An Infinite-Server System with Lévy Shot-Noise Modulation

Moments and Asymptotics

Saxena, M.; Boxma, O.J.; Mandjes, M.

Publication date
2020

Document Version
Submitted manuscript

Published in
Markov Processes and Related Fields

EURANDOM PREPRINT SERIES

2019-003

April 1, 2019

An infinite-server system with Lévy shot-noise modulation: moments and asymptotics

M. Saxena, O. Boxma, M. Mandjes
ISSN 1389-2355

An infinite-server system with Lévy shot-noise modulation: moments and asymptotics

M. Saxena*[a], O.J. Boxma†[a], and M. Mandjes‡[b]

[a] *Department of Mathematics & Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600MB, Eindhoven.*

[b] *Korteweg-de Vries Institute, University of Amsterdam, P.O. Box 94248, 1090GE, Amsterdam.*

April 1, 2019

Abstract

We consider an infinite-server system with as input process a non-homogeneous Poisson process with rate function $\Lambda(t) = \boldsymbol{a}^{\intercal}\boldsymbol{X}(t)$. Here $\{\boldsymbol{X}(t) : t \geq 0\}$ is a generalized multivariate shot-noise process fed by a Lévy subordinator rather than by just a compound Poisson process. We study the transient behavior of the model, analyzing the joint distribution of the number of customers in the queueing system jointly with the multivariate shot-noise process. We also provide a recursive procedure that explicitly identifies transient as well as stationary moments and correlations. Various heavy-tail and heavy-traffic asymptotic results are also derived, and numerical results are presented to provide further insight into the model behavior.

Keywords. Infinite-server queue, Non-homogeneous Poisson process, Lévy subordinator, Modulated shot-noise process.

1 Introduction

In most queueing models the arrival process of customers is assumed to be a homogeneous Poisson process. However, in many real-world examples, the arrival process is non-homogeneous, and the arrival rate of the Poisson process should even be viewed as a stochastic process. The latter arrival process is called a Cox process [12] or doubly stochastic Poisson process. An interesting example is provided by a popular website, where the arrival process of visitors is a Poisson process whose rate may jump up due to a viral event, decay gradually and jump up again at another event. An important characteristic of this type of processes is that their dispersion index, the ratio of the variance to the mean, is greater than one; for the ordinary Poisson process, this ratio is one. Cox processes are well-known within and outside of the

*m.mayank@tue.nl

†o.j.boxma@tue.nl

‡m.r.h.mandjes@uva.nl

queueing community, as they arise naturally in many applications and are typically amenable to exact analysis. In this paper, we consider a Cox arrival process, in which the arrival rate is a weighted sum of d, possibly correlated, shot-noise processes. This arrival process forms the input into an infinite-server queue. There is a rich literature on the M/G/∞ system, see e.g. [21]. However, the Cox/G/∞ queue is much less studied. Our objective in this study is to obtain the joint distribution of the arrival rate and the number of customers in this Cox/G/∞ system.

◦ **Related literature.** Our work partly builds on the recent paper [6], which focuses on the same Cox/G/∞ system. The authors of [6] develop a general framework for the stationary study of infinite-server queues whose input is a Multivariate Cox process. [6] generalizes [17], which studies a Cox/G/∞ system in which the arrival rate is a shot-noise process, where the jumps of the shot-noise process occur according to a homogeneous Poisson process. A similar model is studied by Daw and Pender [8], who perform an exact analysis of an infinite-server queueing system in which the arrivals are driven by a self-exciting Hawkes process and where service follows a phase-type distribution or is deterministic. Such a Hawkes/G/∞ system is also analyzed in [18]. By viewing the Hawkes process as a branching process, it is shown in [18] that the probability generating function of the number of customers in the system can be expressed in terms of the solution of a fixed-point equation. Various asymptotic results are also obtained. Apart from these papers, related work can also be found in [2, 4, 13, 15].

◦ **Main contributions.** We are interested in the performance of an infinite-server queue whose input is a non-homogeneous Poisson process, which is generated by the modulation of a generalized shot-noise process. The rate function of the generalized shot-noise process is defined with respect to a Lévy subordinator. One contribution is the development of the framework of [6] in a more queueing theoretic way. Our second contribution is the derivation of the transient joint transform of the number of customers jointly with the multivariate shot-noise rate function. This transform allows us to obtain the joint transform of the number of customers and the arrival rate. Thirdly, we develop a recursive procedure that explicitly identifies any transient as well as stationary moments. Finally, we also derive various asymptotic results. In particular, we obtain the asymptotics of the queue length process, under assumptions regarding the tail behavior of the shot-noise process. Subsequently, we derive a central limit theorem for the vector of number of customers and arrival rate. We also briefly discuss a functional central limit theorem for the number of customers in the system.

◦ **Organization of the paper.** In Section 2 we describe the model and discuss some preliminary results, which are used in the later sections to obtain the main results of the paper. In Section 3 we derive the transient joint transform of the number of customers and the input rate function. In Section 4 we derive a recursive scheme, which allows us to calculate any mixed transient moments. Section 5 considers the asymptotics of the queue length process, under assumptions regarding the multi-dimensional tail behavior of the shot-noise process. Additionally, a functional central limit theorem for the number of customers is discussed. In Section

6 we present several numerical results for the means, variances and correlation coefficients of the arrival rate and the number of customers, for various parameter settings. Section 7 contains conclusions and suggestions for further research.

2 Model description and preliminary results

∘ **Description.** In this paper we consider an infinite-server queue. We first detail its input process, and then describe the queueing dynamics.

The input process is of Coxian type, i.e., the number of arrivals in $[0, t)$ has a Poisson distribution with a random rate $\Lambda(t)$. To describe how the rate process $\Lambda(\cdot)$ is constructed, we first define the process $\boldsymbol{X}(\cdot)$, as follows. Let $\boldsymbol{J}(\cdot)$ be a d-dimensional subordinator, i.e., a d-dimensional Lévy process which is non-decreasing in all components. For $\boldsymbol{z} \in \mathbb{R}^d_{\geqslant 0}$ we define the Laplace exponent $\eta(\cdot)$ of $J(\cdot)$:

$$\eta(\boldsymbol{z}) := -\log \mathbb{E}[\mathrm{e}^{-\boldsymbol{z}^\intercal \boldsymbol{J}(1)}] = \boldsymbol{c}^\intercal \boldsymbol{z} + \int_0^\infty (1 - \mathrm{e}^{-\boldsymbol{x}^\intercal \boldsymbol{z}})\nu(\mathrm{d}\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{c} \in \mathbb{R}^d_{\geqslant 0}$ and ν is an associated Lévy measure satisfying

$$\nu\left(\left(\mathbb{R}^d_+\right)^\mathsf{c} \cup \{0\}\right) = 0 \quad \text{and} \quad \int_{\mathbb{R}^d_+} \left(\|\boldsymbol{x}\| \wedge 1\right) \nu(\mathrm{d}\boldsymbol{x}) < \infty. \tag{2}$$

Let the matrix $Q = \left(q_{ij}\right)$ be a $(d \times d)$-matrix with non-negative diagonal and non-positive off-diagonal elements, and with all eigenvalues having strictly positive real parts; for more detail see [6]. Now fix the initial state $\boldsymbol{X}(0) = \boldsymbol{x}$, componentwise strictly positive. We then define the process $\boldsymbol{X}(\cdot)$ through

$$\boldsymbol{X}(t) = \mathrm{e}^{-Qt}\boldsymbol{x} + \int_0^t \mathrm{e}^{-Q(t-s)}\mathrm{d}\boldsymbol{J}(s). \tag{3}$$

We thus find that $\boldsymbol{X}(\cdot)$ is the unique solution to the stochastic integral equation

$$\boldsymbol{X}(t) = \boldsymbol{x} + \boldsymbol{J}(t) - Q\int_0^t \boldsymbol{X}(s)\mathrm{d}s. \tag{4}$$

The input rate process, corresponding to our Coxian arrival process, is

$$\Lambda(t) := \boldsymbol{a}^\intercal \boldsymbol{X}(t), \tag{5}$$

where $\boldsymbol{a} \in \mathbb{R}^d_{\geqslant 0}$, and $\intercal$ denotes the transpose of a vector or matrix.

Now that we have defined the arrival process, we can introduce the queueing model under study. This system is of infinite-server type, meaning that all jobs present are served simultaneously (and obviously leave the system after service completion). It is throughout assumed that the service times $G_1, G_2, \ldots$ are independent and identically distributed non-negative random variables, that are in addition independent of $\boldsymbol{J}(\cdot)$; here G denotes a generic random variable with the same distribution as G_1. We let $L(t)$ be the number of customers present at time t.

◦ **Preliminary results.** For the considered model, our first objective is to derive the joint transform of $L(t)$ and $\Lambda(t)$, for any $t \geq 0$. For this purpose, we start with the following lemma, which will be used in the next section.

