European Journal of Operational Research 242 (2015) 455–464

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

The intercept term of the asymptotic variance curve for some queueing output processes

Sophie Hautphenne[a,*], Yoav Kerner[b], Yoni Nazarathy[c], Peter Taylor[d]

[a] Department of Mathematics and Statistics, The University of Melbourne, Melbourne, Vic 3010, Australia
[b] Ben Gurion University, Beer Sheva, Israel
[c] The University of Queensland, Brisbane, Australia
[d] The University of Melbourne, Melbourne, Australia

ARTICLE INFO

Article history:
Received 7 November 2013
Accepted 22 October 2014
Available online 3 November 2014

Keywords:
Queueing
M/M/1/K queue
M/G/1 queue
Markovian Point Process
Output processes

ABSTRACT

We consider the output processes of some elementary queueing models such as the M/M/1/K queue and the M/G/1 queue. An important performance measure for these counting processes is their variance curve $v(t)$, which gives the variance of the number of customers in the time interval $[0, t]$. Recent work has revealed some non-trivial properties dealing with the asymptotic rate at which the variance curve grows. In this paper we add to these results by finding explicit expressions for the intercept term of the linear asymptote.

For M/M/1/K queues our results are based on the deviation matrix of the generator. It turns out that by viewing output processes as Markovian Point Processes and considering the deviation matrix, one can obtain explicit expressions for the intercept term, together with some further insight regarding the BRAVO (Balancing Reduces Asymptotic Variance of Outputs) effect. For M/G/1 queues our results are based on a classic transform of D. J. Daley. In this case we represent the intercept term of the variance curve in terms of the first three moments of the service time distribution. In addition we shed light on a conjecture of Daley, dealing with characterization of stationary M/M/1 queues within the class of stationary M/G/1 queues, based on the variance curve.

1. Introduction

Many models in applied probability and stochastic operations research involve counting processes. Such processes occur in supply chains, health care systems, communication networks as well as many other contexts involving service, logistics and/or technology. The canonical counting process example is the Poisson process. Generalizations include renewal processes, Markovian Point Processes (see for example Latouche & Ramaswami, 1999, Section 3.5 or Asmussen, 2003, Section XI.1), or general simple point processes on the line (see for example Daley & Vere-Jones, 2003).

Sometimes counting processes are used in their own right, while at other times they constitute components of more complicated models such as queues, population processes or risk models. In other instances, counting processes are implicitly defined and constructed through applied probability models. For example, a realization of a queue induces additional counting processes such as the departure process, $\{D(t), t \geq 0\}$, counting the number of serviced customers in the queue until time t.

Departure counting processes of queues have been heavily studied in applied probability and operations research. Classic applied probability surveys are Daley (1976) and Disney and Konig (1985). More recent studies in operations research are Hendricks (1992), Tan (1999) and Tan (1997) where the authors consider departures in and within manufacturing production lines. Indeed, from an operational viewpoint, quantification of the variability of flows within a network is key. A similar comment applies to the flows of finished products at the end of the production process. From a theoretical perspective, there remain some open questions about the ability to characterize $\{D(t)\}$ as a Markovian Point Process, as in Bean and Green (2000), Bean, Green, and Taylor (1998) and Olivier and Walrand (1994). Further, the discovery of the BRAVO effect (Balancing Reduces Asymptotic Variance of Outputs) has motivated research on the variability of departure processes of queues, particularly in critically loaded regimes. Recent papers on this topic are Al Hanbali, Mandjes, Nazarathy, and Whitt (2011), Daley (2011),

* Corresponding author. Tel.: +61 3 8344 9073; fax: +61 3 8344 4599.
E-mail address: sophiemh@unimelb.edu.au (S. Hautphenne).

http://dx.doi.org/10.1016/j.ejor.2014.10.051
0377-2217/

Daley, van Leeuwaarden, and Nazarathy (2014), Nazarathy (2011) and Nazarathy and Weiss (2008).

Next to the mean curve, $m(t) = \mathbb{E}[D(t)]$, an almost equally important performance measure of a counting processes is the variance curve, $v(t) = \mathrm{Var}(D(t))$. For example, for a Poisson process with rate α, the variance curve

$$v(t) = \alpha t$$

is the same as the mean curve. For more complicated counting processes, the variance curve is not as simple and is not the same as the mean curve. For example, for a stationary (also known as *equilibrium*) renewal-process with inter-renewal times distributed as the sum of two independent exponential random variables, each with mean $(2\alpha)^{-1}$, we have

$$m(t) = \alpha t - \frac{1}{4} + \frac{1}{4}e^{-4\alpha t}, \quad v(t) = \alpha\frac{1}{2}t + \frac{1}{8} - \frac{1}{8}e^{-4\alpha t}.$$

For the *ordinary* case of the same renewal process (the first inter-renewal time is distributed as all the rest) the variance curve is

$$v(t) = \alpha\frac{1}{2}t + \frac{1}{16} - te^{-4\alpha t} - \frac{1}{16}e^{-8\alpha t}.$$

These explicit examples are taken from Cox (1962, Section 4.5). In fact, for general, non-lattice, renewal processes (both equilibrium and ordinary), with inter-renewal times having a finite second moment, with squared coefficient of variation c^2, and mean α^{-1}, it is well known that,

$$v(t) = \alpha c^2 t + o(t), \tag{1}$$

as $t \to \infty$ (which is the limiting regime used throughout this paper). However, in general, a finer description of $v(t)$ (through the $o(t)$ term) is typically not as simple as in the examples above.

If the third moment of the inter-renewal time is finite, then

$$v(t) = \begin{cases} \alpha c^2 t + \frac{5}{4}(c^4 - 1) - \frac{2}{3}(\gamma c^3 - 2) + o(1), & \text{for the equilibrium case,} \\ \alpha c^2 t + \frac{1}{2}(c^4 - 1) - \frac{1}{3}(\gamma c^3 - 2) + o(1), & \text{for the ordinary case,} \end{cases} \tag{2}$$

where γ is the skewness coefficient of the inter-renewal time.[1] We remind the reader that for exponential random variables (making the renewal process a Poisson process), $c^2 = 1$ and $\gamma = 2$, and the ordinary and equilibrium versions of a Poisson process are identical. See Asmussen (2003) and Daley and Vere-Jones (2003) for more background on renewal processes. Eq. (2) appears under a slightly different representation in Cox (1962) and was essentially first found in Smith (1959). Generalizations of renewal processes are in Brown and Solomon (1975), Daley and Mohan (1978) and Hunter (1969).

The above examples indicate that, for counting processes in general, it is likely to be fruitful to look for an asymptotic expression for the variance curve of the form

$$v(t) = \bar{\nu}t + \bar{b} + o(1). \tag{3}$$

We refer to $\bar{\nu}$ as the *asymptotic variance rate* and to $\bar{b}$ as the *intercept term*. A point to observe is that, for a renewal process, $\bar{b}$ depends on the version of the renewal process (ordinary vs. equilibrium) while $\bar{\nu}$ does not. Since the latter depends on the initial conditions, we generally employ the notation $\bar{b}_{\mathrm{e}}$ for the stationary (equilibrium) system, $\bar{b}_0$ for systems starting empty and $\bar{b}_\theta$ for systems with arbitrary initial conditions.

[1] The skewness coefficient of a random variable X is $\mathbb{E}[(\frac{X-\mathbb{E}[X]}{\sqrt{\mathrm{Var}(X)}})^3]$.

Moving on from renewal processes to implicitly defined counting processes, the variance curve is typically more complicated to describe and characterize. For example, while the output of a stationary M/M/1 queue with arrival rate λ and service rate μ is simply a Poisson process with rate λ (see Kelly, 1979), the variance curve when the system starts empty at time 0 is much more complicated than $v(t) = \lambda t$. It can be represented in terms of integrals of expressions involving Bessel functions of the first kind, and requires several lines to be written out fully (as in Theorem 5.1 of Al Hanbali et al., 2011). Nevertheless (see Theorem 5.2 in Al Hanbali et al., 2011) the curve can be sensibly approximated as follows:

$$v(t) = \begin{cases} \lambda t - \dfrac{\rho}{(1-\rho)^2} + o(1), & \text{if } \rho < 1, \\ 2\left(1 - \dfrac{2}{\pi}\right)\lambda t - \sqrt{\dfrac{\lambda}{\pi}}\, t^{1/2} + \dfrac{\pi - 2}{4\pi} + o(1), & \text{if } \rho = 1, \\ \mu t - \dfrac{\rho}{(1-\rho)^2} + o(1), & \text{if } \rho > 1, \end{cases} \tag{4}$$

where $\rho = \lambda/\mu$.

As observed from the formula above, it may be initially quite surprising that the asymptotic variance rate is reduced by a factor of $2(1 - 2/\pi) \approx 0.73$ when ρ changes from being approximately 1 to exactly 1. This is a manifestation of the BRAVO effect. BRAVO was first observed for M/M/1/K queues in Nazarathy and Weiss (2008) in which case, as $K \to \infty$, the factor is 2/3, a fact that we confirm in this paper. It was later analyzed for M/M/1 queues and more generally GI/G/1 queues in Al Hanbali et al. (2011). BRAVO has been numerically conjectured for GI/G/1/K queues in Nazarathy (2011), and observed for multi-server M/M/s/K queues in the many-server scaling regime in Daley et al. (2014).

