Queueing Syst
https://doi.org/10.1007/s11134-018-9584-z

Robust heavy-traffic approximations for service systems facing overdispersed demand

Britt W. J. Mathijsen[1] · **A. J. E. M. Janssen**[1] · **Johan S. H. van Leeuwaarden**[1] · **Bert Zwart**[1,2]

Received: 16 February 2017 / Revised: 23 April 2018

Abstract Arrival processes to service systems often display fluctuations that are larger than anticipated under the Poisson assumption, a phenomenon that is referred to as *overdispersion*. Motivated by this, we analyze a class of discrete-time stochastic models for which we derive heavy-traffic approximations that are scalable in the system size. Subsequently, we show how this leads to novel capacity sizing rules that acknowledge the presence of overdispersion. This, in turn, leads to robust approximations for performance characteristics of systems that are of moderate size and/or may not operate in heavy traffic.

Keywords Heavy-traffic approximations · Overdispersion · Saddle point method · Random walk

Mathematics Subject Classification 60K25 · 60G50 · 30E20 · 41A60

1 Introduction

One of the most prevalent assumptions in queueing theory is the assumption that the number of arrivals over any given period is a Poisson random variable with deterministic rate, whose variance equals its expectation. Although natural and convenient from a mathematical viewpoint, the Poisson assumption often fails to be confirmed in practice. Namely, a growing number of empirical studies show that the variance of

✉ Britt W. J. Mathijsen
bwjmathijsen@gmail.com

[1] Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

[2] Centrum Wiskunde and Informatica, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Published online: 11 May 2018

demand typically deviates from the mean significantly. Recent work [24,26] reports variance being strictly less than the mean in health care settings employing appointment booking systems. This reduction of variability can be accredited to the goal of the booking system to create a more predictable arrival pattern. On the other hand, in other scenarios with no control over the arrivals, the variance can dominate the mean; see [4–6,11,12,17,19,23,25,30,31,34,38,41]. The feature that variability is higher than one expects from the Poisson assumption is referred to as *overdispersion* and serves as the primary motivation for this work.

Stochastic models with the Poisson assumption have been widely applied to optimize capacity levels in service systems. When stochastic models, however, do not take into account overdispersion, resulting performance estimates are likely to be overoptimistic. The system then ends up being underprovisioned, which possibly causes severe performance problems, particularly in critical loading.

A significant part of the queueing literature has focused on extending Poisson arrival processes to more bursty arrival processes, and analyzing these models using, for example, matrix-analytic models [29,33]. In this paper, we focus on a different cause of overdispersion in arrival processes, which is *arrival rate uncertainty*. Since model primitives, in particular the arrival rate, are typically estimated through historical data, these are prone to be subject to forecasting errors. In the realm of Poisson processes, this inherent uncertainty can be acknowledged by viewing the arrival rate Λ_n itself as being stochastic. The resulting doubly stochastic Poisson process, also known as a Cox process (first presented in [14]), implies that demand in a given interval $A_{k,n}$ follows a mixed Poisson distribution. In this case, the expected demand per period equals $\mu_n = \mathbb{E}[\Lambda_n]$, while the variance is $\sigma_n^2 = \mathbb{E}[\Lambda_n] + \text{Var}\,\Lambda_n$. By selecting the distribution of the mixing factor Λ_n, the magnitude of overdispersion can be made arbitrarily large, and only a deterministic Λ_n leads to a true Poisson process.

The mixed Poisson model presents a useful way to fit both the mean and variance to real data, particularly in case of overdispersion. The mixing distribution can be estimated parametrically or nonparametrically; see [23,30]. A popular parametric family is the Gamma distribution, which gives rise to an effective data fitting procedure that uses the fact that a Gamma mixed Poisson random variable follows a negative binomial distribution. We will in this paper adopt the assumption of a Gamma–Poisson mixture as the demand process.

We investigate the impact of this modeling assumption within the context of a classical model in queueing theory, which is the reflected random walk. In particular, we consider a sequence of such random walks, indexed by n, with increments $A_{k,n} - s_n$, where $A_{k,n} \sim \text{Pois}(\Lambda_n)$ and s_n denotes the system capacity, and we consider a regime in which the system approaches heavy traffic. We are especially interested in the impact of overdispersion on the way performance measures scale, and how they impact capacity allocation rules.