Lemma 2.1. *Let $L(t)$ be the number of customers in the system at time t given that $L(0) = 0$. Then*

$$\mathbb{E}[w^{L(t)}] = \exp\left(\int_0^t (w-1)\mathbb{P}(G > t-u)\Lambda(u)\mathrm{d}u\right)$$
$$= \exp\left(\int_0^t (w-1)\mathbb{P}(G > t-u)\boldsymbol{a}^\intercal \boldsymbol{X}(u)\mathrm{d}u\right). \tag{6}$$

Proof. Using the argument given [17, Section 3.1], conditional on the path of the rate process $\Lambda(\cdot)$ between 0 and t, it immediately follows that the number of customers $L(t)$ in the system at time t is Poisson distributed with parameter $\int_0^t \mathbb{P}(G > t-u)\Lambda(u)\mathrm{d}u$. □

3 Infinite-server system with Lévy shot-noise modulation

In this section we aim at obtaining the joint transform of $L(t)$ and $\Lambda(t)$ at time $t \geqslant 0$. To this end, we first derive the joint transform of $L(t)$ and $\boldsymbol{X}(t)$.

◦ **Derivation of the joint transform of $L(t)$ and $\boldsymbol{X}(t)$.** Following the line of reasoning of the proof of Lemma 2.1,

$$\begin{aligned}
&\mathbb{E}[w^{L(t)}\mathrm{e}^{-\boldsymbol{b}^\intercal \boldsymbol{X}(t)}|\boldsymbol{X}(0)=\boldsymbol{x}] \\
&= \mathbb{E}\left[\exp\left(\int_0^t (w-1)\mathbb{P}(G>t-u)\Lambda(u)\mathrm{d}u - \boldsymbol{b}^\intercal \boldsymbol{X}(t)\right)\Big|\boldsymbol{X}(0)=\boldsymbol{x}\right] \\
&= \mathbb{E}\left[\exp\left(\int_0^t (w-1)\mathbb{P}(G>t-u)\boldsymbol{a}^\intercal \boldsymbol{X}(u)\mathrm{d}u - \boldsymbol{b}^\intercal \boldsymbol{X}(t)\right)\Big|\boldsymbol{X}(0)=\boldsymbol{x}\right].
\end{aligned} \tag{7}$$

Let $\phi_w(u) := (w-1)\mathbb{P}(G > u)$, and consider the exponent in (7). It follows from (3) that

$$\begin{aligned}
&\int_{u=0}^t \phi_w(t-u)\boldsymbol{a}^\intercal \boldsymbol{X}(u)\mathrm{d}u - \boldsymbol{b}^\intercal \boldsymbol{X}(t) \\
&\quad = \int_{u=0}^t \phi_w(t-u)\left(\boldsymbol{a}^\intercal \mathrm{e}^{-Qu}\boldsymbol{X}(0) + \boldsymbol{a}^\intercal \int_{s=0}^u \mathrm{e}^{-Q(u-s)}\mathrm{d}\boldsymbol{J}(s)\right)\mathrm{d}u \\
&\quad\quad - \boldsymbol{b}^\intercal\left(\mathrm{e}^{-Qt}\boldsymbol{X}(0) + \int_{s=0}^t \mathrm{e}^{-Q(t-s)}\mathrm{d}\boldsymbol{J}(s)\right) \\
&\quad = -\left(\boldsymbol{b}^\intercal \mathrm{e}^{-Qt} - \boldsymbol{a}^\intercal \int_{u=0}^t \mathrm{e}^{-Qu}\phi_w(t-u)\mathrm{d}u\right)\boldsymbol{X}(0) \\
&\quad\quad - \int_{s=0}^t \left(\boldsymbol{b}^\intercal - \boldsymbol{a}^\intercal \int_{u=s}^t \mathrm{e}^{-Q(u-t)}\phi_w(t-u)\mathrm{d}u\right)\mathrm{e}^{-Q(t-s)}\mathrm{d}\boldsymbol{J}(s).
\end{aligned} \tag{8}$$

Combining (7) and (8) and conditioning on $\boldsymbol{X}(0) = \boldsymbol{x}$, we obtain

$$\mathbb{E}[w^{L(t)}\mathrm{e}^{-\boldsymbol{b}^\intercal \boldsymbol{X}(t)}|\boldsymbol{X}(0)=\boldsymbol{x}]$$

$$= \mathbb{E}\Big[\exp\Big\{-\Big(\boldsymbol{b}^{\mathsf{T}}\mathrm{e}^{-Qt} - \boldsymbol{a}^{\mathsf{T}}\int_{u=0}^{t}\mathrm{e}^{-Qu}\phi_w(t-u)\mathrm{d}u\Big)\boldsymbol{x}$$

$$-\int_{s=0}^{t}\Big(\boldsymbol{b}^{\mathsf{T}} - \boldsymbol{a}^{\mathsf{T}}\int_{u=s}^{t}\mathrm{e}^{-Q(u-t)}\phi_w(t-u)\mathrm{d}u\Big)\mathrm{e}^{-Q(t-s)}\mathrm{d}\boldsymbol{J}(s)\Big\}\Big]. \tag{9}$$

Recall that $\eta(\cdot)$ is the Laplace exponent of the subordinator $\boldsymbol{J}(\cdot)$ defined by (1). The following fact for a subordinator, cf. [10, Eqn. (4)], is useful for our purpose:

$$\mathbb{E}\Big[\exp\Big(-\int_0^\infty \boldsymbol{r}^{\mathsf{T}}(s)\mathrm{d}\boldsymbol{J}(s)\Big)\Big] = \exp\Big(-\int_0^\infty \eta(\boldsymbol{r}(s))\mathrm{d}s\Big), \tag{10}$$

for a vector-valued function $\boldsymbol{r}(\cdot)$. We define, for a given $t \geqslant 0$, the row vector

$$\boldsymbol{r}^{\mathsf{T}}(t-s) := \Big(\boldsymbol{b}^{\mathsf{T}} - \boldsymbol{a}^{\mathsf{T}}\int_{u=s}^{t}\mathrm{e}^{-Q(u-t)}\phi_w(t-u)\mathrm{d}u\Big)\mathrm{e}^{-Q(t-s)}$$

$$= \boldsymbol{b}^{\mathsf{T}}\mathrm{e}^{-Q(t-s)} - (w-1)\boldsymbol{a}^{\mathsf{T}}\int_{g=0}^{t-s}\mathrm{e}^{-Q(t-s-g)}\mathbb{P}(G>g)\mathrm{d}g, \tag{11}$$

if $s\in[0,t]$, and 0 if $s>t$. It now follows from (9)-(11) that

$$\mathbb{E}[w^{L(t)}\mathrm{e}^{-\boldsymbol{b}^{\mathsf{T}}\boldsymbol{X}(t)}|\boldsymbol{X}(0)=\boldsymbol{x}] = \exp\Big\{-\Big(\boldsymbol{b}^{\mathsf{T}}\mathrm{e}^{-Qt} - (w-1)\boldsymbol{a}^{\mathsf{T}}\int_{g=0}^{t}\mathrm{e}^{-Q(t-g)}\mathbb{P}(G>g)\mathrm{d}g\Big)\boldsymbol{x}$$

$$-\int_{s=0}^{t}\eta\Big(\Big(\boldsymbol{b}^{\mathsf{T}}\mathrm{e}^{-Qs} - (w-1)\boldsymbol{a}^{\mathsf{T}}\int_{g=0}^{s}\mathrm{e}^{-Q(s-g)}\mathbb{P}(G>g)\,\mathrm{d}g\Big)^{\mathsf{T}}\Big)\mathrm{d}s\Big\}. \tag{12}$$

Note that (12) is a generalization of [17, Lemma 2.1]. In [17] the driving Lévy process was compound Poisson, whereas here it is a subordinator. In the compound Poisson case $\eta(v) = \lambda(1-\beta(v))$, with λ the arrival rate of jumps, the sizes of which have Laplace-Stieltjes transform $\beta(\cdot)$.

Recalling the concept of the *residual* of a random variable, we have for the residual G^{res} of G:

$$\int_{g=0}^{y}\mathbb{P}(G>g)\,\mathrm{d}g = \mathbb{E}[G]\,\mathbb{P}(G^{\mathrm{res}}\leqslant y), \tag{13}$$

so that

$$\mathbb{E}[G]\ \mathbb{E}[\mathrm{e}^{-Q(y-G^{\mathrm{res}})}\ 1_{\{G^{\mathrm{res}}\leq y\}}] = \int_{g=0}^{y}\mathrm{e}^{-Q(y-g)}\mathbb{P}(G>g)\mathrm{d}g. \tag{14}$$

Upon combining the above results, we arrive at

$$\mathbb{E}[w^{L(t)}\mathrm{e}^{-\boldsymbol{b}^{\mathsf{T}}\boldsymbol{X}(t)}|\boldsymbol{X}(0)=\boldsymbol{x}] = \exp\Big\{-\boldsymbol{r}^{\mathsf{T}}(t)\boldsymbol{x} - \int_{s=0}^{t}\eta\left(\boldsymbol{r}(s)\right)\mathrm{d}s\Big\}, \tag{15}$$

where $\boldsymbol{r}^{\mathsf{T}}(s)$ can be written as $\boldsymbol{b}^{\mathsf{T}}\mathrm{e}^{-Qs} - (w-1)\,\mathbb{E}[G]\,\boldsymbol{a}^{\mathsf{T}}\,\mathbb{E}[\mathrm{e}^{-Q(s-G^{\mathrm{res}})}\ 1_{\{G^{\mathrm{res}}\leq s\}}]$.

