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

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

We refer to $\bar{v}$ 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 $\bar{v}$ does not. Since the latter depends on the initial conditions, we generally employ the notation $\bar{b}_e$ 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:

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

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 $\bar{b}_e = 0$. 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, $\bar{v}$, and the intercept term, $\bar{b}$, need not exist for every counting process. Nevertheless, for a variety of models and situations, both $\bar{v}$ 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 $\bar{b}_e$, $\bar{b}_0$ 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