A sensible candidate capacity allocation rule is $s_n = \mu_n + \beta\sigma_n + o(\sigma_n)$ for some $\beta > 0$, which is asymptotically equivalent to the scaling

$$\frac{\mu_n}{\sigma_n}(1 - \rho_n) \to \beta, \quad \text{for } n \to \infty,$$

where $\rho_n := \mu_n/s_n$ denotes the utilization. We will verify mathematically that this is asymptotically the appropriate choice and our methods allow us to quantify the accuracy of the resulting performance formulae for finite systems. Studies that have addressed similar capacity allocation problems with stochastic arrival rates include [28,30,39,40]. Of the aforementioned papers, our work best relates to [30], in the sense that we also assess the asymptotic performance of a queueing system having a stochastic arrival rate in heavy traffic. We therefore expand the paradigm of the quality-and-efficiency-driven (QED) regime, which relies on the popular square-root staffing rule $s_n = \mu_n + \beta\sqrt{\mu_n}$, in order to have it accommodated for overdispersed demand that follows from a doubly stochastic Poisson process.

The first part of our analysis relates to [37], in which a sequence of cyclically thinned queues, denoted by $G_n/G_n/1$ queues, is considered. Here, G_n indicates that only every nth point of the original point process is considered. In this framework, it is shown that the stationary waiting time can be characterized as the maximum of a random walk, in which the increments grow indefinitely. Under appropriate heavy-traffic scaling, the authors prove convergence to a Gaussian random walk and moreover characterize the limits of the stationary waiting time moments. Our work differs with respect to [37] in the sense that we study a discrete-time model, rather than the continuous-time $G_n/G_n/1$ queue. Also, the presence of the overdispersion requires us to employ an alternative scaling.

Furthermore, our approach through Pollaczek's formula allows us to derive estimates for performance measures in pre-limit, i.e., large but finite-size, systems. Mathematically, this second part of our analysis is related to previous work [22]. In particular, we use a refinement of the saddle point technique to establish our asymptotic estimates. The associated analysis is substantially more involved in the present situation, as we will explain in Sect. 4.

Structure of the paper The remainder of this paper is structured as follows. Our model is introduced in Sect. 2 together with some preliminary results. In Sect. 3, we derive the classical heavy-traffic scaling limits for the queue length process in the presence of overdispersed arrivals both for the moments and the distribution itself. Section 4 presents our main theoretical result, which provides a robust refinement to the heavy-traffic characterization of the queue length measures in pre-limit systems. In Sect. 5, we describe the numerical results and demonstrate the heavy-traffic approximation.

2 Model description and preliminaries

We consider a sequence of discrete stochastic models, indexed by n, in which time is divided into periods of equal length. At the beginning of each period $k = 1, 2, 3, \ldots$, new demand $A_{k,n}$ arrives to the system. The demands per period $A_{1,n}, A_{2,n}, \ldots$ are assumed independent and equal in distribution to some nonnegative integer-valued random variable A_n. For brevity, we define $\mu_n := \mathbb{E}A_n$ and $\sigma_n^2 = \mathrm{Var}\, A_n$. The system has a service capacity $s_n \in \mathbb{N}$ per period, so we have the recursion

$$Q_{k+1,n} = \max\{Q_{k,n} + A_{k,n} - s_n, 0\}, \quad k = 0, 1, 2, \ldots, \tag{1}$$

with $Q_{0,n} = 0$. The duality principle for random walks, see, for example [35, Sec. 7.1], shows that this expression is equivalent to

$$Q_{k+1,n} \stackrel{d}{=} \max_{0 \le j \le k} \left\{ \sum_{i=1}^{j} (A_{i,n} - s_n) \right\}, \quad k = 0, 1, 2, \ldots, \tag{2}$$

i.e., the maximum of the first k steps of a random walk with steps distributed as $A_n - s_n$. Even more, we can characterize Q_n, the stationary queue length, as

$$Q_n \stackrel{d}{=} \max_{k \ge 0} \left\{ \sum_{i=1}^{k} (A_{i,n} - s_n) \right\}. \tag{3}$$

The behavior of $Q_{k,n}$ greatly depends on the characteristics of A_n and s_n. First, note that $\mu_n < s_n$ is a necessary condition for the maximum to be finite and therefore for the queue to be stable. This random variable is finite a.s. if $\mathbb{E}[A_{i,n}] < s_n$, which is guaranteed by our Assumption 1 (in particular $\beta > 0$) below. Before continuing the analysis of Q_n, we impose a set of conditions on the asymptotic properties of s_n, μ_n and σ_n, which are assumed to hold throughout the remainder of this paper.