Our focus in this paper is on the more subtle intercept term $\bar{b}$. For a stationary M/M/1 queue, $\{D(t)\}$ is a Poisson process and thus $\bar{b}_{\mathrm{e}} = 0$. As opposed to that, for an $M/M/1$ queue starting empty, it follows from (4) that $\bar{b}_0 = -\rho/(1-\rho)^2$ as long as $\rho \neq 1$. When $\rho = 1$, we see from (4) that the variance curve does not have the asymptotic form (3). This can happen more generally. If, for example, there is sufficient long range dependence in the counting process, then the variance can grow super-linearly (see Daley & Vesilo, 1997 for some examples). This demonstrates that the asymptotic variance rate, $\bar{\nu}$, and the intercept term, $\bar{b}$, need not exist for every counting process. Nevertheless, for a variety of models and situations, both $\bar{\nu}$ and $\bar{b}$ exist, and thus the linear asymptote is well-defined. In such cases, having a closed formula is beneficial for performance analysis of the model at hand.

We are now faced with the challenge of finding the intercept term for other counting processes generated by queues. In this paper we carry out such an analysis for two models related to the M/M/1 queue: a finite capacity M/M/1/K queue, and an infinite capacity M/G/1 queue. Besides obtaining explicit formulas for $\bar{b}_{\mathrm{e}}$, $\bar{b}_0$ and $\bar{b}_\theta$, our investigation also pinpoints some of the analytical challenges involved and raises some open questions. Here is a summary of our main contributions.

1.1. M/M/1/K queues

In this case the departure process is a Markovian Point Process. The linear asymptote is then given by formulas based on the matrix $\Lambda^- = (\mathbf{1}\boldsymbol{\pi} - \Lambda)^{-1}$, where Λ is the generator matrix of the (finite) birth-death process, $\boldsymbol{\pi}$ is its stationary distribution taken as a row vector, and $\mathbf{1}$ is a column vector of 1's. In the case where $\rho = 1$, the distribution $\boldsymbol{\pi}$ is uniform and an explicit expression for Λ^- was

previously found, which in turn yielded the equilibrium version of the intercept term,

$$\overline{b}_{\mathrm{e}} = \frac{7K^4 + 28K^3 + 37K^2 + 18K}{180(K+1)^2},$$

in Proposition 4.4 of Nazarathy and Weiss (2008). When $\rho \neq 1$, the form of the inverse Λ^- is more complicated and an expression for $\overline{b}$ has not been previously known. We are now able to find such an expression for both the stationary version and for arbitrary initial conditions. Our results are based on relating Λ^- to the matrix

$$\Lambda^\sharp = \int_0^\infty (P(t) - \mathbf{1}\boldsymbol{\pi})\, dt,$$

where $P(\cdot)$ is the transition probability kernel of the birth-death process. The matrix $\Lambda^\sharp$ is called the *deviation matrix* (also known as the *Drazin inverse* of $-\Lambda$), and we are able to provide explicit expressions for the entries of this matrix. Our contribution also includes some useful results regarding the asymptotic covariance between the count and phase in arbitrary Markovian Point Processes which, to the best of our knowledge, have not appeared elsewhere.

1.2. Stable M/G/1 queues with finite third moment of G

When the third moment of the service time distribution is finite, then it is known that the stationary queue length has a finite variance. When the service time distribution is not exponential, the form of $\overline{b}$ has not previously been known. Our contribution is an exact expression for the $\overline{b}$ term based on the first three moments of G. We begin with $\overline{b}_{\mathrm{e}}$, after which we employ a simple coupling argument to find $\overline{b}_\theta$ and $\overline{b}_0$.

The structure of the rest of the paper is as follows: In Section 2 we present our M/M/1/K queue results for $\overline{b}$ together with a discussion of the deviation matrix and its application to Markovian Point Processes. In Section 3 we present our M/G/1 queue results for $\overline{b}$, and discuss a related conjecture of Daley, dealing with a characterization of the M/M/1 queue within the class of stationary M/G/1 queues. We conclude in Section 4.

2. The M/M/1/K queue

We begin our investigation with the M/M/1/K queue, where K denotes the total capacity of the system. In this case, it is well known that the departure process $\{D(t)\}$ is a Markovian Point Process and is a renewal processes only when $K = 1$ or $K = \infty$. Some standard references on Markovian Point Processes are Asmussen (2003, Section XI.1) and Latouche and Ramaswami (1999, Section 3.5).

Denote the arrival rate by $\lambda > 0$, the service rate by $\mu > 0$ and let $\rho = \lambda/\mu$ be the traffic intensity. The queue length process, $\{Q(t)\}$, is a continuous-time Markov chain on the state space $\{0, 1, \ldots, K\}$, with generator matrix Λ and stationary distribution (row) vector $\boldsymbol{\pi}$ given by

$$\Lambda = \begin{bmatrix} -\lambda & \lambda & & & 0 \\ \mu & -(\mu+\lambda) & \lambda & & \\ & \ddots & \ddots & \ddots & \\ & & \mu & -(\mu+\lambda) & \lambda \\ 0 & & & \mu & -\mu \end{bmatrix},$$

$$\boldsymbol{\pi} = \begin{cases} \dfrac{1-\rho}{1-\rho^{K+1}}[1, \rho, \rho^2, \ldots, \rho^K], & \text{for } \rho \neq 1, \\ \dfrac{1}{K+1}\mathbf{1}', & \text{for } \rho = 1. \end{cases}$$

The departure process $\{D(t)\}$ is a Markovian Point Process of which the phase-process is $\{Q(t)\}$, and the event intensity matrix Λ_1 is given by

$$\Lambda_1 = \begin{bmatrix} 0 & 0 & & & 0 \\ \mu & 0 & 0 & & \\ & \ddots & \ddots & \ddots & \\ & & \mu & 0 & 0 \\ 0 & & & \mu & 0 \end{bmatrix}.$$

In brief, Λ_1 indicates which transitions of $\{Q(t)\}$ will count as increments of $\{D(t)\}$.

The *fundamental matrix* $\Lambda^- := (\mathbf{1}\boldsymbol{\pi} - \Lambda)^{-1}$ is a generalized inverse (that is, a matrix X such that

$$(-\Lambda)X(-\Lambda) = (-\Lambda), \tag{5}$$

see Hunter (1982)) for the negative of the generator Λ of a continuous-time Markov chain. Generalized inverses are not uniquely determined by (5); however, by specifying other relationships, specific classes of generalized inverses can be defined. The *deviation matrix*

$$\Lambda^\sharp := \int_0^\infty (e^{\Lambda t} - \mathbf{1}\boldsymbol{\pi})\, dt$$

is the so-called *group inverse* (or *Drazin inverse*) of $-\Lambda$, which, by definition, is the unique solution of (5), $X(-\Lambda)X = X$, and $(-\Lambda)X = X(-\Lambda)$. The deviation matrix can be interpreted as a measure of the total deviation from the limiting probabilities, see for instance Coolen-Schrijner and van Doorn (2002) as well as Koole and Spieksma (2001). The fundamental matrix Λ^- and the deviation matrix $\Lambda^\sharp$ are related by the expression

$$\Lambda^- = \Lambda^\sharp + \mathbf{1}\boldsymbol{\pi}. \tag{6}$$

For finite state space continuous-time Markov chains, such as the M/M/1/K queue, the fundamental matrix and the deviation matrix always exist. The deviation matrix satisfies the properties

$$\Lambda^\sharp \mathbf{1} = \mathbf{0}, \tag{7}$$

$$\boldsymbol{\pi}\Lambda^\sharp = \mathbf{0}, \tag{8}$$

$$\Lambda^\sharp\Lambda = \Lambda\Lambda^\sharp = \mathbf{1}\boldsymbol{\pi} - I,$$

as well as

$$\Lambda^\sharp_{i,j} = \pi_j\,(m^{\mathrm{e}}_j - m_{i,j}),$$

where $m_{i,j}$ is the mean first entrance time from state i to state j, and m^{e}_j is the mean first entrance time to state j from the stationary distribution, that is,

$$m_{i,j} = \mathbb{E}[\inf\{t : Q(t) = j\} | Q(0) = i], \quad m^{\mathrm{e}}_j = \sum_{i=0}^{K} \pi_i m_{i,j},$$

see Coolen-Schrijner and van Doorn (2002). Note that we take the indices of the matrices and vectors of size $K+1$ used here to run on range $\{0, \ldots, K\}$.

2.1. M/M/1/K queue: explicit formulas related to the deviation matrix

As will be evident below, we are particularly interested in the bottom left entry of the deviation matrix and that of its square,

$$\overline{d}_{\mathrm{v}} = \Lambda^\sharp_{K,0} = \int_0^\infty (\mathbb{P}(Q(t) = 0 | Q(0) = K) - \pi_0)\, dt = \pi_0(m^{\mathrm{e}}_0 - m_{K,0}),$$

$$\overline{d}_{\mathrm{b}} = (\Lambda^\sharp\Lambda^\sharp)_{K,0} = \pi_0 \sum_{j=0}^{K} \pi_j\,(m^{\mathrm{e}}_j - m_{K,j})(m^{\mathrm{e}}_0 - m_{j,0}).$$

Finding explicit expressions for these quantities is tedious yet possible for the M/M/1/K queue.