$\circ$ **Derivation of the joint transform of $L(t)$ and $\Lambda(t)$.** The above joint transform of $L(t)$ and $\boldsymbol{X}(t)$, conditional on $\boldsymbol{X}(0)=\boldsymbol{x}$, now leads to the following two theorems. Define

$$H(v,w,y) := v\mathrm{e}^{-Qy} - (w-1)\,\mathbb{E}[G]\,\Omega(y); \quad \Omega(y) := \mathbb{E}[\mathrm{e}^{-Q(y-G^{\mathrm{res}})}\,1_{\{G^{\mathrm{res}}\leq y\}}]. \tag{16}$$

Theorem 3.1. *Let $L(0)=0$. Then, with $\tilde{\boldsymbol{r}}^\intercal(y) := \boldsymbol{a}^\intercal H(v,w,y)$,*

$$\mathbb{E}[w^{L(t)}\mathrm{e}^{-v\Lambda(t)}|\boldsymbol{X}(0)=\boldsymbol{x}] = \exp\Big\{-\tilde{\boldsymbol{r}}^\intercal(t)\boldsymbol{x} - \int_{y=0}^{t}\eta\left(\tilde{\boldsymbol{r}}(y)\right)\mathrm{d}y\Big\}. \tag{17}$$

Proof. Plug in $\boldsymbol{b}=v\boldsymbol{a}$ in (15). ☐

The next two corollaries present extensions of Theorem 3.1 to the case that $L(0)=n_0$, taking into account the residual service times of those n_0 customers. Let G_j^{res} be the residual service time of the j-th customer present at time $t=0$, $j=1,\dots,n_0$.

Corollary 3.2. *Let $L(0)=n_0$. Then, with $\tilde{\boldsymbol{r}}^\intercal(y) = \boldsymbol{a}^\intercal H(v,w,y)$,*

$$\begin{aligned}\mathbb{E}[w^{L(t)}\mathrm{e}^{-v\Lambda(t)}&|L(0)=n_0,\boldsymbol{X}(0)=\boldsymbol{x},G_1^{\mathrm{res}}=g_1,\dots,G_{n_0}^{\mathrm{res}}=g_{n_0}]\\ &= \left(\prod_{j=1}^{n_0} w^{\mathbb{1}_{\{g_j>t\}}}\right)\exp\Big\{-\tilde{\boldsymbol{r}}^\intercal(t)\boldsymbol{x} - \int_{y=0}^{t}\eta\left(\tilde{\boldsymbol{r}}(y)\right)\mathrm{d}y\Big\}.\end{aligned} \tag{18}$$

Proof. Define $L(t) := L_{\mathrm{old}}(t)+L_{\mathrm{new}}(t)$, where $L_{\mathrm{old}}(t)$ and $L_{\mathrm{new}}(t)$ are the numbers of customers still present at time t, that were present in the system at time $t=0$ and that have joined the system in $[0,t]$, respectively. Conditional on $L(0)=n_0$, $\boldsymbol{X}(0)=\boldsymbol{x}$, $G_1^{\mathrm{res}}=g_1,\dots,G_{n_0}^{\mathrm{res}}=g_{n_0}$, the random variables $L_{\mathrm{old}}(t)$ and $L_{\mathrm{new}}(t)$ are independent. The corollary now follows immediately from Theorem 3.1. ☐

We now consider the case of exponential service times with parameter μ. In this case $H(v,w,y)$ simplifies to

$$H_{\exp}(v,w,y) := v\mathrm{e}^{-Qy} - (w-1)(\mu\mathbb{I}-Q)^{-1}(\mathrm{e}^{-Qy}-\mathrm{e}^{-\mu y}\mathbb{I}), \tag{19}$$

with $\mathbb{I}$ the d-dimensional identity matrix,

Corollary 3.3. *Let $L(0)=n_0$, and suppose that G is exponentially distributed with parameter μ. Then, with $\tilde{\boldsymbol{r}}^\intercal(y) = \boldsymbol{a}^\intercal H_{\exp}(v,w,y)$,*

$$\begin{aligned}\mathbb{E}[w^{L(t)}\mathrm{e}^{-v\Lambda(t)}&|L(0)=n_0,\boldsymbol{X}(0)=\boldsymbol{x}]\\ &= \Big(1+(w-1)\mathrm{e}^{-\mu t}\Big)^{n_0}\exp\Big\{-\tilde{\boldsymbol{r}}^\intercal(t)\boldsymbol{x} - \int_{y=0}^{t}\eta\left(\tilde{\boldsymbol{r}}(y)\right)\mathrm{d}y\Big\}.\end{aligned} \tag{20}$$

Proof. First observe that as G is exponentially distributed with parameter μ, all G_j^{res} (for $j=1,2,\dots,n_0$) are i.i.d. and also exponentially distributed with the same parameter μ; this implies $\mathbb{E}[w^{\mathbb{1}_{\{G^{\mathrm{res}}>t\}}}] = 1+(w-1)\mathrm{e}^{-\mu t}$. After verifying

$$\mathbb{E}[G]\,\mathbb{E}[\mathrm{e}^{-Q(t-G^{\mathrm{res}})}\,1_{\{G^{\mathrm{res}}\leq t\}}] = (\mu\mathbb{I}-Q)^{-1}(\mathrm{e}^{-Qt}-\mathrm{e}^{-\mu t}\mathbb{I}), \tag{21}$$

the claim in the corollary follows. ☐

Remark 1. Taking the limit $t \to \infty$ in Theorem 3.1 gives us the transform of the stationary number of customers jointly with the stationary rate of the shot-noise process, i.e.,

$$\mathbb{E}[w^{L(\infty)}\mathrm{e}^{-v\Lambda(\infty)}|\boldsymbol{X}(0)=\boldsymbol{x}] = \exp\left\{-\int_{y=0}^{\infty}\eta\left(\tilde{\boldsymbol{r}}(y)\right)\mathrm{d}y\right\}; \tag{22}$$

to see this, observe that $H(v,w,y) \to 0$ as $y \to \infty$. Formula (22) was already derived in [6, Eqn. (62)]. Our Theorem 3.1 generalizes this to the transient case.

4 Derivation of recursive moment relations

In this section we provide an iterative scheme to determine mixed moments of $(L(t), \Lambda(t))$ (provided that they exist), under the assumption that $(L(0), \boldsymbol{X}(0)) = (0, \boldsymbol{x})$. Our starting point is Theorem 3.1, which gives

$$\Phi(w,v) := \mathbb{E}[w^{L(t)}\mathrm{e}^{-v\Lambda(t)}|L(0)=0, \boldsymbol{X}(0)=\boldsymbol{x}] = \mathrm{e}^{g(w,v)}, \tag{23}$$

where, with $\tilde{\boldsymbol{r}}^{\intercal}(y) = \boldsymbol{a}^{\intercal}H(v,w,y)$,

$$g(w,v) := -\tilde{\boldsymbol{r}}^{\intercal}(t)\boldsymbol{x} - \int_{y=0}^{t}\eta\left(\tilde{\boldsymbol{r}}(y)\right)\mathrm{d}y. \tag{24}$$

We introduce the notation

$$\mathbb{M}_n^r(w,v) := \mathfrak{d}_{r,n}\Phi(w,v), \quad \text{where } \mathfrak{d}_{r,n} := \frac{\partial^r}{\partial w^r}\frac{\partial^n}{\partial v^n}. \tag{25}$$

Notice that

$$\mathbb{M}_n^r(1,0) = (-1)^n\mathbb{E}[L(t)(L(t)-1)\dots(L(t)-r+1)\,(\Lambda(t))^n], \tag{26}$$

so knowledge of these terms leads to expressions for the cross-moments $\mathbb{E}[(L(t))^r(\Lambda(t))^n]$. The nice form of (23) enables us to evaluate the $\mathbb{M}_r^n(w,v)$ recursively; plugging in $w=1$ and $v=0$ eventually leads to moment expressions. Below we first present and derive the recurrence relations, and then provide explicit expressions for the first two moments.

Proposition 4.1. *For $n = 0, 1, \dots$, $r = 0, 1, \dots$,*

$$\begin{aligned}
\mathbb{M}_{n+1}^{r+1}(w,v) &= \sum_{k=0}^{n}\binom{n}{k}\sum_{l=0}^{r+1}\binom{r+1}{l}\left(\mathfrak{d}_{\ell,k+1}g(w,v)\right)\ \mathbb{M}_{n-k}^{r+1-l}(w,v), &(27)\\
&\equiv \sum_{l=0}^{r}\binom{r}{l}\sum_{k=0}^{n+1}\binom{n+1}{k}\left(\mathfrak{d}_{k,\ell+1}g(w,v)\right)\ \mathbb{M}_{n+1-k}^{r-l}(w,v), &(28)\\
\mathbb{M}_{0}^{r+1}(w,v) &= \sum_{l=0}^{r}\binom{r}{l}\left(\mathfrak{d}_{\ell+1,0}g(w,v)\right)\ \mathbb{M}_{0}^{r-l}(w,v), &(29)\\
\mathbb{M}_{n+1}^{0}(w,v) &= \sum_{k=0}^{n}\binom{n}{k}\left(\mathfrak{d}_{0,k+1}g(w,v)\right)\ \mathbb{M}_{n-k}^{0}(w,v). &(30)
\end{aligned}$$

Proof. We have, for $n = 0, 1, \ldots, r = 0, 1, \ldots$:

$$\begin{aligned}\mathbb{M}_{n+1}^{r+1}(w,v) &= \left(\mathfrak{d}_{r+1,n+1}\mathrm{e}^{g(w,v)}\right) = \mathfrak{d}_{r+1,n}\left(\frac{\partial g(w,v)}{\partial v}\mathrm{e}^{g(w,v)}\right) \\ &= \frac{\partial^{r+1}}{\partial w^{r+1}}\left(\sum_{k=0}^{n}\binom{n}{k}\frac{\partial^{k+1}g(w,v)}{\partial v^{k+1}}\mathbb{M}_{n-k}^{0}(w,v)\right) \\ &= \sum_{k=0}^{n}\binom{n}{k}\left[\sum_{l=0}^{r+1}\binom{r+1}{l}\frac{\partial^{l}}{\partial w^{l}}\left(\frac{\partial^{k+1}g(w,v)}{\partial v^{k+1}}\right)\mathbb{M}_{n-k}^{r+1-l}(w,v)\right].\end{aligned} \tag{31}$$

This proves (27); (28) follows by symmetry. Eqns. (29) and (30) follow analogously. □

Similar recursive relations for higher moments have been derived in, e.g., [1,19].