Assumption 1 (a) (Asymptotic growth)

$$\mu_n, \sigma_n \to \infty, \quad \text{for } n \to \infty.$$

(b) (Persistence of overdispersion)

$$\sigma_n^2 / \mu_n \to \infty, \quad \text{for } n \to \infty.$$

(c) (Heavy-traffic condition) The utilization $\rho_n := \mu_n / s_n \to 1$ as $n \to \infty$ according to

$$(1 - \rho_n) \frac{\mu_n}{\sigma_n} \to \beta, \quad \text{for } n \to \infty, \tag{4}$$

for some $\beta > 0$.

By Assumption 1(a), we insist that the expected demand per period grows infinitely large, which allows us to develop approximations for systems with large yet finite arrival volumes. Moreover, it is assumed that the order of stochastic variability of the arrival process relative to the mean arrival volume does not vanish in the limit. In fact, the assumption on the persistence of overdispersion says that the variance of the demand per period is of higher order than its mean as n grows large. We note that the scenario with $\sigma_n^2 / \mu_n \to \gamma$ for some $\gamma > 0$ is asymptotically equivalent to the process studied in [22], in which case overdispersion of the arrival process does not play a role in the limit as $n \to \infty$. In order to establish heavy-traffic approximations for large systems that do face overdispersion we need to construct an asymptotic regime in which overdispersion continues to play a dominant role as $n \to \infty$, which

is secured by Assumption 1(b). The subsequent analysis will clarify why the heavy-traffic condition in Assumption 1(c) is the correct one for our purposes. Note that Assumption 1(c) is satisfied for the capacity allocation rule

$$s_n = \mu_n + \beta\,\sigma_n. \tag{5}$$

Since we are mainly interested in the system behavior in heavy traffic, it is appropriate to study the queue length process in a scaled form. Substituting s_n as in Assumption 1(c), and dividing both sides of (3) by σ_n, gives

$$\frac{Q_n}{\sigma_n} = \max_{k\geq 0}\left\{\sum_{i=1}^{k}\left(\frac{A_{i,n}-\mu_n}{\sigma_n} - \beta\right)\right\}. \tag{6}$$

By defining $\hat{Q}_n := Q_n/\sigma_n$ and

$$\hat{A}_{i,n} := (A_{i,n} - \mu_n)/\sigma_n, \tag{7}$$

we see that the scaled queue length process is in distribution equal to the maximum of a random walk with i.i.d. increments distributed as $\hat{A}_n - \beta$. Besides $\mathbb{E}\hat{A}_n = 0$ and $\text{Var}\,\hat{A}_n = 1$, the scaled and centered arrival counts $\hat{A}_n$ have a few other nice properties which we turn to later in this section.

The model in (1) is valid for any distribution of A_n, also for the original case where the number of arrivals follows a Poisson distribution with fixed parameter λ_n, but Assumption 1(b) does not hold then. Instead, we assume A_n to be Poisson distributed with uncertain arrival rate rendered by the nonnegative random variable Λ_n. This Λ_n is commonly referred to as the *prior* distribution, while A_n is given the name of a Poisson mixture; see [18]. Given that the moment generation function of Λ_n, denoted by $M_n^{\Lambda}(\cdot)$, exists, we are able to express the probability generating function (pgf) of A_n through the former. Namely,

$$\tilde{A}_n(z) = \mathbb{E}[\mathbb{E}[z^{A_n}|\Lambda_n]] = \mathbb{E}[\exp(\Lambda_n(z-1))] = M_n^{\Lambda}(z-1). \tag{8}$$

From (8), we get

$$\mu_n = \mathbb{E}A_n = \mathbb{E}\Lambda_n, \quad \sigma_n^2 = \text{Var}\,A_n = \text{Var}\,\Lambda_n + \mathbb{E}\Lambda_n, \tag{9}$$

so that $\mu_n < \sigma_n^2$ if Λ_n is non-deterministic. Assumption 1(b) hence translates to

$$\text{Var}\,\Lambda_n/\mathbb{E}\Lambda_n \to \infty, \quad n \to \infty.$$

The next result relates the convergence behavior of the centered and scaled Λ_n to that of $\hat{A}_n$.