Lemma 1. *For the M/M/1/K queue length continuous-time Markov chain, the $(K,0)$ elements of the deviation matrix and its square are*

$$\overline{d_{\mathrm{v}}} = \begin{cases} -\mu^{-1}\,\dfrac{K(1-\rho)(1+\rho^{K+1})-2\rho(1-\rho^K)}{(1-\rho)\,(1-\rho^{K+1})^2}, & \rho\neq 1,\\ -\mu^{-1}\,\dfrac{K(K+2)}{6(K+1)}, & \rho=1. \end{cases}$$

$$\overline{d}_{\mathrm{b}} = \begin{cases} -\mu^{-2}\left\{\dfrac{[6(1+\rho^2)(1+K)^2-4\rho(1+6K+3K^2)]\rho^{K+1}+\rho^2(2+3K+K^2)(1+\rho^{2K})}{2(1-\rho)^3(1-\rho^{K+1})^3}\right.\\ \qquad\left.-\dfrac{2\rho(3+2K+K^2)(1+\rho^{2(K+1)})-K(1+K)(1+\rho^{2(K+2)})}{2(1-\rho)^3(1-\rho^{K+1})^3}\right\}, & \rho\neq 1,\\ -\mu^{-2}\,\dfrac{7K^4+28K^3+37K^2+18K}{360\,(K+1)}, & \rho=1. \end{cases}$$

Proof. For the M/M/1/K queue, a standard application of "first step analysis" leads to the following recurrence equations for $m_{i,j}$:

$$m_{i,j} = \begin{cases} 0, & i=j,\\ \lambda^{-1}+m_{1,j}, & i=0, j\neq i,\\ \mu^{-1}+m_{K-1,j}, & i=K, j\neq i,\\ (\lambda+\mu)^{-1}+\dfrac{\lambda}{\lambda+\mu}m_{i+1,j}+\dfrac{\mu}{\lambda+\mu}m_{i-1,j}, & \text{otherwise.} \end{cases}$$

When $\rho\neq 1$, the solution is

$$m_{i,j} = \begin{cases} \mu^{-1}\left(\dfrac{\rho^{-j}-\rho^{-i}}{(1-\rho)^2}+\dfrac{i-j}{1-\rho}\right), & 0\le i\le j,\\ \mu^{-1}\left(\rho^{K+1}\dfrac{\rho^{-j}-\rho^{-i}}{(1-\rho)^2}+\dfrac{i-j}{1-\rho}\right), & j\le i\le K, \end{cases}$$

and, when $\rho=1$, the solution is

$$m_{i,j} = \begin{cases} \mu^{-1}\,\dfrac{j\,(j+1)-i\,(i+1)}{2}, & 0\le i\le j,\\ \mu^{-1}\,\dfrac{(K-j)\,[(K-j)+1]-(K-i)\,[(K-i)+1]}{2}, & j\le i\le K. \end{cases}$$

Averaging over $\boldsymbol{\pi}$ we get,

$$m_j^{\mathrm{e}} = \begin{cases} \mu^{-1}\,\dfrac{\rho^{-j}-(1+2j)(1-\rho)-[1+2(K-j)](1-\rho)\rho^{K+1}-\rho^{2(K+1)-j}}{(1-\rho)^2(1-\rho^{K+1})}, & \rho\neq 1,\\ \mu^{-1}\left(j^2-Kj+\dfrac{K^2}{3}+\dfrac{K}{6}\right), & \rho=1. \end{cases}$$

Combining the above we get the desired results. □

We note that related results were found for a discrete time version of the queue in Koole (1998) and in Koole and Spieksma (2001).

2.2. M/M/1/K queue: the stationary case

When the queue length process $\{Q(t)\}$ is stationary, the Markovian Point Process $\{D(t)\}$ is a (time) stationary point process (see Asmussen, 2003). In this case, the asymptotic variance rate, $\bar{v}$, and the y-intercept, $\bar{b}_{\mathrm{e}}$, are respectively given by

$$\bar{v} = \boldsymbol{\pi}\Lambda_1\mathbf{1} - 2(\boldsymbol{\pi}\Lambda_1\mathbf{1})^2 + 2\boldsymbol{\pi}\Lambda_1\Lambda^{-}\Lambda_1\mathbf{1}, \tag{9}$$

$$\bar{b}_{\mathrm{e}} = 2(\boldsymbol{\pi}\Lambda_1\mathbf{1})^2 - 2\boldsymbol{\pi}\Lambda_1\Lambda^{-}\Lambda^{-}\Lambda_1\mathbf{1}, \tag{10}$$

see for instance Narayana and Neuts (1992), Asmussen (2003) or the summary in Nazarathy and Weiss (2008). By substituting (6) into (9) and (10) we obtain a simpler expression for $\bar{v}$ and $\bar{b}_{\mathrm{e}}$ in terms of the deviation matrix:

$$\bar{v} = \boldsymbol{\pi}\Lambda_1\mathbf{1} + 2\boldsymbol{\pi}\Lambda_1\Lambda^{\sharp}\Lambda_1\mathbf{1} = \lambda^* + 2\boldsymbol{\pi}\Lambda_1\Lambda^{\sharp}\Lambda_1\mathbf{1}, \tag{11}$$

$$\bar{b}_{\mathrm{e}} = -2\boldsymbol{\pi}\Lambda_1\Lambda^{\sharp}\Lambda^{\sharp}\Lambda_1\mathbf{1}, \tag{12}$$

where

$$\lambda^* = \lim_{t\to\infty}\frac{\mathbb{E}[D(t)]}{t} = \boldsymbol{\pi}\Lambda_1\mathbf{1} = \begin{cases} \mu\rho\,\dfrac{1-\rho^K}{1-\rho^{K+1}}, & \text{for } \rho\neq 1,\\ \mu\,\dfrac{K}{K+1}, & \text{for } \rho=1. \end{cases}$$

We can go further in the simplification of the expressions by using (7) and (8). Let $\boldsymbol{e}_i$ denote the column vector of which the only nonzero entry is the entry corresponding to state i, which is equal to 1 ($0\le i\le K$). First, observe that $\boldsymbol{\pi}\Lambda_1$ and $\Lambda_1\mathbf{1}$ take simple forms in our case:

$$\boldsymbol{\pi}\Lambda_1 = \begin{cases} \mu\rho\boldsymbol{\pi}-\rho^{K+1}\,\dfrac{1-\rho}{1-\rho^{K+1}}\,\mu\boldsymbol{e}_K', & \rho\neq 1,\\ \mu\boldsymbol{\pi}-\dfrac{1}{K+1}\,\mu\boldsymbol{e}_K', & \text{for } \rho=1. \end{cases}, \qquad \Lambda_1\mathbf{1} = \mu(\mathbf{1}-\boldsymbol{e}_0).$$

Then, since $\boldsymbol{\pi}$ and $\mathbf{1}$ are respectively the left and right eigenvectors of $\Lambda^{\sharp}$ corresponding to the eigenvalue 0, we obtain

$$\boldsymbol{\pi}\Lambda_1\Lambda^{\sharp}\Lambda_1\mathbf{1} = \begin{cases} \mu^2\rho^{K+1}\,\dfrac{1-\rho}{1-\rho^{K+1}}\,\overline{d}_{\mathrm{v}}, & \text{for } \rho\neq 1,\\ \mu^2\,\dfrac{1}{K+1}\,\overline{d}_{\mathrm{v}}, & \text{for } \rho=1. \end{cases}$$

We thus obtain

$$\bar{v} = \begin{cases} \lambda^*+2\mu^2\rho^{K+1}\,\dfrac{1-\rho}{1-\rho^{K+1}}\,\overline{d}_{\mathrm{v}}, & \rho\neq 1,\\ \lambda^*+\dfrac{2\mu^2}{K+1}\,\overline{d}_{\mathrm{v}}, & \rho=1, \end{cases}$$

and similarly,

$$\bar{b}_{\mathrm{e}} = \begin{cases} -2\mu^2\,\rho^{K+1}\,\dfrac{1-\rho}{1-\rho^{K+1}}\,\overline{d}_{\mathrm{b}}, & \rho\neq 1,\\ -2\mu^2\,\dfrac{1}{K+1}\,\overline{d}_{\mathrm{b}}, & \rho=1. \end{cases}$$

Combining the above with the results of Lemma 1, and manipulating the expressions, we obtain our main result for M/M/1/K queues:

Proposition 2. *For the stationary M/M/1/K queue, $v(t)=\bar{v}t+\bar{b}_{\mathrm{e}}+o(1)$ where*

$$\bar{v} = \begin{cases} \lambda\,\dfrac{(1+\rho^{K+1})(1-(1+2K)\rho^K(1-\rho)-\rho^{2K+1})}{(1-\rho^{K+1})^3}, & \rho\neq 1\\ \lambda\left(\dfrac{2}{3}-\dfrac{3K+2}{3(K+1)^2}\right), & \rho=1 \end{cases},$$

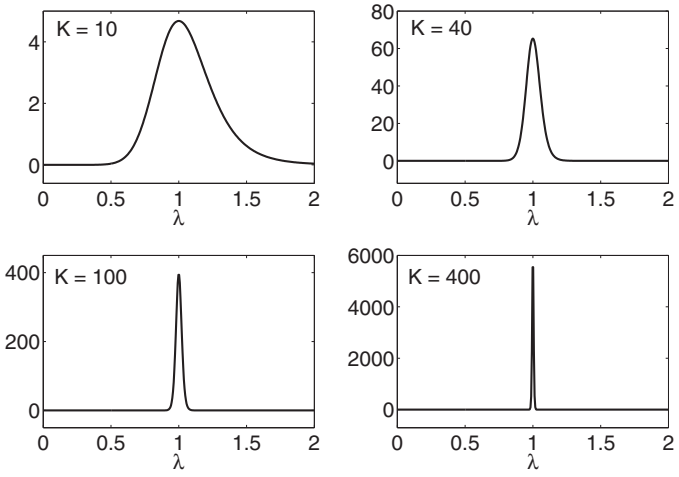

Fig. 1. The intercept term $\bar{b}_e$ as a function of λ when $\mu = 1$ and for $K = 10, 40, 100, 400$.

and

$$\bar{b}_e = \begin{cases} \rho^{K+1}\left\{\dfrac{(6(1+\rho^2)(1+K)^2 - 4\rho(1+6K+3K^2))\,\rho^{K+1} + \rho^2(2+3K+K^2)(1+\rho^{2K})}{(1-\rho)^2\left(1-\rho^{K+1}\right)^4}\right. \\ \qquad \left. - \dfrac{2\rho(3+2K+K^2)(1+\rho^{2(K+1)}) - K(1+K)(1+\rho^{2(K+2)})}{(1-\rho)^2(1-\rho^{K+1})^4}\right\}, & \rho \neq 1, \\ \dfrac{7K^4+28K^3+37K^2+18K}{180\,(K+1)^2}, & \rho = 1. \end{cases}$$

respectively.

