2 < \infty$.

We 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, $Q_t \ge Q_t^\star$ for all $t \ge 0$ also. We also have $Q_t - Q_t^\star \le Q_0 - Q_0^\star$ for all $t \ge 0$. With $\mathbf{1}\{\tau > t,\ \tau^\star > t\} \le \mathbf{1}\{\tau > t\}$, we can use the bounds above. We again find that $\kappa = -\zeta$ and $\kappa^\star = 0$ yield 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 4.1. *It holds that* $\lim_{t\to\infty} t^{-1}\cdot\log n^{(\mathrm{IS})}(t) = 0$.

https://doi.org/10.1239/jap/1300198139 Published online by Cambridge University Press

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

$$Q_t = X_t + \max\Big\{-\inf_{0\le s\le t} X_s,\, Q_0\Big\}.$$

It is a matter of straightforward verification that Q_0 is exp(2)-distributed, i.e. has an exponential distribution 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

$$\mathrm{P}\Big(-\inf_{0\le s\le t} X_s \le x \;\Big|\; X_t = z\Big) = \exp\Big(-2\frac{x}{t}(x+z)\Big).$$

Then it can be verified that

$$Y_z := \Big(-\inf_{0\le s\le t} X_s \;\Big|\; X_t = z\Big) \stackrel{\mathrm{D}}{=} -\frac{z}{2} + \frac{1}{2}\sqrt{z^2 - 2t\log U},$$

where U has a uniform distribution over $(0, 1]$ and '$\stackrel{\mathrm{D}}{=}$' denotes equality in distribution. 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, when simulating $r_{++}(t)$ and $r_{+-}(t)$, we let the initial workload $Q_0^\star$ be sampled from exp(2) and Q_0 be sampled from exp(1), leading to the likelihood ratio $L_1 := 2\mathrm{e}^{-Q_0}$; when simulating $r_{-+}(t)$, we do this vice versa, resulting in $L_1 := 2\mathrm{e}^{-Q_0^\star}$. Then we simulate X_t from a normal distribution with mean 0 and variance t. Supposing that X_t has value z, we sample Y_z as explained above. This yields the likelihood ratio

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

Then in each run the simulation output $Q_0(Q_t - Q_t^\star)$ needs to be multiplied by L_1L_2.

Table 1 (in which 10^8 runs were performed per experiment) 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.

https://doi.org/10.1239/jap/1300198139 Published online by Cambridge University Press

TABLE 1: Numerical results for reflected Brownian motion.

t	$r(t)$	Relative error (%)		
		Naive approach	Coupling approach	Importance sampling approach
10	7.91×10^{-4}	35	0.85	0.038
12	2.21×10^{-4}	75	1.50	0.042
14	6.75×10^{-5}	133	2.82	0.045
16	2.17×10^{-5}	151	4.99	0.049
18	6.83×10^{-6}	160	8.4	0.054
20	2.27×10^{-6}	188	11.9	0.057

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

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

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,

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

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, except that we now present the results for the covariances $\bar{r}(\cdot)$ rather than the correlations $r(\cdot)$;

TABLE 2: Numerical results for the M/M/1 queue.

t	$\bar{r}(t)$	Relative error (%)		
		Naive approach	Coupling approach	Importance sampling approach
50	6.25×10^{-3}	18	7.0	0.53
60	2.26×10^{-3}	41	12.6	0.52
70	8.20×10^{-4}	65	18.7	0.54
80	3.01×10^{-4}	76	31.8	0.59
90	1.15×10^{-4}	87	46.4	0.61
100	4.20×10^{-5}	101	69.1	0.62

https://doi.org/10.1239/jap/1300198139 Published online by Cambridge University Press

here the number of runs per experiment is 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, Chapter XII].