Lemma 1 *Let $\mu_n, \sigma_n^2 \to \infty$ and $\sigma_n^2/\mu_n \to \infty$. If*

$$\hat{\Lambda}_n := \frac{\Lambda_n - \mu_n}{\sigma_n} \overset{d}{\Rightarrow} N(0,1), \quad \text{for } n \to \infty, \tag{10}$$

where $N(0,1)$ denotes a standard normal variable, then $\hat{A}_n$ converges weakly to a standard normal variable as $n \to \infty$.

The proof can be found in Appendix A.

The prevalent choice for Λ_n is the Gamma distribution. The Gamma–Poisson mixture turns out to provide a good fit to arrival counts observed in service systems, as was observed by [23,30]. Assuming Λ_n to be of Gamma type with scale and rate parameters a_n and $1/b_n$, respectively, we get

$$\tilde{A}_n(z) = \left(\frac{1}{1 + b_n(1-z)}\right)^{a_n}, \tag{11}$$

in which we recognize the pgf of a negative binomial distribution with parameters a_n and $1/(b_n+1)$, so that

$$\mu_n = a_n b_n, \quad \sigma_n^2 = a_n b_n (b_n + 1). \tag{12}$$

Note that in the context of a Gamma prior, the restrictions in Assumption 1 reduce to only two rules. For completeness, we include the revised list below.

Assumption 2 (a) (Asymptotic regime and persistence of overdispersion)

$$a_n, b_n \to \infty, \quad \text{for } n \to \infty.$$

(b) (Heavy-traffic condition) Let

$$s_n = a_n b_n + \beta\sqrt{a_n b_n (b_n+1)} + o\big(\sqrt{a_n} b_n\big),$$

for some $\beta > 0$, or equivalently

$$(1-\rho_n)\sqrt{a_n} \to \beta, \quad \text{for } n \to \infty.$$

The next result follows from the fact that Λ_n is a Gamma random variable:

Corollary 1 *Let $\Lambda_n \sim \text{Gamma}(a_n, 1/b_n)$, $A_n \sim \text{Poisson}(\Lambda_n)$ and $a_n, b_n \to \infty$. Then, $\hat{A}_n$ converges weakly to a standard normal random variable as $n \to \infty$.*

Proof With Lemma 1 in mind, it is sufficient to prove that $\hat{\Lambda}_n \Rightarrow N(0,1)$ for this particular choice of Λ_n. We do this by proving the pointwise convergence of the characteristic function (cf) of $\hat{\Lambda}_n$ to $\exp(-t^2/2)$, the cf of the standard normal distribution. Let $\varphi_G(\cdot)$ denote the characteristic function of a random variable G. By basic

properties of the cf,

$$\begin{aligned}\varphi_{\hat{\Lambda}_n}(t) &= \mathrm{e}^{-i\mu_n t/\sigma_n}\,\varphi_{\Lambda_n}(t/\sigma_n) = \mathrm{e}^{-i\mu_n t/\sigma_n}\left(1-\frac{ib_n t}{\sigma_n}\right)^{-a_n}\\ &= \exp\left[-\frac{i\mu_n t}{\sigma_n} - a_n \ln\left(1-\frac{ib_n t}{\sigma_n}\right)\right]\\ &= \exp\left[-\frac{i\mu_n t}{\sigma_n} - a_n\left(-\frac{i\,b_n t}{\sigma_n}+\frac{b_n^2 t^2}{2\sigma_n^2}+O(b_n^3/\sigma_n^3)\right)\right]\\ &= \exp\left[-\frac{b_n\,t^2}{2(b_n+1)}+O\left(1/\sqrt{a_n}\right)\right]\to\exp\left(-t^2/2\right),\end{aligned}\tag{13}$$

for $n\to\infty$. Since $b_n/\sigma_n = a_n^{-1/2}(1+1/b_n)^{-1/2}\to 0$ as $n\to\infty$, we can take the principal value in (13) for the logarithm when t is in any compact set and n is large enough. By Lévy's continuity theorem, see, for example, [16, Thm. 18.21], this implies $\hat{\Lambda}_n$ is indeed asymptotically standard normal. □

The characterization of the arrival process as a Gamma–Poisson mixture is of vital importance in later sections.