Here are some observations:

- With the exception of $\bar{b}_e$ for $\rho \neq 1$, all of the expressions in Proposition 2 appeared in Nazarathy and Weiss (2008). Yet, while working on Nazarathy and Weiss (2008), the authors were not able to obtain $\bar{b}_e$ when $\rho \neq 1$, as obtained here.
- We illustrate the intercept term for different values of K and λ in Fig. 1. It is straightforward to see that

$$\lim_{K\to\infty} \bar{b}_e = \begin{cases} 0, & \rho \neq 1, \\ \infty, & \rho = 1, \end{cases}$$

  and further, for $\rho = 1$, $\bar{b}_e = O(K^2)$.
- It is insightful to see the role of $\bar{d}_v$ and $\bar{d}_b$ in the above derivations. In fact, the spikes in $\bar{v}$ and $\bar{b}$ that occur at $\rho \approx 1$ are are because of similar spikes in $\bar{d}_v$ and $\bar{d}_b$.

2.3. Some further useful results on Markovian Point Processes

Our derivation of $\bar{v}$ and $\bar{b}_e$ above is based on (9) and (10) respectively, or alternatively on their deviation matrix based forms, (11) and (12). On route to calculating additional performance measures for the M/M/1/K queue, we first derive some further results for Markovian Point Processes, which to the best of our knowledge have not appeared elsewhere. These results are of independent interest.

Consider an arbitrary Markovian Point Process with an $n \times n$ irreducible generator matrix $\Lambda = \Lambda_0 + \Lambda_1$, where Λ_1 is the event intensity matrix and Λ_0 is assumed to be non-singular. Such a process corresponds to a two-dimensional Markov chain $\{(N(t), \varphi(t)), t \geq 0\}$, where $N(t)$ denotes the number of events in the interval $[0, t]$ and $\varphi(t)$ denotes the *phase* at time t, taking values in $\{1, \ldots, n\}$. We assume that $N(0) = 0$ almost surely and denote by $\boldsymbol{\theta}$ the distribution of $\varphi(0)$. Further, we denote by $\boldsymbol{\pi}$ the stationary distribution corresponding to Λ, or equivalently to the phase process $\{\varphi(t)\}$.

If $\boldsymbol{\theta} = \boldsymbol{\pi}$ then $N(t)$ is a time-stationary point process. This implies that for any sequence of intervals $(t_1, s_1), \ldots, (t_\ell, s_\ell)$ and for any τ,

$$[N(s_1) - N(t_1), \ldots, N(s_\ell) - N(t_\ell)] =^d [N(s_1+\tau) - N(t_1+\tau), \ldots, N(s_\ell+\tau) - N(t_\ell+\tau)],$$

where the equality is in distribution (see Asmussen, 2003, chap. XI, Proposition 1.2).

Another interesting initial distribution is $\boldsymbol{\alpha} = \boldsymbol{\pi}\Lambda_1/(\boldsymbol{\pi}\Lambda_1\mathbf{1})$. This is the invariant distribution of a discrete time jump chain, $\{Y_k\}$ where Y_k is the value of $\varphi(t)$ just after the kth arrival. The probability transition matrix of this Markov chain is $-\Lambda_0^{-1}\Lambda_1$. As shown in Asmussen (2003, chap. XI, Proposition 1.4), setting $\boldsymbol{\theta} = \boldsymbol{\alpha}$ makes $\{N(t)\}$ an event-stationary point process. That is, if T_k denotes the time interval between the $(k-1)$st and the kth event in the Markovian Point Process, then the joint distribution of $(T_k, T_{k+1}, \ldots, T_{k+\ell})$ is the same as the joint distribution of $(T_{k'}, T_{k'+1}, \ldots, T_{k'+\ell})$ for all integer k, k', ℓ.

Since $\boldsymbol{\theta}$ affects the point process in such a manner, it is natural to ask what is its effect on $v(t)$ and related quantities. We now have the following:

Proposition 3. *For an arbitrary Markovian Point Process with initial distribution $\boldsymbol{\theta}$, $Var(N(t)) = \bar{v}t + \bar{b}_\theta + o(1)$, where the y-intercept is given by*

$$\bar{b}_\theta = \bar{b}_e - (2\boldsymbol{\pi}\Lambda_1\mathbf{1}\boldsymbol{\theta}(\Lambda^\sharp)^2\Lambda_1\mathbf{1} - 2\boldsymbol{\theta}\Lambda^\sharp\Lambda_1\Lambda^\sharp\Lambda_1\mathbf{1} + (\boldsymbol{\theta}\Lambda^\sharp\Lambda_1\mathbf{1})^2), \tag{13}$$

and $\bar{v}$ and $\bar{b}_e$ are respectively given by (11) and (12).

Proof. The variance curve of an arbitrary Markovian Point Process with initial distribution $\boldsymbol{\theta}$ is given by

$$\mathrm{Var}(N(t)) = \boldsymbol{\theta}M_2(t)\mathbf{1} + \boldsymbol{\theta}M_1(t)\mathbf{1} - (\boldsymbol{\theta}M_1(t)\mathbf{1})^2, \tag{14}$$

where $M_1(t)$ and $M_2(t)$ denote the matrices of the first two factorial moments of the number of events in a non-stationary Markovian Point Process, that is,

$$[M_1(t)]_{ij} = \mathbb{E}[N(t)\,\mathbb{1}_{\{\varphi(t)=j\}}|\varphi(0) = i]$$
$$[M_2(t)]_{ij} = \mathbb{E}[N(t)(N(t)-1)\,\mathbb{1}_{\{\varphi(t)=j\}}|\varphi(0) = i].$$

Narayana and Neuts (1992) showed that $M_1(t)$ has a linear asymptote in that there exist constant matrices A_0 and A_1 such that

$$M_1(t) = A_0 t + A_1 + O(e^{-\eta t}t^{2r-1}) \quad \text{as } t \to \infty, \tag{15}$$

where $-\eta$ is the real part of η^*, the non-zero eigenvalue of Λ with maximum real part, and r is the multiplicity of η^*. Similarly, $M_2(t)$ has a quadratic asymptote in that there exist constant matrices B_0, B_1 and B_2 such that

$$M_2(t) = B_0 t^2 + 2B_1 t + 2B_2 + O(e^{-\eta t}t^{3r-1}) \quad \text{as } t \to \infty. \tag{16}$$

The expressions for the coefficient matrices A_0, A_1, B_0, B_1, and B_2 given in Narayana and Neuts (1992) are in terms of the fundamental

matrix Λ^-. After rewriting them in terms of the deviation matrix via the relation (6), we see that

$$
\begin{aligned}
A_0 &= (\boldsymbol{\pi}\Lambda_1\mathbf{1})\,\mathbf{1}\boldsymbol{\pi},\\
A_1 &= \Lambda^\sharp\Lambda_1\mathbf{1}\boldsymbol{\pi} + \mathbf{1}\boldsymbol{\pi}\Lambda_1\Lambda^\sharp,\\
B_0 &= (\boldsymbol{\pi}\Lambda_1\mathbf{1})^2\,\mathbf{1}\boldsymbol{\pi},\\
B_1 &= (\boldsymbol{\pi}\Lambda_1\mathbf{1})\,\Lambda^\sharp\Lambda_1\mathbf{1}\boldsymbol{\pi} + (\boldsymbol{\pi}\Lambda_1\mathbf{1})\,\mathbf{1}\boldsymbol{\pi}\Lambda_1\Lambda^\sharp + (\boldsymbol{\pi}\Lambda_1\Lambda^\sharp\Lambda_1\mathbf{1})\,\mathbf{1}\boldsymbol{\pi},\\
B_2 &= -\mathbf{1}\boldsymbol{\pi}(\boldsymbol{\pi}\Lambda_1(\Lambda^\sharp)^2\Lambda_1\mathbf{1}) + \Lambda^\sharp\Lambda_1\mathbf{1}\boldsymbol{\pi}\Lambda_1\Lambda^\sharp - (\boldsymbol{\pi}\Lambda_1\mathbf{1})(\Lambda^\sharp)^2\Lambda_1\mathbf{1}\boldsymbol{\pi}\\
&\quad - (\boldsymbol{\pi}\Lambda_1\mathbf{1})\mathbf{1}\boldsymbol{\pi}\Lambda_1(\Lambda^\sharp)^2 + \mathbf{1}\boldsymbol{\pi}\Lambda_1\Lambda^\sharp\Lambda_1\Lambda^\sharp + \Lambda^\sharp\Lambda_1\Lambda^\sharp\Lambda_1\mathbf{1}\boldsymbol{\pi}.
\end{aligned}
$$

By injecting (15) and (16) in terms of $\Lambda^\sharp$ into (14), and making a few simplifications, we obtain the fact that $\mathrm{Var}(N(t))$ has a linear asymptote whose intercept term is

$$
\begin{aligned}
\bar{b}_\theta &= -2\boldsymbol{\pi}\Lambda_1\Lambda^\sharp\Lambda^\sharp\Lambda_1\mathbf{1} - 2\boldsymbol{\pi}\Lambda_1\mathbf{1}\boldsymbol{\theta}(\Lambda^\sharp)^2\Lambda_1\mathbf{1} + 2\boldsymbol{\theta}\Lambda^\sharp\Lambda_1\Lambda^\sharp\Lambda_1\mathbf{1}\\
&\quad - (\boldsymbol{\theta}\Lambda^\sharp\Lambda_1\mathbf{1})^2,
\end{aligned}
$$

which proves the theorem. □

Here are some observations:

- For $\boldsymbol{\theta} = \boldsymbol{\pi}$, the correction term $\bar{b}_\theta - \bar{b}_\mathrm{e}$ vanishes as expected.
- The correction term does not depend only on the variance of the initial distribution $\boldsymbol{\theta}$, as we show to be the case for the M/G/1 queue (see Proposition 7). Indeed, consider $n = 3$, $\boldsymbol{\theta}_1 = [1, 0, 0]$ and $\boldsymbol{\theta}_2 = [0, 1, 0]$ which have the same variance equal to zero; however it is easy to find an example of a Markovian Point Process for which the value of $\bar{b}_\theta - \bar{b}_\mathrm{e}$ is different for the two initial distributions.
- In the specific case where $\Lambda_1\mathbf{1} = \beta\mathbf{1}$ (where β is a constant), the correction term $\bar{b}_\theta - \bar{b}_\mathrm{e}$ vanishes for all initial distributions $\boldsymbol{\theta}$ because of the property (7). This shows that $\boldsymbol{\theta} = \boldsymbol{\pi}$ is a sufficient but not necessary condition for having $\bar{b}_\theta = \bar{b}_\mathrm{e}$.