4.1 First two transient moments

Relying on Proposition 4.1, we now compute the first and (joint) second moments explicitly. Below $\boldsymbol{J}$ denotes $\boldsymbol{J}(1)$. In addition, from the definition of $\eta(\cdot)$, we have that the gradient vector $\nabla^{(1)}\eta(\mathbf{0})$ equals $\mathbb{E}[\boldsymbol{J}]$. The Hessian $\nabla^{(2)}\eta(\mathbf{0})$ is $-\Sigma$, with $\Sigma_{ij} := \mathbb{E}[J_i(1)J_j(1)]$.

Corollary 4.2. *For $t \geqslant 0$,*

$$\mathbb{E}[\Lambda(t)] = \boldsymbol{a}^\intercal\,\mathrm{e}^{-Qt}\,\boldsymbol{x} + \boldsymbol{a}^\intercal Q^{-1}\left(\mathbb{I} - \mathrm{e}^{-Qt}\right)\mathbb{E}[\boldsymbol{J}], \tag{32}$$

$$\mathbb{E}[L(t)] = \mathbb{E}[G]\,\boldsymbol{a}^\intercal\left(\Omega(t)\boldsymbol{x} + \int_{y=0}^{t}\Omega(y)\mathrm{d}y\,\mathbb{E}[\boldsymbol{J}]\right), \tag{33}$$

where $\mathbb{I}$ is the $d \times d$ identity matrix.

Proof. See Appendix A. □

Remark 2. If the service time G is exponentially distributed with parameter μ, then it can be verified that $\Omega(t) = -\mu Q_\mu^{-1}(\mathrm{e}^{-Qt} - e^{-\mu t}\mathbb{I})$, with $Q_\mu^{-1} := (Q - \mu\mathbb{I})^{-1}$. In addition,

$$\begin{aligned}\int_{y=0}^{t}\Omega(y)\,\mathrm{d}y &= Q_\mu^{-1}\big(1 - \mathrm{e}^{-\mu t}\big) - \mu Q_\mu^{-1}Q^{-1}\left(\mathbb{I} - \mathrm{e}^{-Qt}\right) \\ &= Q^{-1} + \mu Q_\mu^{-1}Q^{-1}\mathrm{e}^{-Qt} - Q_\mu^{-1}\mathrm{e}^{-\mu t}\mathbb{I}.\end{aligned} \tag{34}$$

We thus obtain

$$\mathbb{E}[L(t)] = \boldsymbol{a}^\intercal\Big(\frac{1}{\mu}Q^{-1}\mathbb{E}[\boldsymbol{J}] - Q_\mu^{-1}\left(\mathrm{e}^{-Qt} - e^{-\mu t}\mathbb{I}\right)\boldsymbol{x} + Q_\mu^{-1}\Big(Q^{-1}\mathrm{e}^{-Qt} - \frac{\mathrm{e}^{-\mu t}}{\mu}\mathbb{I}\Big)\,\mathbb{E}[\boldsymbol{J}]\Big). \tag{35}$$

Corollary 4.3. *For $t \geqslant 0$,*

$$\mathbb{V}\mathrm{ar}[\Lambda(t)] = \boldsymbol{a}^\intercal\left(\int_0^t \mathrm{e}^{-Qy}\,\Sigma\,\mathrm{e}^{-Q^\intercal y}\mathrm{d}y\right)\boldsymbol{a}, \tag{36}$$

$$\mathbb{C}\mathrm{ov}(\Lambda(t), L(t)) = \mathbb{E}[G]\,\boldsymbol{a}^\intercal\left(\int_0^t \Omega(y)\,\Sigma\mathrm{e}^{-Q^\intercal y}\,\mathrm{d}y\right)\boldsymbol{a}, \tag{37}$$

$$\mathbb{V}\mathrm{ar}[L(t)] = \mathbb{E}[L(t)] + (\mathbb{E}[G])^2\,\boldsymbol{a}^\intercal\left(\int_0^t \Omega(y)\,\Sigma\,\Omega^\intercal(y)\,\mathrm{d}y\right)\boldsymbol{a}. \tag{38}$$

Proof. See Appendix B. □

To obtain explicit expressions for higher moments, we need to assume that Q and Σ commute. If Q and Σ do not commute, then the integrals appearing in the variances and the covariance do not allow easy computation; see [23].

Remark 3. If the service time G is exponentially distributed with parameter μ and the matrices Q and Σ commute, then straightforward computations reveal that, with $\bar{Q}_\mu^{-1} := (Q+\mu\mathbb{I})^{-1}$ and $Q_+ := Q + Q^\intercal$,

$$\mathbb{V}\mathrm{ar}[\Lambda(t)] = \boldsymbol{a}^\intercal\, \Sigma\, Q_+^{-1}\, (\mathbb{I} - \mathrm{e}^{-Q_+t})\, \boldsymbol{a}, \tag{39}$$

$$\mathbb{C}\mathrm{ov}(\Lambda(t), L(t)) = -\,\boldsymbol{a}^\intercal Q_\mu^{-1}\, \Sigma\, \Big[Q_+^{-1}\, (\mathbb{I} - \mathrm{e}^{-Q_+t}) - (\bar{Q}_\mu^{-1})^\intercal\, (\mathbb{I} - \mathrm{e}^{-(Q^\intercal+\mu\mathbb{I})t})\Big]\boldsymbol{a}, \tag{40}$$

$$\begin{aligned}\mathbb{V}\mathrm{ar}[L(t)] = \mathbb{E}[L(t)] + \boldsymbol{a}^\intercal Q_\mu^{-1}\, \Sigma\, \Big[&Q_+^{-1}\, (\mathbb{I} - \mathrm{e}^{-Q_+t}) - \bar{Q}_\mu^{-1}(\mathbb{I} - \mathrm{e}^{-(Q+\mu\mathbb{I})t}) \\ &-(\bar{Q}_\mu^{-1})^\intercal(\mathbb{I} - \mathrm{e}^{-(Q^\intercal+\mu\mathbb{I})t}) + \frac{1-\mathrm{e}^{-2\mu t}}{2\mu}\mathbb{I}\Big](Q_\mu^{-1})^\intercal\boldsymbol{a}.\end{aligned} \tag{41}$$

4.2 First two stationary moments

In this subsection we consider stationary moments. Stationary moments can be easily derived from the previous subsection by taking the limit $t \to \infty$. Define by $\Lambda(\infty)$ and $L(\infty)$ the stationary versions of $\Lambda(t)$ and $L(t)$, respectively. The results in the corollaries below agree with those obtained in [6].

Corollary 4.4. *In stationarity,*

$$\mathbb{E}[\Lambda(\infty)] = \boldsymbol{a}^\intercal Q^{-1}\, \mathbb{E}[\boldsymbol{J}], \tag{42}$$

$$\mathbb{E}[L(\infty)] = \mathbb{E}[G]\, \boldsymbol{a}^\intercal \left(\int_{y=0}^{\infty} \Omega(y)\, \mathrm{d}y\right) \mathbb{E}[\boldsymbol{J}] = \mathbb{E}[G]\, \boldsymbol{a}^\intercal Q^{-1}\, \mathbb{E}[\boldsymbol{J}]. \tag{43}$$

Proof. Taking the limit $t \to \infty$ in Corollary 4.2 yields the desired results. □

Corollary 4.5. *In stationarity,*

$$\mathbb{V}\mathrm{ar}[\Lambda(\infty)] = \boldsymbol{a}^\intercal \left(\int_{y=0}^{\infty} \mathrm{e}^{-Qy}\, \Sigma\, \mathrm{e}^{-Q^\intercal y}\mathrm{d}y\right) \boldsymbol{a}, \tag{44}$$

$$\mathbb{C}\mathrm{ov}(\Lambda(\infty), L(\infty)) = \mathbb{E}[G]\, \boldsymbol{a} \left(\int_0^{\infty} \Omega(y)\, \Sigma\, \mathrm{e}^{-Q^\intercal y}\mathrm{d}y\right) \boldsymbol{a}, \tag{45}$$

$$\mathbb{V}\mathrm{ar}[L(\infty)] = \mathbb{E}[L(\infty)] + (\mathbb{E}[G])^2\, \boldsymbol{a}^\intercal \left(\int_0^{\infty} \Omega(y)\, \Sigma\, \Omega^\intercal(y)\, \mathrm{d}y\right)\boldsymbol{a}. \tag{46}$$

Proof. Taking the limit $t \to \infty$ in Corollary 4.3 yields the desired results. □

5 Asymptotic analysis

In this section we derive the asymptotics of the queue length process, under assumptions regarding the tail behavior of the shot-noise process. Additionally, we discuss a central limit theorem for $(\Lambda(t), L(t))$ for a given $t \geqslant 0$, as well as the associated functional central limit theorem.