In addition, we 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 the Pollaczek–Khinchine formula), and we could use those methods described in [9] to generate samples. An alternative, but which is 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 we 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; and (iii) a negative drift. Restricting ourselves to *phase-type* jobs, it is readily seen from the generalized Pollaczek–Khinchine formula that 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) We could try to apply the coupling idea to 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. Proof of Lemma 4.1

In this appendix we present the proof of Lemma 4.1. Take $\varepsilon > 0$ arbitrary. Let m denote $-\mathrm{E}\, X_1 > 0$ 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 the fact that $\mathrm{P}(\tau(x) > t)$ increases monotonically in x,

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

https://doi.org/10.1239/jap/1300198139 Published online by Cambridge University Press

$$\leq \sum_{i=0}^{m_\varepsilon} ((i+1)\varepsilon t)^k \, \mathrm{P}(\tau((i+1)\varepsilon t) > t) \, \mathrm{P}(Q_0 > i\varepsilon t)$$
$$+ \sum_{i=m_\varepsilon+1}^{\infty} ((i+1)\varepsilon t)^k \, \mathrm{P}(Q_0 > i\varepsilon t).$$

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

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

for all $x < mt$. In addition, [7, Remark 5.3] yields $\mathrm{P}(Q_0 > x) \leq \mathrm{e}^{-\xi x}$, 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

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

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

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

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

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

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

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

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 find that, for sufficiently small ε, 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 $I(0) = -\vartheta^\star$, completes the proof.

https://doi.org/10.1239/jap/1300198139 Published online by Cambridge University Press

Acknowledgements

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.

References

[1] Abate, J. and Whitt, W. (1994). Transient behavior of the M/G/1 workload process. *Operat. Res.* **42,** 750–764.
[2] Abate, J. and Whitt, W. (1997). Asymptotics for M/G/1 low-priority waiting-time tail probabilities. *Queueing Systems* **25,** 173–233.
[3] Asmussen, S. (2003). *Applied Probability and Queues*, 2nd edn. Springer, New York.
[4] Asmussen, S. and Glynn, P. W. (2007). *Stochastic Simulation: Algorithms and Analysis*. Springer, New York.
[5] Beneš, V. E. (1957). On queues with Poisson arrivals. *Ann. Math. Statist.* **28,** 670–677.
[6] Bernstein, S. N. (1929). Sur les fonctions absolument monotones. *Acta Math.* **52,** 1–66.
[7] Dębicki, K., Es-Saghouani, A. and Mandjes, M. (2010). Transient asymptotics of Lévy-driven queues. *J. Appl. Prob.* **47,** 109–129.
[8] Dembo, A. and Zeitouni, O. (1998). *Large Deviations Techniques and Applications*, 2nd edn. Springer, New York.
[9] Devroye, L. (1986). *Nonuniform Random Variate Generation*. Springer, Berlin.
[10] Doney, R. A. (2005). Some excursion calculations for spectrally one-sided Lévy processes. In *Séminaire de Probabilités XXXVIII* (Lecture Notes Math. **1857**), Springer, Berlin, pp. 5–15.
[11] Ensor, K. and Glynn, P. (2000). Simulating the maximum of a random walk. *J. Statist. Planning Infer.* **85,** 127–135.
[12] Es-Saghouani, A. and Mandjes, M. (2008). On the correlation structure of a Lévy-driven queue. *J. Appl. Prob.* **45,** 940–952.
[13] Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, Vol. II, 2nd edn. John Wiley, New York.
[14] Kella, O., Boxma, O. J. and Mandjes, M. (2006). A Lévy process reflected at a Poisson age process. *J. Appl. Prob.* **43,** 221–230.
[15] Kyprianou, A. E. (2006). *Introductory Lectures on Fluctuations of Lévy Processes with Applications*. Springer, Berlin.
[16] Mandjes, M. (2007). *Large Deviations for Gaussian Queues*. John Wiley, Chichester.
[17] Morse, P. M. (1955). Stochastic properties of waiting lines. *Operat. Res.* **3,** 255–261.
[18] Ott, T. J. (1977). The covariance function of the virtual waiting-time process in an M/G/1 queue. *Adv. Appl. Prob.* **9,** 158–168.
[19] Pistorius, M. R. (2004). On exit and ergodicity of the spectrally one-sided Lévy process reflected at its infimum. *J. Theoret. Prob.* **17,** 183–220.
[20] Reynolds, J. F. (1975). The covariance structure of queues and related processes—a survey of recent work. *Adv. Appl. Prob.* **7,** 383–415.
[21] Williams, D. (1991). *Probability with Martingales*. Cambridge University Press.
[22] Zolotarev, V. M. (1964). The first passage time of a level and the behaviour at infinity for a class of processes with independent increments. *Theoret. Prob. Appl.* **9,** 653–661.