A further performance measure of interest is the asymptotic covariance between the number of points and the phase of a Markovian Point Process. As shown in the following proposition, the deviation matrix also plays a role in this asymptotic quantity.

Proposition 4. *Let $\{N(t)\}$ and $\{\varphi(t)\}$ be the number of points in $[0, t]$ and the phase at time t of a Markovian Point Process with initial phase distribution $\boldsymbol{\theta}$. Then,*

$$
\lim_{t\to\infty} Cov(N(t), \varphi(t)) = \sum_{i=1}^{n} i(\boldsymbol{\pi}\Lambda_1\Lambda^\sharp)_i - (\sum_{i=1}^{n} i\,\pi_i)\,\boldsymbol{\theta}\Lambda^\sharp\Lambda_1\mathbf{1}.
$$

Further, in the time-stationary case ($\boldsymbol{\theta} = \boldsymbol{\pi}$), the term with the second sum vanishes.

Proof. For a Markovian Point Process with initial distribution $\boldsymbol{\theta}$, define $M_i^\theta(t) = \mathbb{E}[N(t)\,\mathbb{1}_{\{\varphi(t)=i\}}]$, and $\boldsymbol{M}^\theta(t) = [M_1^\theta(t), \ldots, M_n^\theta(t)]$. From Asmussen (2003, chap. XI, Proposition 1.7), we have

$$
\boldsymbol{M}^\theta(t) = \boldsymbol{\theta}\int_0^t e^{\Lambda u}\,\Lambda_1\,e^{\Lambda(t-u)}\,du.
$$

Let us define the *transient* deviation matrix as

$$
\Lambda^\sharp(t) = \int_0^t (e^{\Lambda u} - \mathbf{1}\boldsymbol{\pi})\,du = \int_0^t e^{\Lambda u}\,du - \mathbf{1}\boldsymbol{\pi}\,t,
$$

so that $\Lambda^\sharp = \lim_{t\to\infty}\Lambda^\sharp(t)$. With this,

$$
\mathbb{E}[N(t)] = \boldsymbol{M}^\theta(t)\mathbf{1} = \boldsymbol{\theta}\int_0^t e^{\Lambda u}du\,\Lambda_1\mathbf{1} = \boldsymbol{\pi}\Lambda_1\mathbf{1}t + \boldsymbol{\theta}\Lambda^\sharp(t)\Lambda_1\mathbf{1}.
$$

Next, note that $\mathbb{E}[N(t)\,\varphi(t)] = \sum_{i=1}^n i\,M_i^\theta(t)$. Therefore, since $\mathbb{E}[\varphi(t)] = \sum_{i=1}^n i\,(\boldsymbol{\theta}e^{\Lambda t})_i$, we obtain

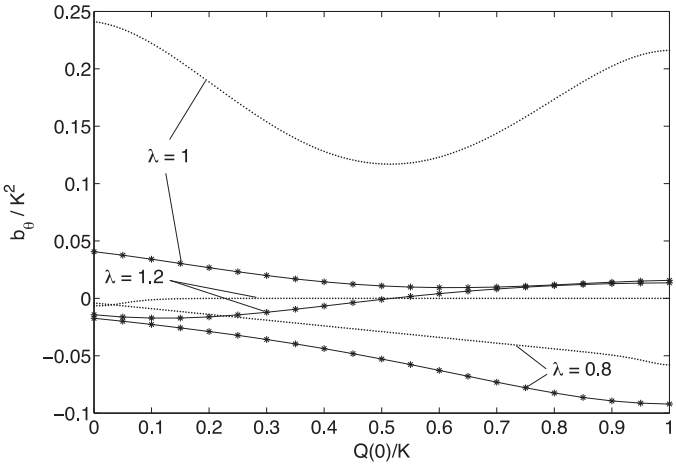

Fig. 2. The scaled intercept term $\bar{b}_\theta/K^2$ when $\mu = 1$ and $\boldsymbol{\theta} = \mathbf{e}_i'$ (that is, $Q(0) = i$ almost surely) as a function of $i/K \in \{0, 1/K, 2/K, \ldots, 1\}$, for $\lambda = 0.8$, $\lambda = 1$, and $\lambda = 1.2$, for $K = 20$ (stars) and $K = 200$ (dots).

$$
\begin{aligned}
\mathrm{Cov}(N(t), \varphi(t)) &= \mathbb{E}[N(t)\,\varphi(t)] - \mathbb{E}[N(t)]\,\mathbb{E}[\varphi(t)]\\
&= \sum_{i=1}^{n} i\left\{\left[\boldsymbol{\theta}e^{\Lambda t}\int_0^t e^{-\Lambda u}\Lambda_1 e^{\Lambda u}du\right]_i\right.\\
&\quad \left. - (\boldsymbol{\pi}\Lambda_1\mathbf{1}t + \boldsymbol{\theta}\Lambda^\sharp(t)\Lambda_1\mathbf{1})[\boldsymbol{\theta}e^{\Lambda t}]_i\right\}.
\end{aligned}
$$

Finally, we take $t \to \infty$ in the last expression, and we use the fact that $\lim_{t\to\infty}\boldsymbol{\theta}e^{\Lambda t} = \boldsymbol{\pi}$, $\boldsymbol{\pi}e^{-\Lambda u} = \boldsymbol{\pi}$, and $\int_0^t e^{\Lambda u}du = \Lambda^\sharp(t) + \mathbf{1}\boldsymbol{\pi}t$. After some algebraic simplifications, we obtain the statement of the proposition. □

2.4. M/M/1/K queue: arbitrary initial conditions

We now make use of Proposition 3 to investigate the intercept term $\bar{b}_\theta$ of $v(t)$ for an arbitrary M/M/1/K queue where the distribution of $Q(0)$ is $\boldsymbol{\theta}$.

In Fig. 2, we show the (scaled) value of $\bar{b}_\theta$ for the particular initial distributions $\boldsymbol{\theta} = \mathbf{e}_i'$, that is, for M/M/1/K queues which start with i customers at time $t = 0$ almost surely, for $0 \le i \le K$. We observe that when $\rho < 1$, $\bar{b}_\theta$ is a monotonically decreasing function of i, while when $\rho \ge 1$, $\bar{b}_\theta$ exhibits a minimum. We also observe that for $\rho > 1$, when K increases and $Q(0)/K \to 1$, $\bar{b}_\theta \to 0$.

In Fig. 3, we consider the behavior of the intercept term $\bar{b}_\theta$ in the event-stationary case, that is, when $\boldsymbol{\theta} = \boldsymbol{\alpha}$. The four graphs show the correction term $\bar{b}_\theta - \bar{b}_\mathrm{e}$ as a function of ρ (or, more precisely, as a function of λ for a fixed value of μ) for increasing values of K. We see that the curves have an interesting shape with two local minima centered around $\rho = 1$. As K increases, the dips become narrow and deep; the correction term converges to zero everywhere, except at $\rho = 1$ where further computation has shown that it decreases approximately linearly. As a consequence, the effect of event-stationarity on the intercept term becomes indistinguishable from the effect of time-stationarity as $K \to \infty$ for all values of ρ except in the balanced case.

Note that the explicit expressions for $\bar{b}_\theta$ when $\boldsymbol{\theta} = \mathbf{e}_i'$ or $\boldsymbol{\theta} = \boldsymbol{\alpha}$ can also be obtained, but since they are quite cumbersome and do not bring more information, we do not present these here.

2.5. M/M/1/K queue: asymptotic covariance

Making use of Proposition 4 and the particular structure of the vector $\boldsymbol{\pi}$ and the matrix Λ_1, in the stationary case, we can express

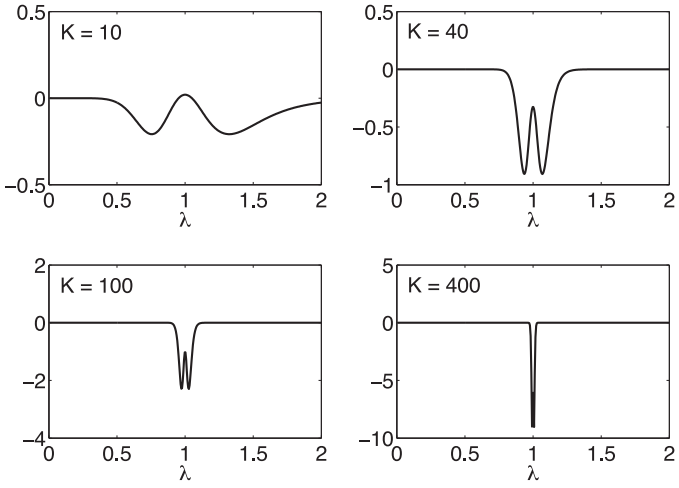

Fig. 3. The correction term $\bar{b}_\theta - \bar{b}_e$ when $\boldsymbol{\theta} = \boldsymbol{\alpha}$ as a function of λ when $\mu = 1$ and for $K = 10, 40, 100, 400$.