5.1 Tail asymptotics

In this subsection we show that regularly varying behavior of $\boldsymbol{J}(\cdot)$ leads to regularly varying behavior of $L(t)$. To do so, we first give the definition of a regularly varying random variable, cf. [5].

Definition 5.1. A random variable X on $[0,\infty)$ is called regularly varying of index $-\nu$, denoted by $\mathscr{R}(-\nu)$, with $\nu>0$, if

$$\mathbb{P}(X>x)=x^{-\nu}\ell(x),\ x\geqslant 0, \tag{47}$$

with $\ell(\cdot)$ a slowly varying function at infinity, i.e., $\ell(\gamma x)/\ell(x)\to 1$ as $x\to\infty$ for all $\gamma>0$.

We assume the subordinator $\boldsymbol{J}$ is multivariate regularly varying at infinity of index $-\nu\in(1,2)$, in the sense that, for some $\tilde{\boldsymbol{c}}\in\mathbb{R}^d_{\geqslant 0}$, as $\boldsymbol{z}\downarrow 0$,

$$\eta(\boldsymbol{z})=\tilde{\boldsymbol{c}}^\intercal\boldsymbol{z}+z_i^\nu\ell(z_i^{-1})+o(\|\boldsymbol{z}\|^\nu), \tag{48}$$

for some $i\in\{1,\dots,d\}$. Using Theorem 3.1 we thus have that, as $w\uparrow 1$,

$$\begin{aligned}\mathbb{E}[w^{L(t)}|\boldsymbol{X}(0)=\boldsymbol{x}]-1&=\exp\Big\{-(1-w)\mathbb{E}[G]\,\boldsymbol{a}^\intercal\,\Omega(t)\,\boldsymbol{x}-\int_{y=0}^t\eta\big((1-w)\mathbb{E}[G]\,\Omega^\intercal(y)\boldsymbol{a}\big)\mathrm{d}y\Big\}-1\\&=-(1-w)\mathbb{E}[G]\,\boldsymbol{a}^\intercal\,\Omega(t)\,\boldsymbol{x}-\int_{y=0}^t(1-w)\mathbb{E}[G]\,\boldsymbol{a}^\intercal\,\Omega(y)\,\tilde{\boldsymbol{c}}\,\mathrm{d}y\\&\quad-(1-w)^\nu\ell\left(\frac{1}{1-w}\right)\int_{y=0}^t\big(\mathbb{E}[G]\,\boldsymbol{a}^\intercal\,\Omega(y)\,\boldsymbol{e}_i\big)^\nu\,\mathrm{d}y+o((1-w)^\nu),\end{aligned} \tag{49}$$

and hence, as $w\uparrow 1$,

$$\begin{aligned}&\mathbb{E}[w^{L(t)}|\boldsymbol{X}(0)=\boldsymbol{x}]-1+(1-w)\mathbb{E}[L(t)|\boldsymbol{X}(0)=x]\\&=-(1-w)^\nu\ell\left(\frac{1}{1-w}\right)\int_{y=0}^t\big(\mathbb{E}[G]\,\boldsymbol{a}^\intercal\,\Omega(y)\,\boldsymbol{e}_i\big)^\nu\,\mathrm{d}y+o((1-w)^\nu),\end{aligned} \tag{50}$$

with $\boldsymbol{e}_i$ the i-th unit vector. Applying a Tauberian theorem, see [5, Thm 8.1.6], yields the following result.

Proposition 5.1. *Assuming* (48), *$L(t)$ has a regularly varying tail at infinity of index $-\nu$.*

We treat two examples, one in which the heaviest component of $\boldsymbol{J}(\cdot)$ is compound Poisson, and one in which the heaviest component of $\boldsymbol{J}(\cdot)$ is α-stable.

We start with the compound Poisson case, in which we assume that $J_i(\cdot)$ corresponds to a compound Poisson process with arrival rate λ_i and service requirements that are distributed like a random variable B_i such that $\mathbb{P}(B_i>x)=x^{-\nu}\ell(x)$. In addition, we assume that the tails of the other components of $\boldsymbol{J}(\cdot)$ are lighter, so that we can write that $\eta(\boldsymbol{z})$ is necessarily of the form, as $\boldsymbol{z}\downarrow 0$,

$$\boldsymbol{c}^\intercal\boldsymbol{z}+\lambda_i\Big(z_i\,\mathbb{E}[B_i]+\Gamma(1-\nu)z_i^\nu\ell(z_i^{-1})+o(\|\boldsymbol{z}\|^\nu)\Big)=\tilde{\boldsymbol{c}}^\intercal\boldsymbol{z}+\lambda_i\Gamma(1-\nu)z_i^\nu\ell(z_i^{-1})+o(\|\boldsymbol{z}\|^\nu). \tag{51}$$

We conclude from Proposition 5.1 that $L(t)$ inherits the regular variation of index $-\nu$.

In the second example, the heaviest component (say the i-th) of $\boldsymbol{J}(\cdot)$ is α-stable with parameter ν. As we assumed that $\boldsymbol{J}$ corresponds to an increasing subordinator, we have to take the skewness parameter of the stable process (which we refer to as β) equal to 1; as we assumed the subordinator to have finite mean, we take the index of stability (commonly referred to as α) in $(1, 2)$. We thus have [11, Prop. 2.25] that $\eta(\boldsymbol{z})$ is of the form, as $\boldsymbol{z} \downarrow 0$,

$$\tilde{\boldsymbol{c}}^{\mathsf{T}} \boldsymbol{z} - \frac{1}{\cos(\pi(\alpha/2 - 1))} z_i^{\alpha} + o(\|\boldsymbol{z}\|^{\alpha}). \tag{52}$$

Again, by Proposition 5.1, $L(t)$ inherits the regular variation, in this case of index $-\alpha$.

5.2 Diffusion limits

In this subsection we focus on diffusion limits. First we establish an ordinary central limit theorem (CLT) for the number of customers $L(t)$ jointly with the shot-noise driven rate $\Lambda(t)$. Later we point out how to extend this to a functional version.

◦ **CLT result.** Assuming the first two moments of the subordinator $\boldsymbol{J}(\cdot)$ are finite (which means $\mathbb{E}[J_i^2] < \infty$ for all $i = 1, \ldots, d$), we apply a linear scaling to the initial shot-noise rate and the Lévy exponent, i.e., $\boldsymbol{x} \mapsto N\boldsymbol{x}$ and $\eta(\cdot) \mapsto N\eta(\cdot)$, respectively. The scaling of $\eta(\cdot)$ means that $\boldsymbol{c} \mapsto N\boldsymbol{c}$ and $\nu(\cdot) \mapsto N\nu(\cdot)$ in (1). We denote the N-scaled version of $(L(t), \Lambda(t))$ by $(L_N(t), \Lambda_N(t))$; also, $\boldsymbol{X}_N(\cdot)$ is the counterpart of $\boldsymbol{X}(\cdot)$ under the scaling.
By Theorem 3.1, we observe that the transform of $(L_N(t), \Lambda_N(t))$ is exponential in N. In other words, under the scaling imposed, $(L_N(t), \Lambda_N(t))$ can be seen as the sum of N i.i.d. random vectors. We conclude that we are in the setting of the classical (bivariate version of the) central limit theorem.
More specifically, define

$$K_N(t) := N^{-1/2} \left(\begin{pmatrix} \Lambda_N(t) \\ L_N(t) \end{pmatrix} - N \begin{pmatrix} \mathbb{E}[\Lambda(t)] \\ \mathbb{E}[L(t)] \end{pmatrix} \right), \tag{53}$$

and

$$C(t) := \begin{pmatrix} \mathbb{V}\mathrm{ar}\Lambda(t) & \mathbb{C}\mathrm{ov}(\Lambda(t), L(t)) \\ \mathbb{C}\mathrm{ov}(\Lambda(t), L(t)) & \mathbb{V}\mathrm{ar}L(t) \end{pmatrix}; \tag{54}$$

the means, variances, and covariance in these expressions have been determined in Corollaries 4.2 and 4.3. Then the above yields the following result, in self-evident notation.