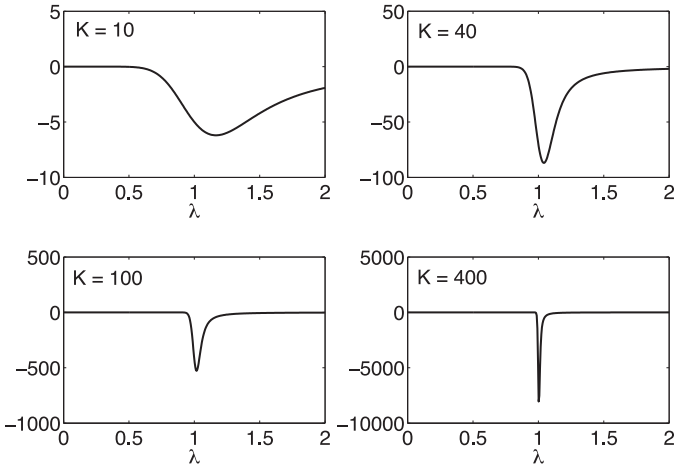

Fig. 4. The asymptotic covariance between $D(t)$ and $Q(t)$ as a function of λ when $\mu = 1$ and for $K = 10, 40, 100, 400$.

the asymptotic covariance between the number of departures and the queue size explicitly.

Corollary 5. *Consider the stationary M/M/1/K queue with output process* $\{D(t)\}$ *and queue level process* $\{Q(t)\}$. *Then,*

$$\lim_{t\to\infty} Cov(D(t), Q(t)) =$$
$$\begin{cases} \rho^{K+1}\left\{\dfrac{K^2(\rho-1)^2\,(1+3\rho^{K+1}) - 2\rho(\rho^K-1)(-2+\rho+\rho^{K+2})}{2(\rho-1)^2(\rho^{K+1}-1)^3}\right. \\ \qquad \left. + \dfrac{K(\rho-1)\,(-1+3\rho-7\rho^{K+1}+5\rho^{K+2})}{2(\rho-1)^2(\rho^{K+1}-1)^3}\right\}, & \rho \neq 1, \\ -K(K+2)/24, & \rho = 1. \end{cases}$$

Proof. We use Proposition 4 with $N(t) = D(t)$ and $\varphi(t) = Q(t)$, together with the fact that $(\boldsymbol{\pi}\Lambda_1\Lambda^\sharp)_i = -C\,\Lambda^\sharp_{Ki}$, where

$$C = \begin{cases} \mu\dfrac{\rho^{K+1}(1-\rho)}{1-\rho^{K+1}}, & \text{for } \rho \neq 1, \\ \dfrac{\mu}{K+1}, & \text{for } \rho = 1, \end{cases}$$

and $\Lambda^\sharp_{Ki} = \pi_i\,(m^e_i - m_{K,i})$. The entries $\Lambda^\sharp_{Ki}$ of the deviation matrix are then computed explicitly for $0 \le i \le K$ using the expressions for $\boldsymbol{\pi}$ and $m_{i,j}$ derived in the proof of Lemma 1. □

Note that Proposition 4 indicates that the difference between the stationary and the non-stationary cases is the correction term $-(\sum_i i\pi_i)\boldsymbol{\theta}\Lambda^\sharp\Lambda_1\mathbf{1}$, where $\sum_i i\pi_i = K/2$ for $\rho = 1$, and for $\rho \neq 1$,

$$\sum_{i=0}^{K} i\pi_i = \frac{\rho(1-(1+K)\rho^K + K\rho^{1+K})}{(1-\rho)(1-\rho^{1+K})}.$$

In Fig. 4, we illustrate the asymptotic covariance between $D(t)$ and $Q(t)$ in the stationary case, as a function of ρ and for increasing values of K. We see that the asymptotic covariance curves exhibit a similar behavior to (the negative of) those of the intercept term in the stationary case (see Fig. 1), but in the present case the curves are more skewed with respect to $\rho = 1$.

3. The M/G/1 queue

We now consider the departure process of the M/G/1 queue. In this case, the departure process $\{D(t)\}$ is generally not a Markovian Point Process, and the analysis is more complicated. Nevertheless we are able to obtain some partial results about the linear asymptote of $v(t)$. Our approach is to first assume the existence of a linear asymptote and then to find an elegant formula for the intercept term under this assumption, generalizing the expression for the M/M/1 queue in (4) for the stable case. Further we conjecture that our assumption holds when the third moment of the service time is finite.

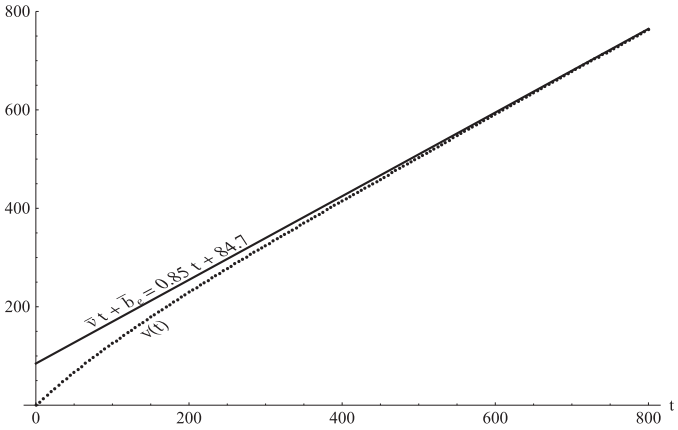

Fig. 5. The variance curve and its linear asymptote for a stationary M/G/1 queue with $\lambda = 0.85$, $\mu = 1$ and G following a log-normal distribution with $c^2 = 2$.

Denote the arrival rate by λ, the service time distribution by $G(\cdot)$ and its kth moment by g_k. In this case, $\mu = g_1^{-1}$, and we assume that $\rho = \lambda/\mu < 1$. The squared coefficient of variation and skewness coefficient are given by $c^2 = g_2/g_1^2 - 1$ and $\gamma = (2g_1^3 - 3g_1g_2 + g_3)/((g_2 - g_1^2)^{3/2})$ respectively.

Consider the output from a simulated numerical example in Fig. 5.[2] It depicts the variance curve and its linear asymptote for a stationary M/G/1 queue with $\lambda = 0.85$ and $\mu = 1$ ($\rho = 0.85$). The service time distribution is taken to be log-normal with $c^2 = 2$. This implies that $\gamma = 10/\sqrt{2}$. It is visually evident that for non-small t, $v(t) \approx \bar{v}t + \bar{b}_e$.

[2] We simulated 10^6 realizations of the queueing process, recording and estimating the variance of $D(t)$ over the grid $t = 0, 5, 10, \ldots, 800$. Prior to time $t = 0$ we simulated each realization for 3×10^4 units so as to begin in approximate steady-state. The simulation is coded in C to allow for efficient computation. During the simulation run, roughly 26×10^9 jobs were processed in the simulated M/G/1 queue.

Indeed, for the rest of this paper, we shall assume that such a linear asymptote exists. This is stated in the assumption below.

Assumption 1. *There exist $\bar{v}$ and $\bar{b}_\theta$ such that,*

$$v(t) = \bar{v}t + \bar{b}_\theta + o(1). \tag{17}$$

The $\bar{b}_e$ term in Fig. 5 was calculated from the formula in Proposition 6 below and is a function of μ, c^2 and γ. As is attested by the figure and by further extensive numerical experiments, we believe that such a term exists for all M/G/1 queues with $\rho \neq 1$ in which the service time distribution has a finite third moment, yet we have not been able to prove this.

Conjecture 1. *For $\lambda \neq \mu$, consider an M/G/1 queue operating under any work-conserving non-pre-emptive policy, in which the service time distribution has a finite third moment. Assume furthermore that the variance of the queue length at time 0 is finite. Then there exists a finite $\bar{b}_\theta$ such that (17) holds, with*

$$\bar{v} = \begin{cases} \lambda, & \lambda < \mu, \\ \mu c^2, & \lambda > \mu. \end{cases}$$

To get insight into the asymptotic variance rate, $\bar{v}$, in Conjecture 1, consider first the case where $\lambda > \mu$. In this case, after some time τ which is almost surely finite, the server never stops operating and thus for $t > \tau$, $\{D(t)\}$ is effectively a renewal process with asymptotic variance rate μc^2 (see Eq. (1)). To the best of our knowledge, this intuitive argument cannot easily be made rigorous.

For the case $\lambda < \mu$, consider the relation $D(t) = A(t) + Q(0) - Q(t)$, where $D(t)$ is the number of service completions during $[0, t]$, $\{A(t)\}$ is the Poisson arrival process, counting arrivals during $[0, t]$, and $Q(t)$ is the number of customers in the system at time t. Taking the variance of both sides of the last equation, observing that $A(t)$ is independent of $Q(0)$, dividing by t, and letting $t \to \infty$, we get

$$\lim_{t\to\infty} \frac{v(t)}{t} = \lim_{t\to\infty} \frac{\mathrm{Var}(A(t))}{t} + \lim_{t\to\infty} \frac{\mathrm{Var}(Q(0))}{t} + \lim_{t\to\infty} \frac{\mathrm{Var}(Q(t))}{t} - 2\lim_{t\to\infty} \frac{\mathrm{Cov}(Q(0), Q(t))}{t} - 2\lim_{t\to\infty} \frac{\mathrm{Cov}(A(t), Q(t))}{t},$$

whenever the limits exist. Now the first limit on the right hand side equals λ, the second limit vanishes by assumption, the third limit should vanish since the stationary variance is finite (due to a finite third moment – see Eq. (25)), the fourth limit should vanish since in fact $\mathrm{Cov}(Q(0), Q(t))$ vanishes (this is not trivial to establish in general, yet was communicated to us for the FCFS case through personal communication with Brian Fralix), and finally, for the fifth limit observe that

$$|\mathrm{Cov}(A(t), Q(t))| \leq \sqrt{\lambda t \mathrm{Var}(Q(t))} = \mathrm{O}(\sqrt{t}),$$

and thus the limit vanishes. This implies that the departure asymptotic variance equals the arrival asymptotic variance. To get insight into our belief of the importance of $g_3 < \infty$ for the existence of $\bar{b}_\theta$ (at least for the case $\lambda < \mu$), see the proof of Proposition 6 below.