Proposition 5.2. *As $N \to \infty$,*

$$K_N(t) \xrightarrow{\mathrm{d}} \mathcal{N}\Big(\mathbf{0}, C(t)\Big). \tag{55}$$

◦ **CLT results at multiple points in time.** We now point out how the above central limit theorem extends to multiple points in time and to a functional variant. We start by considering times $t_1 \geqslant 0$ and $t_2 \geqslant t_1$. To this end, we first define

$$\begin{aligned} \psi_{\boldsymbol{w}}(u) := {} & (w_1 - 1)\mathbb{P}(t_1 - u < G \leqslant t_2 - u)\, 1_{\{u \in [0, t_1)\}} + \\ & (w_1 w_2 - 1)\mathbb{P}(G > t_2 - u)\, 1_{\{u \in [0, t_1)\}} + (w_2 - 1)\mathbb{P}(G > t_2 - u)\, 1_{\{u \in [t_1, t_2)\}}. \end{aligned} \tag{56}$$

Using the same argumentation as before, for our scaled process,

$$
\begin{aligned}
&\mathbb{E}[w_1^{L_N(t_1)} w_2^{L_N(t_2)} \mathrm{e}^{-\boldsymbol{b}_1^\intercal \boldsymbol{X}_N(t_1) - \boldsymbol{b}_2^\intercal \boldsymbol{X}_N(t_2)} | \boldsymbol{X}_N(0) = N\boldsymbol{x}] \\
&\qquad = \mathbb{E}\left[\exp\left(\int_0^{t_2} \psi_{\boldsymbol{w}}(u) \boldsymbol{a}^\intercal \boldsymbol{X}_N(u)\,\mathrm{d}u - \boldsymbol{b}_1^\intercal \boldsymbol{X}_N(t_1) - \boldsymbol{b}_2^\intercal \boldsymbol{X}_N(t_2)\right) \Big| \boldsymbol{X}_N(0) = N\boldsymbol{x}\right] \\
&\qquad = \left(\mathbb{E}\left[\exp\left(\int_0^{t_2} \psi_{\boldsymbol{w}}(u) \boldsymbol{a}^\intercal \boldsymbol{X}(u)\,\mathrm{d}u - \boldsymbol{b}_1^\intercal \boldsymbol{X}(t_1) - \boldsymbol{b}_2^\intercal \boldsymbol{X}(t_2)\right) \Big| \boldsymbol{X}(0) = \boldsymbol{x}\right]\right)^N .
\end{aligned} \tag{57}
$$

We thus conclude that in this setting, corresponding to two time epochs, again the transform is exponential in N. This implies that we can write the random vector under consideration as the sum of N i.i.d. random vectors. We thus have that, after centering and normalizing by $\sqrt{N}$, the concatenation of the vectors $(\Lambda_N(t_1), L_N(t_1))^\intercal$ and $(\Lambda_N(t_2), L_N(t_2))^\intercal$ converges to a normally distributed vector with the appropriate covariance matrix.

This argument naturally extends to any finite-dimensional distribution. Transforms of the type, for $0 \leqslant t_1 \leqslant \cdots \leqslant t_I$,

$$
\mathbb{E}\Big[\prod_{i=1}^{I} w_i^{L(t_i)} \exp\Big(-\sum_{i=1}^{I} \boldsymbol{b}_i^\intercal \boldsymbol{X}(t_i)\Big) | \boldsymbol{X}(0) = \boldsymbol{x}\Big] \tag{58}
$$

can be evaluated in the same manner as their single-dimensional counterpart in Section 3; for the N-scaled model these transforms are exponential in N. By differentiation all covariances, such as $\mathbb{C}\mathrm{ov}(L(t_i), L(t_j))$ for $i \neq j$, can be found.

A further extension concerns the functional version of the above central limit theorems. For the case of exponentially distributed service times (say, with parameter $\mu > 0$), this can be done by extending the approach discussed in [17], relying on the martingale central limit theorem and extensive use of Poisson processes with random time change. In the setup of [17] the rate process $\Lambda(\cdot)$ is (ordinary) shot-noise, whereas in this paper we consider a richer class of rate processes. Just like in [17], the limiting process of the number of customers in the system is an Ornstein–Uhlenbeck process driven by a superposition of a standard Brownian motion and an integrated Ornstein–Uhlenbeck process.

As argued in [17, Remark 3.6], if the service time distribution is not exponential, then the results are less clean; cf. [20] for results on a related model. Then there is convergence to a Gaussian process of the Kiefer type [7]; the underlying covariance structure aligns with the covariances mentioned above when discussing the finite-dimensional central limit theorem.

6 Numerical results

In this section we provide some numerical results for the case where the service time distribution is exponential with parameter μ. In Example 1 we numerically detemine the stationary and transient moments of the number of customers and the correlation coefficient between arrival rate and number of customers, for various parameter choices. Example 2 considers the sensitivity of the distribution of the number of customers for the service time distribution. In Example 3 we numerically verify the accuracy of using the marginal and bivariate CLT approximation for the number of customers and the input rate function.

Example 1. In this example we plot the means, the variances and the correlation of the arrival rate and the number of customers in the system, both in the transient and the stationary case. We consider a 2-dimensional shot-noise process in which we choose the parameters such that the marginal shot-noise processes are dependent. We start with an empty system and $\boldsymbol{X}(0) = \boldsymbol{x}$ and the parameters $\boldsymbol{x} = (2,2)^\intercal, \boldsymbol{a} = (1,1)^\intercal,\ \mu = \frac{3}{2}$,

$$Q = \begin{bmatrix} 4 & -1 \\ -1 & 4 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix},$$

and $\mathbb{E}[\boldsymbol{J}] = (1,1)^\intercal$. Observe that Remark 3 applies.

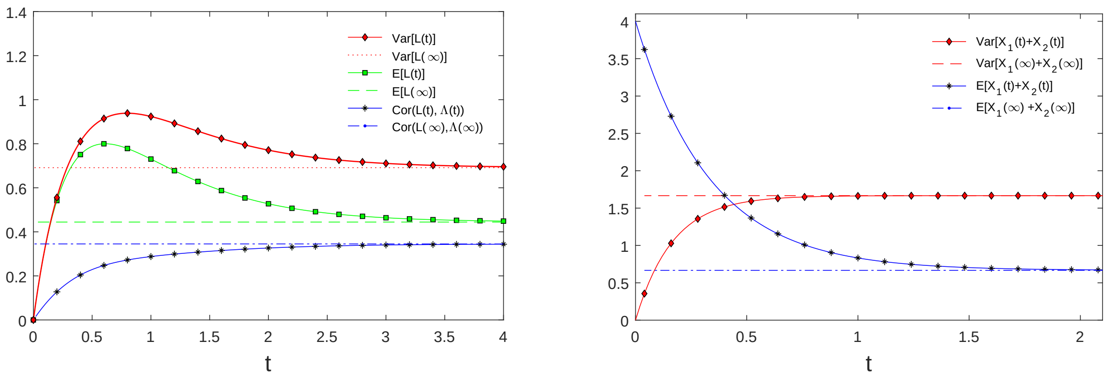


Figure 1: Transient and stationary expectations, variances and correlation of Example 1

Numerical observations. Figure 1 shows how the mean number of customers increases as t grows and converges to the stationary value; the same is true for the variance of the number of customers. The figure also confirms that the number of customers and shot-noise processes are overdispersed (meaning that the variance of the number of arrivals exceeds the mean). Also the transient correlation is plotted; with increasing t, it monotonously tends to the stationary value.

In the right part of Figure 1 we display the mean and variance of $\Lambda(t) = X_1(t) + X_2(t)$.

Example 2. In this example, we assess the sensitivity of the number of customers in the system with respect to the service time distribution. In an ordinary M/G/∞ system the distribution of the number of customers is insensitive to the service time distribution, given its mean, and therefore we are interested in a quantification of the sensitivity when the input process is Coxian. We compare the variance of $L(t)$ for two choices of the service time distribution: exponential $\mathbb{P}(G_1 > t) = \mathrm{e}^{-\mu t}$ and hyperexponential $\mathbb{P}(G_2 > t) = P\mathrm{e}^{-\mu_1 t} + (1-P)\mathrm{e}^{-\mu_2 t}$. The parameters for the service time distributions are chosen such that the expectations of both distributions are the same, i.e., $\mathbb{E}[G_1] = \mathbb{E}[G_2]$. The parameters of the hyperexponential distribution are $\mu_1 = 2\mu P$ and $\mu_2 = 2\mu(1-P)$ for $P \in (0,1)$ (i.e., $\frac{P}{\mu_1} = \frac{1-P}{\mu_2}$, which is called balanced means, see [22, p. 359]). The other required parameters are the same as in Example 1.

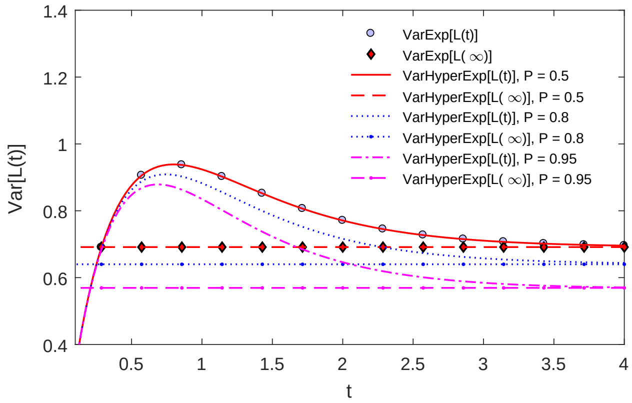


Figure 2: Transient and stationary variance of the number of customers for different values of P and increasing t

Numerical observations. In Figure 2, we have plotted the variance of the number of customers for different values of P. When $P = 1/2$ both service time distributions match, and hence the variance as well. For other values of P, we observe that the variance of the number of customers in our system varies, despite the fact that the mean service times coincide. In other words: there is no insensitivity (as in the M/G/∞ system). Surprisingly, a higher service time variance seems to lead to a lower variance of the number of customers.

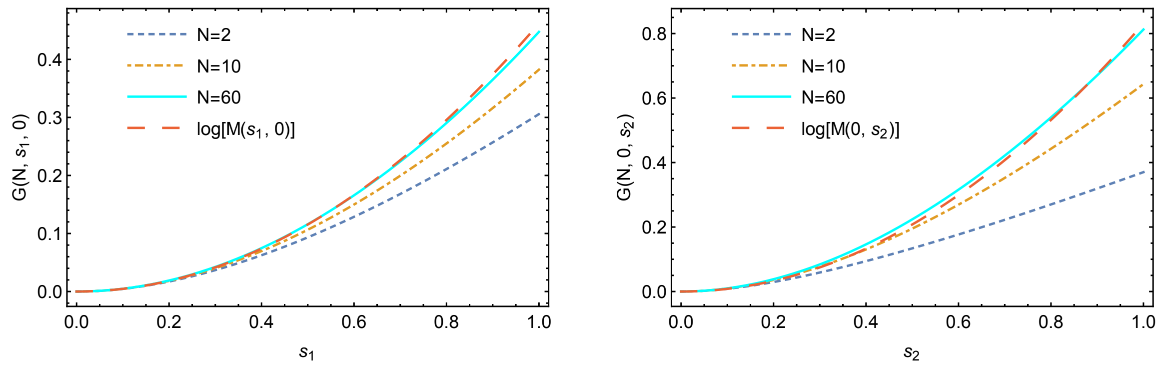
Example 3. In this example we numerically verify Proposition 5.2 for the marginal and bivariate CLT approximations for the number of customers and input rate function. From Corollary 3.3, we know an explicit expression for the joint transform of the number of customers and the input rate function, in the case of exponential service time distribution with parameter μ. Therefore we define the functions, for $\boldsymbol{s}^\intercal = (s_1, s_2)$, $\boldsymbol{Z}^\intercal = (Z_1, Z_2) \sim \mathcal{N}(\boldsymbol{0}, C(t))$,

$$\begin{aligned} \mathrm{G}(N, s_1, s_2) &= \log(\mathbb{E}[\mathrm{e}^{-\boldsymbol{s}^\intercal K_N(t)}]) \\ &= \log(\mathbb{E}[\mathrm{e}^{-N^{-1/2}s_1 L_N(t) - N^{-1/2}s_2 \Lambda_N(t)}) + s_1\sqrt{N}\,\mathbb{E}[L(t)] + s_2\sqrt{N}\,\mathbb{E}[\Lambda(t)], \end{aligned} \tag{59}$$

$$\log[\mathrm{M}(s_1, s_2)] = \log(\mathbb{E}[\mathrm{e}^{-\boldsymbol{s}^\intercal \boldsymbol{Z}}]) = \frac{1}{2}\boldsymbol{s}^\intercal C(t)\boldsymbol{s}, \tag{60}$$

where $K_N(t)$ and $C(t)$ are given by (53) and (54), respectively. In this example, our aim is to show the convergence of the function $G(N, s_1, s_2)$ for large values of N to the function $\log[\mathrm{M}(s_1, s_2)]$. To do so, we again consider a 2-dimensional shot-noise process with the Laplace exponent of the subordinator $\eta(\alpha_1, \alpha_2) := \lambda(1 - \frac{\mu_1}{\mu_1+\alpha_1}\frac{\mu_2}{\mu_2+\alpha_2})$. The parameters we choose are as follows: $t = 1, \lambda = \frac{1}{2}$, and $\mu_1 = \mu_2 = \frac{1}{2}$. The other required parameters are the same as in Example 1.

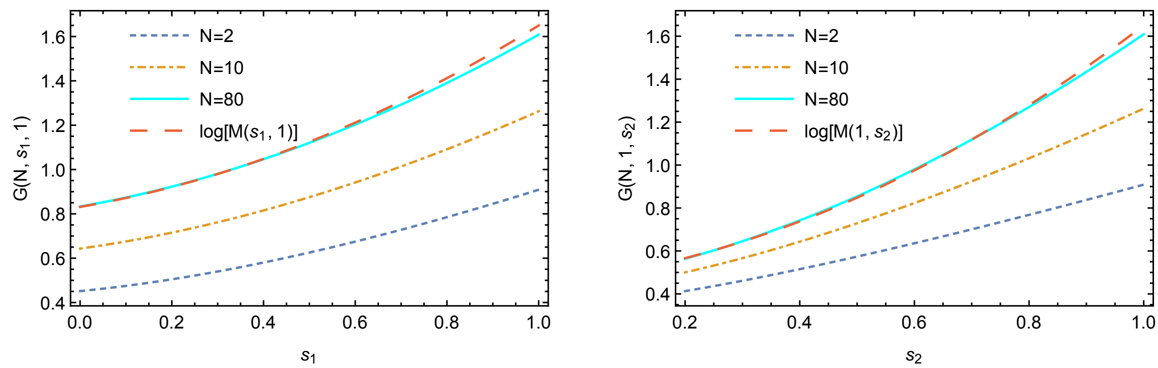
Numerical observations. For larger values of N, from Figures 3a and 3b, we can observe the convergence of the functions $G(N, s_1, 0)$ and $G(N, 0, s_2)$ to $\log[\mathrm{M}(s_1, 0)]$ and $\log[\mathrm{M}(0, s_2)]$, respectively. This means that the distribution of the scaled-normalized number of customers and input rate function both converge to a normal distribution with zero means and variances $\mathbb{V}ar[L(t)]$ and $\mathbb{V}ar[\Lambda(t)]$, respectively. Similarly, for large N, from Figures 4a and 4b we can also observe that the function $G(N, s_1, s_2)$, for some fixed values s_1 and s_2, converges to the function $\log[\mathrm{M}(s_1, s_2)]$. This supports the proposition that the joint distribution of the scaled-



(a) CLT for $L(t)$

(b) CLT for $\Lambda(t)$

Figure 3: Marginal CLT approximation for the number of customers and input rate function



(a) Bivariate CLT for $(L(t), \Lambda(t))$

(b) Bivariate CLT for $(L(t), \Lambda(t))$

Figure 4: Bivariate CLT approximation for the number of customers and input rate function.

normalized number of customers in the system and input rate function converges to a bivariate normal distribution with zero mean and covariance matrix $C(t)$.

7 Conclusions and future work

In this paper we have studied an infinite-server queue in which the input is a non-homogeneous Poisson process, which is generated by the modulation of a generalized shot-noise process. The rate function of the generalized shot-noise process is defined with respect to a Lévy subordinator. For this model, we have derived an explicit expression for the joint transient transform of the number of customers and the input rate function. Further, we have formulated a recursive procedure that explicitly identifies any transient as well as stationary moments. As asymptotic results we have studied the tail asymptotics of the queue length process, under the assumption that the heaviest component of subordinator $\boldsymbol{J}(\cdot)$ is *compound Poisson* or α*-stable*. Additionally, we derived a multivariate central limit theorem for $(\Lambda(t), L(t))$ at multiple points in time. We have also discussed how one can derive the functional central limit theorem for the number of customers in this system. In the end, we have added some numerical examples to provide insight into the behavior of the model.

Finally, we would like to list some suggestions for further research. (i) It would be interesting to extend the tail asymptotics of Proposition 5.1 to more general multivariate regularly varying subordinators. (ii) One should be able to prove a functional central limit theorem for the number of customers in the system, for *general* service time distributions. (iii) One could study the large deviations behavior of the system, as in [3,9].

Acknowledgments

The authors gratefully acknowledge useful discussions with Sem Borst and Offer Kella. The research for this paper is partly funded by the NWO Gravitation Project NETWORKS, Grant Number 024.002.003 (Boxma, Mandjes) and an NWO Top Grant, Grant Number 613.001.352 (Boxma, Mandjes, Saxena).

A Proof of Corollary 4.2

Proof. Letting $r = 0$ and $n = 0$ in (29) yields

$$\mathbb{E}[\Lambda(t)] \;\;=\;\; -\frac{\partial}{\partial v} g(1,v)\Big|_{v=0}, \quad \mathbb{E}[L(t)] = \frac{\partial}{\partial w} g(w,0)\Big|_{w=1}. \tag{61}$$

Firstly we compute $\mathbb{E}[\Lambda(t)]$. We obtain

$$\begin{aligned}\frac{\partial}{\partial v} g(1,v) &= \left(-\boldsymbol{a}^{\intercal} \mathrm{e}^{-Qt} \boldsymbol{x} - \int_{s=0}^{t} \frac{\partial}{\partial v} \eta \left(v \mathrm{e}^{-Q^{\intercal} y} \boldsymbol{a}\right) \mathrm{d}y\right) \\ &= -\left(\boldsymbol{a}^{\intercal} \mathrm{e}^{-Qt} \boldsymbol{x} + \int_{s=0}^{t} \left(\mathrm{e}^{-Q^{\intercal} y} \boldsymbol{a}\right)^{\intercal} \nabla^{(1)} \eta \left(v \mathrm{e}^{-Q^{\intercal} y} \boldsymbol{a}\right) \mathrm{d}y\right); \end{aligned} \tag{62}$$

substituting $v = 0$ in the above equation yields

$$\frac{\partial}{\partial v} g(1,v)\Big|_{v=0} = -\left(\boldsymbol{a}^{\intercal} \mathrm{e}^{-Qt} \boldsymbol{x} + \left(\int_{s=0}^{t} \boldsymbol{a}^{\intercal} \mathrm{e}^{-Qy} \mathrm{d}y\right) \nabla^{(1)} \eta(\boldsymbol{0})\right). \tag{63}$$