Our focus for the rest of this section is on the stable case ($\lambda < \mu$) in which we are able to find explicit expressions for $\bar{b}_e$ and $\bar{b}_\theta$ (including $\bar{b}_0$).

3.1. M/G/1 queue: the stationary case

In Daley (1975) (see also Daley, 1976) Daley found the Laplace–Stieltjes transform (LST) of the variance curve for the stationary case. After some minor rearrangement, Daley's formula may be written as

$$v^*(s) = \int_0^\infty e^{-st} dv(t) = \frac{\lambda}{s} + b^*(s), \tag{18}$$

where,

$$b^*(s) = \frac{2\lambda}{s}\left(\frac{G^*(s)}{1-G^*(s)}\left(1 - \frac{s\Pi(\Gamma(s))}{s+\lambda(1-\Gamma(s))}\right) - \frac{\lambda}{s}\right), \tag{19}$$

and the LST exists for $\Re(s) > 0$. Here, $G^*(\cdot)$ is the Laplace–Stieltjes transform (LST) of $G(\cdot)$, $\Pi(\cdot)$ is the probability generating function of the stationary number of customers in the system, with

$$\Pi(z) = (1-\rho)\frac{(1-z)G^*(\lambda(1-z))}{G^*(\lambda(1-z)) - z}, \tag{20}$$

for $\Re(z) \leq 1$. Further, $\Gamma(s)$ is the LST of the busy period at s, with $\Re(s) > 0$. It is obtained as the minimal non-negative solution of

$$\Gamma(s) = G^*(s + \lambda(1 - \Gamma(s))). \tag{21}$$

For relevant standard queueing background see for example Prabhu (1998).

We now have the following:

Proposition 6. *Consider the stationary M/G/1 queue having $g_3 < \infty$. If Assumption 1 holds, then the intercept term in (17) is given by*

$$\bar{b}_e = L_e \frac{\rho}{(1-\rho)^2}, \quad \text{with}$$

$$L_e = \frac{(3c^4 - 4\gamma c^3 + 6c^2 - 1)\rho^3 + (4\gamma c^3 - 12c^2 + 4)\rho^2 + (6c^2 - 6)\rho}{6}.$$

Proof. Let $\tilde{b}(\cdot)$ be such that $v(t) = \lambda t + \tilde{b}(t)$. By Assumption 1, $\lim_{t\to\infty} \tilde{b}(t) = \bar{b}_e$. Since the LST of λt is λ/s and since the LST is a linear operator, we have from (18) that for $\Re(s) > 0$ the LST of $\tilde{b}(\cdot)$ is

$$b^*(s) = \int_0^\infty e^{-st} d\tilde{b}(t).$$

By the Final Value Theorem (see for example Widder, 1959 for a rigorous reference) we have

$$\lim_{s\to 0} b^*(s) = \bar{b}_e. \tag{22}$$

The remainder of the derivation deals with evaluation of the limit (22) by using (19) together with (20) in order to find $\bar{b}_e$. This is a combination of straightforward classic queuing calculations together with five applications of L'Hospital's rule. It requires us to evaluate the first three moments of the busy period by taking derivatives of (21) and setting $s \to 0$, and we get

$$b_1 = (1 + \lambda b_1)g_1,$$
$$b_2 = \lambda b_2 g_1 + (1 + \lambda b_1)^2 g_2,$$
$$b_3 = \lambda b_3 g_1 + 3\lambda(1 + \lambda b_1) b_2 g_2 + (1 + \lambda b_1)^3 g_3,$$

where $b_1 = -\Gamma'(0)$, $b_2 = \Gamma''(0)$ and $b_3 = -\Gamma'''(0)$. These values are well known and appear in many queueing texts, yet we present them here for completeness:

$$b_1 = \mu^{-1}\frac{1}{1-\rho},$$

$$b_2 = \frac{g_2}{(1-\rho)^3} = \mu^{-2}\frac{c^2+1}{(1-\rho)^3},$$

$$b_3 = \frac{g_3(1-\rho) + 3\lambda g_2^2}{(1-\rho)^5} = \mu^{-3}\frac{3c^4\rho + c^3\gamma(1-\rho) + 3c^2(1+\rho) + 2\rho + 1}{(1-\rho)^5}.$$ □

Here are some observations:

- If G follows an exponential distribution, then $c = 1$ and $\gamma = 2$, yielding $\bar{b}_e = 0$, as is expected since in this case $\{D(t)\}$ is a Poisson process.

- L_e (and thus $\bar{b}_e$) is monotone increasing in both c and γ.
- As $\rho \to 1$, $L_e \to (c^4 - 1)/2$. This gives some insight into the form of the variance curve of heavy traffic systems, showing how the squared coefficient of variation of G plays a role when $\rho \approx 1$: When $c^2 > 1$ the intercept term is positive, and when $c^2 < 1$ the intercept term is negative. Further, as typical for heavy traffic systems, only the first two moments of the service time play a role. The skewness coefficient γ does not matter when $\rho \approx 1$.

3.2. M/G/1 queue: on a conjecture by Daley

In the M/M/1 queue case, $v^*(s)$ in (18) is equal to λ/s, which corresponds to the variance curve $v(t) = \lambda t$. This is expected since, for the stationary M/M/1 queue, $\{D(t)\}$ is a Poisson process. In Daley (1975), Daley conjectures that the reverse implication is also true:

Having $v(t) = \lambda t$ implies that the service time distribution is exponential.

Restated in terms of LSTs using (18), the conjecture is that a $b^*(s)$ of the form (19), where $G^*(s)$ is the LST of a probability distribution, can be equal to zero only if $G^*(s) = 1/(1 + sg_1)$ with $g_1 > 0$. Proving Daley's conjecture would strengthen a result by Finch (1959), stating that all *stationary* M/G/1 queues with an Poisson output process are M/M/1 queues.

Our expression for $\bar{b}_e$ in Proposition 6 gives a necessary condition for $v(t) = \lambda t$:

$$v(t) = \lambda t \quad \textit{only if} \quad L_e = 0. \tag{23}$$

At this point it is tempting to believe that if $L_e = 0$ (i.e. $\bar{b}_e = 0$) then $v(t) = \lambda t$. This could then be used to disprove Daley's conjecture, since L_e only depends on the first three moments of G and it is known that the exponential distribution is not characterized (within the class of non-negative distributions) by the first three moments. For example, consider a mixture of an exponential random variable with point masses at $1/2$, $3/2$, $5/2$ and $9/2$, with LST:

$$G^*(s) = \frac{1}{384}\left(192\frac{1}{1+s} + 147e^{-\frac{1}{2}s} + 8e^{-\frac{3}{2}s} + 30e^{-\frac{5}{2}s} + 7e^{-\frac{9}{2}s}\right).$$

Elementary calculation yields $-G^{*\prime}(0) = 1, G^{*\prime\prime}(0) = 2, -G^{*\prime\prime\prime}(0) = 6$, as is the case for a unit mean exponential distribution, and thus $L_e = 0$. However, via numerical evaluation of $b^*(s)$ for the case where $\lambda = 3/4$ and $\mu = 1$, we verified that $b^*(s) \neq 0$ and thus $v(t) \neq \lambda t$, but rather, $v(t) = \lambda t + o(1)$, where the $o(1)$ term is not identically 0. Note that in this case, $\Gamma(s)$ was found by iterating (21) for fixed s over a fine grid of s. Thus we conclude that the necessary condition (23) is not sufficient. Daley's conjecture remains open.

3.3. M/G/1 queue: arbitrary initial conditions

We are now able to use the steady state intercept term to obtain the intercept term for a system with an arbitrary distribution of the initial state.

Proposition 7. *Consider the M/G/1 queue with $\rho < 1$, $g_3 < \infty$ and an arbitrary initial distribution of the number of customers having a finite variance. If Assumption 1 holds, then the intercept term in (17) is given by*

$$\bar{b}_\theta = \sigma_0^2 - \sigma_\pi^2 + \bar{b}_e,$$

where $\sigma_0^2 = Var(Q(0))$, σ_π^2 is the steady state variance of the number of customers in the system, and $\bar{b}_e$ is as in Proposition 6.

Proof. We use a coupling argument. Consider the following two systems 0 and θ under the same sample path of the arrival process and service times. System 0 starts empty and system θ starts with $Q(0)$ customers in the system. Operate system θ by giving low priority to the initial $Q(0)$ customers, that is, these customers are served only when there are no other customers in system and pre-empted if an arrival occurs during their service time. This implies that the first $Q(0)$ customers of System θ are being served only at times that coincide with idle periods of System 0. After a finite time T, the trajectories of the queue lengths of the two systems coincide. Thus for $t \geq T$, $D_\theta(t) = D_0(t) + Q(0)$, where $D_i(\cdot)$ denotes the relevant output counting process. This yields $\lambda t + \bar{b}_\theta = \lambda t + \bar{b}_0 + o(1) + \mathrm{Var}(Q(0))$. Taking $t \to \infty$ we obtain,

$$\bar{b}_\theta = \bar{b}_0 + \mathrm{Var}(Q(0)). \tag{24}$$

Now selecting $Q(0)$ to be distributed according to the steady state distribution we get from (24), $\bar{b}_0 = \bar{b}_e - \sigma_\pi^2$. Applying this again in (24) we obtain the result. Note that a similar coupling argument also holds for the non-preemptive case. □

As is well known, σ_π^2 can be obtained directly from $\Pi(\cdot)$, yet it is a cumbersome calculation. We state it here for completeness:

$$\sigma_\pi^2 = \left(\left(\frac{1}{4}c^4 - \frac{1}{3}\gamma c^3 + \frac{1}{2}c^2 - \frac{1}{12}\right)\rho^3 + \left(\frac{1}{3}\gamma c^3 - \frac{3}{2}c^2 + \frac{5}{6}\right)\rho^2 + \left(\frac{3}{2}c^2 - \frac{3}{2}\right)\rho + 1\right)\frac{\rho}{(1-\rho)^2}. \tag{25}$$

As a result of the above, the intercept term for a system that starts empty is:

$$\bar{b}_0 = -(1 - L_0)\frac{\rho}{(1-\rho)^2}, \quad \text{with}$$

$$L_0 = \frac{(3c^4 - 4\gamma c^3 + 6c^2 - 1)\rho^3 + (4\gamma c^3 - 6c^2 - 2)\rho^2 + (-6c^2 + 6)\rho}{12}.$$

Here are some observations:

- In the M/M/1 queue, $L_0 = 0$ and thus $\bar{b}_0 = -\rho/(1-\rho)^2$. As expected, this is in agreement with the case $\rho < 1$ in (4).
- As $\rho \to 1$, $L_0 \to (c^2 - 1)^2/4$. This implies that in heavy-traffic, c^2 plays a similar role with respect to the sign of the intercept term as it did in the stationary case: In this case when $c^2 > 3$ the intercept term is positive, and when $c^2 \leq 3$ the intercept term is negative. Compare with the remarks following Proposition 6.

4. Conclusion

In going through the detailed Markovian Point Process derivations for M/M/1/K queues, we have illustrated how asymptotic quantities such as $\bar{b}$ may be obtained explicitly. The key is to have explicit expressions for mean hitting times in the underlying Markov chain. By using the deviation matrix we have gained some further insight into the BRAVO effect. In plotting the graphs of $\bar{d}_v$ and $\bar{d}_b$, we observe that spikes occur when $\lambda \approx \mu$ in a similar fashion to the BRAVO effect.

For the M/G/1 queue, the formal calculations are of a different flavor, but are also generally tedious. Nevertheless, in stating our results, we have had to resort to Assumption 1. We believe that this assumption holds whenever G has a finite third moment (as specified by Conjecture 1). Toward this end, it is worthwhile to refer the reader to Fralix (2012) and Fralix and Riaño (2010), where similar transient

analysis of the M/G/1 queue is undertaken; see also Pakes (1971) for classic results in the stationary case.

Besides the "transient moment problems" associated with Assumption 1, our work has highlighted other open questions about M/G/1 queues: (i) Daley's conjecture discussed in detail in Section 3.2 above, remains open. (ii) We have not been able to find $\bar{b}_\theta$ when $\rho > 1$. (iii) In the case $\rho = 1$ we conjecture that the variance curve can be written in the form $v(t) = \bar{v}t + a\sqrt{t} + b + o(\sqrt{t})$. Our belief stems from our knowledge of M/M/1 queues (see Eq. (4)):

$$v(t) = 2\left(1 - \frac{2}{\pi}\right)\lambda\, t + \sqrt{\frac{\lambda}{\pi}}\,\sqrt{t} + o(\sqrt{t}).$$

Hence it is natural to believe that the critical M/G/1 queue also has such a $\sqrt{t}$ term, perhaps with a constant differing from $\sqrt{\lambda/\pi}$. Note that the GI/G/1 results of Al Hanbali et al. (2011) actually imply that for the M/G/1 queue,

$$\bar{v} = (1 + c^2)\left(1 - \frac{2}{\pi}\right)\lambda,$$

where c^2 is the squared coefficient of variation of the service time. But to date, nothing is known about the a and b terms for the general M/G/1 queue.

Acknowledgments

We thank two anonymous referees for their comments. We also thank Onno Boxma, Brian Fralix and Guy Latouche for useful discussions and advice. Yoni Nazarathy is supported by Australian Research Council (ARC) grants DP130100156 and DE130100291. Yoav Kerner is supported by Israeli Science Foundation (ISF) grant 1319/11. Yoav Kerner and Yoni Nazarathy also thank EURANDOM for hosting and support. Sophie Hautphenne and Peter Taylor are supported by Australian Research Council (ARC) grant DP110101663 and Laureate Fellowship FL130100039.

References

Al Hanbali, A., Mandjes, M., Nazarathy, Y., & Whitt, W. (2011). The asymptotic variance of departures in critically loaded queues. *Advances in Applied Probability*, *43*(1), 243–263.

Asmussen, S. (2003). *Applied probability and queues*. Springer-Verlag.

Bean, N. G., & Green, D. A. (2000). When is a MAP Poisson? *Mathematical and Computer Modelling*, *31*(10), 31–46.

Bean, N. G., Green, D. A., & Taylor, P. (1998). The output process of an MMPP/M/1 queue. *Journal of Applied Probability*, *35*(4), 998–1002.

Brown, M., & Solomon, H. (1975). A second-order approximation for the variance of a renewal reward process. *Stochastic Processes and their Applications*, *3*(3), 301–314.

Coolen-Schrijner, P., & van Doorn, E. A. (2002). The deviation matrix of a continuous-time Markov chain. *Probability in the Engineering and Informational Sciences*, *16*(3), 351–366. 10.1017/S0269964802163066

Cox, D. R. (1962). *Renewal theory* (Vol. 1). London: Methuen.

Daley, D. J. (1975). Further second-order properties of certain single-server queueing systems. *Stochastic Processes and their Applications*, *3*, 185–191.

Daley, D. J. (1976). Queueing output processes. *Advances in Applied Probability*, *8*, 395–415.

Daley, D. J., & Mohan, N. R. (1978). Asymptotic behaviour of the variance of renewal processes and random walks. *The Annals of Probability*, 516–521.

Daley, D. J., & Vere-Jones, D. (2003). *An introduction to the theory of point processes*. Springer.

Daley, D. J. (2011). Revisiting queueing output processes: a point process viewpoint. *Queueing Systems*, *68*(3–4), 395–405.

Daley, D. J., van Leeuwaarden, J. S. H., & Nazarathy, Y. (2014). BRAVO for many-server QED systems with finite buffers. *Advances in Applied Probability*, in press.

Daley, D. J., & Vesilo, R. (1997). Long range dependence of point processes, with queueing examples. *Stochastic Processes and Their Applications*, *70*(2), 265–282.

Disney, R. L., & Konig, D. (1985). Queueing networks: A survey of their random processes. *SIAM Review*, *27*(3), 335–403.

Finch, P. D. (1959). The output process of the queueing system M/G/1. *Journal of the Royal Statistical Society, Series B (Methodological)*, *21*(2), 375–380.

Fralix, B. H. (2012). On the time-dependent moments of Markovian queues with reneging. *Queueing Systems*, 1–20.

Fralix, B. H., & Riaño, G. (2010). A new look at transient versions of Little's law, and M/G/1 preemptive Last-Come-First-Served queues. *Journal of Applied Probability*, *47*(2), 459–473.

Hendricks, K. B. (1992). The output processes of serial production lines of exponential machines with finite buffers. *Operations Research*, *40*(6), 1139–1147.

Hunter, J. (1969). On the moments of Markov renewal processes. *Advances in Applied Probability*, 188–210.

Hunter, J. J. (1982). Generalized inverses and their application to applied probability problems. *Linear Algebra and Its Applications*, *45*, 157–198.

Kelly, F. (1979). *Reversibility and stochastic networks*. John Wiley & Sons.

Koole, G. M. (1998). The deviation matrix of the M/M/1/ and M/M/1/N queue, with applications to controlled queueing models. In *Proceedings of the 37th IEEE conference on decision and control*.

Koole, G. M., & Spieksma, F. M. (2001). On deviation matrices for birth–death processes. *Probability in the Engineering and Informational Sciences*, *15*(2), 239–258.

Latouche, G., & Ramaswami, V. (1999). *Introduction to matrix analytic methods in stochastic modeling*. PA: SIAM.

Narayana, S., & Neuts, M. F. (1992). The first two moment matrices of the counts for the Markovian arrival process. *Stochastic Models*, *8*(3), 459–477. 10.1080/15326349208807234

Nazarathy, Y. (2011). The variance of departure processes: puzzling behavior and open problems. *Queueing Systems*, *68*(3–4), 385–394.

Nazarathy, Y., & Weiss, G. (2008). The asymptotic variance rate of finite capacity birth-death queues. *Queueing Systems*, *59*(2), 135–156.

Olivier, C., & Walrand, J. (1994). On the existence of finite-dimensional filters for Markov-modulated traffic. *Journal of Applied Probability*, 515–525.

Pakes, A. G. (1971). The correlation coefficients of the queue lengths of some stationary single server queues. *Journal of the Australian Mathematical Society*, *2*(1), 35–46.

Prabhu, N. (1998). *Stochastic storage processes: Queues, insurance risk, dams, and data communication*. Springer.

Smith, W. L. (1959). On the cumulants of renewal processes. *Biometrika*, *46*(1–2), 1–29.

Tan, B. (1997). Variance of the throughput of an N–station production line with no intermediate buffers and time dependent failures. *European Journal of Operational Research*, *101*(3), 560–576.

Tan, B. (1999). Variance of the output as a function of time: Production line dynamics. *European Journal of Operational Research*, *117*(3), 470–484.

Widder, D. V. (1959). The Laplace transform. 1946. *Zbl0139*, *29504*.