Evaluating the above integral and then combining (63) and (61) yields (32).
We now compute $\mathbb{E}[L(t)]$. We obtain

$$\begin{aligned}\frac{\partial g(w,0)}{\partial w} &= \mathbb{E}[G] \boldsymbol{a}^{\intercal}\, \Omega(t)\, \boldsymbol{x} - \int_{y=0}^{t} \frac{\partial}{\partial w} \eta\big(-(w-1)\,\mathbb{E}[G] \Omega^{\intercal}(y) \boldsymbol{a}\big) \mathrm{d}y \\ &= \mathbb{E}[G] \boldsymbol{a}^{\intercal}\, \Omega(t)\, \boldsymbol{x} + \int_{y=0}^{t} \mathbb{E}[G] \boldsymbol{a}^{\intercal} \Omega(y) \cdot \nabla^{(1)} \eta\big(-(w-1)\; \mathbb{E}[G] \Omega^{\intercal}(y) \boldsymbol{a}\big) \mathrm{d}y. \end{aligned} \tag{64}$$

Substituting $w = 1$ in the above equation gives

$$\frac{\partial}{\partial w} g(w,0)\Big|_{w=1} = \mathbb{E}[G] \boldsymbol{a}^{\intercal}\, \Omega(t)\, \boldsymbol{x} + \mathbb{E}[G]\; \boldsymbol{a}^{\intercal} \int_{y=0}^{t} \Omega(y)\, \mathrm{d}y\; \nabla^{(1)} \eta(\boldsymbol{0}). \tag{65}$$

This immediately yields (33). □

B Proof of Corollary 4.3

Proof. By (30), (27) and (29), respectively, we find

$$\mathbb{E}[\Lambda^2(t)] = \left.\frac{\partial g(1,v)}{\partial v}\right|_{v=0} \mathbb{M}_0^1(0,1) + \left.\frac{\partial^2 g(1,v)}{\partial v^2}\right|_{v=0} \mathbb{M}_0^0(0,1), \tag{66}$$

$$\mathbb{E}[\Lambda(t)L(t)] = -\left.\frac{\partial g(w,0)}{\partial w}\right|_{w=1} \mathbb{M}_0^1(0,1) - \left.\frac{\partial}{\partial v}\left(\frac{\partial g(w,v)}{\partial w}\right)\right|_{v=0,w=1} \mathbb{M}_0^0(0,1), \tag{67}$$

$$\mathbb{E}[L(t)(L(t)-1)] = \left.\frac{\partial g(w,0)}{\partial w}\right|_{w=1} \mathbb{M}_1^0(0,1) + \left.\frac{\partial^2 g(w,0)}{\partial w^2}\right|_{w=1} \mathbb{M}_0^0(0,1). \tag{68}$$

We thus obtain

$$\mathbb{E}[\Lambda^2(t)] = (\mathbb{E}[\Lambda(t)])^2 + \left.\frac{\partial^2 g(1,v)}{\partial v^2}\right|_{v=0}, \tag{69}$$

$$\mathbb{E}[\Lambda(t)L(t)] = \mathbb{E}[\Lambda(t)]\mathbb{E}[L(t)] - \left.\frac{\partial}{\partial v}\left(\frac{\partial g(w,v)}{\partial w}\right)\right|_{v=0,w=1}, \tag{70}$$

$$\mathbb{E}[L(t)(L(t)-1)] = (\mathbb{E}[L(t)])^2 + \left.\frac{\partial^2 g(w,0)}{\partial w^2}\right|_{w=1}. \tag{71}$$

First we compute the second moment of $\Lambda(t)$, i.e., $\mathbb{E}[\Lambda^2(t)]$. Elementary computations yield

$$\begin{aligned}\mathbb{E}[\Lambda^2(t)] &= (\mathbb{E}[\Lambda(t)])^2 + \left.\frac{\partial^2 g(1,v)}{\partial v^2}\right|_{v=0} \\ &= (\mathbb{E}[\Lambda(t)])^2 - \boldsymbol{a}^\intercal \left(\int_{y=0}^t \mathrm{e}^{-Qy} \left(\nabla^{(2)}\eta(\mathbf{0})\right) \mathrm{e}^{-Q^\intercal y} \mathrm{d}y\right)\boldsymbol{a},\end{aligned} \tag{72}$$

yielding (36) (recall that $\nabla^{(2)}\eta(\mathbf{0}) = -\Sigma$). Analogously, we compute the mixed moment of $\Lambda(t)$ and $L(t)$, i.e., $\mathbb{E}[\Lambda(t)L(t)]$. From (70),

$$\begin{aligned}\mathbb{E}[\Lambda(t)L(t)] &= \mathbb{E}[\Lambda(t)]\mathbb{E}[L(t)] - \left.\frac{\partial}{\partial v}\left(\frac{\partial g(w,v)}{\partial w}\right)\right|_{v=0,w=1} \\ &= \mathbb{E}[\Lambda(t)]\mathbb{E}[L(t)] - \mathbb{E}[G]\, \boldsymbol{a}^\intercal \left(\int_0^t \mathbb{E}[\mathrm{e}^{-Q(y-G^{\mathrm{res}})} 1_{\{G^{\mathrm{res}}\le y\}}] \left(\nabla^{(2)}\eta(\mathbf{0})\right)\mathrm{e}^{-Q^\intercal y}\, \mathrm{d}y\right)\boldsymbol{a},\end{aligned} \tag{73}$$

thus yielding (37). Finally, elementary computations reveal that

$$\begin{aligned}\mathbb{E}[L(t)(L(t)-1)] &= \big(\mathbb{E}[L(t)]\big)^2 + \left.\frac{\partial^2 g(w,0)}{\partial w^2}\right|_{w=1} \\ &= \big(\mathbb{E}[L(t)]\big)^2 - (\mathbb{E}[G])^2\, \boldsymbol{a}^\intercal \left(\int_0^t \Omega(y)\big(\nabla^{(2)}\eta(\mathbf{0})\big)\Omega^\intercal(y)\, \mathrm{d}y\right)\boldsymbol{a},\end{aligned} \tag{74}$$

leading to (38). □

References

[1] Asmussen, S., and Kella, O. Rate modulation in dams and ruin problems. *Journal of Applied Probability* **33(2)**, 523-535, 1996.

[2] Blom, J., De Turck, K., and Mandjes, M. Analysis of Markov-modulated infinite-server queues in the central-limit regime. *Probability in the Engineering and Informational Sciences* **29**, 433-459, 2015.

[3] Blom, J., De Turck, K. and Mandjes, M. Refined large deviations asymptotics for Markov-modulated infinite-server systems. *European Journal of Operational Research* **259(3)**, 1036-1044, 2017.

[4] Bar-Lev, S. K., Boxma, O., Mathijsen, B., and Perry, D. A blood bank model with perishable blood and demand impatience. *Stochastic Systems* **7**, 237-263, 2017.

[5] Bingham, N. H., Goldie, C. M., and Teugels, J. L. *Regular Variation*. Cambridge University Press, Cambridge, 1989.

[6] Boxma, O., Kella, O., and Mandjes, M. Infinite-server systems with Coxian arrivals. Eurandom report 2018-009; submitted, 2018.

[7] Csörgő, M., and Révész, P. *Strong Approximations in Probability and Statistics*. Academic Press, New York, 1981.

[8] Daw, A., and Pender, J. Queues driven by Hawkes processes, ArXiv: `1707.05143v1`. To appear in *Stochastic Systems*, 2018.

[9] De Turck, K., and Mandjes, M. Large deviations of an infinite-server system with a linearly scaled background process. *Performance Evaluation* **75**, 36-49, 2014.

[10] Eliazar, I., and Klafter, J. Lévy, Ornstein–Uhlenbeck, and subordination: spectral vs. jump description. *Journal of Statistical Physics* **119**, 165-196, 2005.

[11] Furrer, H. Risk theory and heavy-tailed Lévy processes. PhD thesis ETH Zürich. `https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/143398/eth-22556-02.pdf`, 1997.

[12] Lando, D. On Cox processes and credit risky securities. *Review of Derivatives Research* **2**, 99-120, 1998.

[13] Kella, O., and Whitt, W. Linear stochastic fluid networks. *Journal of Applied Probability* **36**, 244-260, 1999.

[14] Kella, O., and Stadje, W. Markov-modulated linear fluid networks with Markov additive input. *Journal of Applied Probability* **39**, 413-420, 2002.

[15] Kella. O., and Stadje. W. On hitting times for compound Poisson dams with exponential jumps and linear release rate. *Journal of Applied Probability* **38**, 781-786, 2001.

[16] Kim, S. H., and Whitt, W. Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing Service Operations Management* **16**, 464-480, 2014.

[17] Koops, D. T., Boxma, O. J., and Mandjes, M. R. H. Networks of $\cdot/G/\infty$ queues with shot-noise-driven arrival intensities. *Queueing Systems* **86**, 301-325, 2017.

[18] Koops, D. T., Saxena, M., Boxma, O. J., and Mandjes, M. R. H. Infinite server queues with Hawkes input. To appear in *Journal of Applied Probability*, 2019.

[19] Mandjes, M., Starreveld, N., and Bekker, R. Queues on a dynamically evolving graph. *arXiv preprint arXiv:1712.01573*, 2017.

[20] Pang, G. and Whitt, W. Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems* **65**, 325-364, 2010.

[21] Takács, L. *Introduction to the Theory of Queues*. Oxford University Press, New York, 1962.

[22] Tijms, H. C. *Stochastic Models: An Algorithmic Approach.* John Wiley & Sons Inc, 1994.

[23] Van Loan, C. Computing integrals involving the matrix exponential. *IEEE Transactions on Automatic Control* **23**, 395-404, 1978.