2.1 Expressions for the stationary distribution

Our main focus is on the stationary queue length distribution, denoted by

$$\mathbb{P}(Q_n=i)=\lim_{k\to\infty}\mathbb{P}(Q_{k,n}=i).$$

Denote the pgf of Q_n by

$$\tilde{Q}_n(w)=\sum_{i=0}^{\infty}\mathbb{P}(Q_n=i)w^i.\tag{14}$$

To continue our analysis of Q_n, we need one more condition on A_n.

Assumption 3 The pgf of A_n, denoted by $\tilde{A}_n(w)$, exists within $|w|<r_0$, for some $r_0>1$, so that all moments of A_n are finite.

We next recall two characterizations of $\tilde{Q}_n(w)$ that play prominent roles in the remainder of our analysis. The first characterization of $\tilde{Q}_n(w)$ originates from a random walk perspective. As we see from (3), the (scaled) stationary queue length is equal in distribution to the all-time maximum of a random walk with i.i.d. increments distributed as A_n-s_n (or $\hat{A}_n-\beta$ in the scaled setting). Spitzer's identity, see, for example, [3, Theorem VIII4.2], then gives

$$\tilde{Q}_n(w)=\exp\left\{\sum_{k=1}^{\infty}\frac{1}{k}\left(\mathbb{E}\left[w^{\left(\sum_{i=1}^{k}\{A_{i,n}-s_n\}\right)^+}\right]-1\right)\right\},\tag{15}$$

where $(x)^+ = \max\{x, 0\}$. Hence,

$$\mathbb{E}Q_n = \tilde{Q}_n'(1) = \sum_{k=1}^{\infty} \frac{1}{k} \mathbb{E}\left[\sum_{i=1}^{k} (A_{i,n} - s_n)\right]^+, \tag{16}$$

$$\text{Var } Q_n = \tilde{Q}_n''(1) + Q_n'(1) - \left(\tilde{Q}_n'(1)\right)^2 = \sum_{k=1}^{\infty} \frac{1}{k} \mathbb{E}\left[\left(\sum_{i=1}^{k} (A_{i,n} - s_n)\right)^+\right]^2, \tag{17}$$

$$\mathbb{P}(Q_n = 0) = \tilde{Q}_n(0) = \exp\left\{-\sum_{k=1}^{\infty} \frac{1}{k} \mathbb{P}\left(\sum_{i=1}^{k} (A_{i,n} - s_n) > 0\right)\right\}. \tag{18}$$

A second characterization follows from Pollaczek's formula, see [1,22]:

$$\tilde{Q}_n(w) = \exp\left\{\frac{1}{2\pi i} \int_{|z|=1+\varepsilon} \ln\left(\frac{w-z}{1-z}\right) \frac{(z^{s_n} - \tilde{A}_n(z))'}{z^{s_n} - \tilde{A}_n(z)}\, dz\right\}, \tag{19}$$

which is analytic for $|w| < r_0$, for some $r_0 > 1$. Therefore, $\varepsilon > 0$ has to be chosen such that $|w| < 1 + \varepsilon < r_0$. This gives

$$\mathbb{E}Q_n = \frac{1}{2\pi i} \int_{|z|=1+\varepsilon} \frac{1}{1-z} \frac{(z^{s_n} - \tilde{A}_n(z))'}{z^{s_n} - \tilde{A}_n(z)}\, \mathrm{d}z, \tag{20}$$

$$\text{Var } Q_n = \frac{1}{2\pi i} \int_{|z|=1+\varepsilon} \frac{-z}{(1-z)^2} \frac{(z^{s_n} - \tilde{A}_n(z))'}{z^{s_n} - \tilde{A}_n(z)}\, \mathrm{d}z, \tag{21}$$

$$\mathbb{P}(Q_n = 0) = \exp\left\{\frac{1}{2\pi i} \int_{|z|=1+\varepsilon} \ln\left(\frac{z}{z-1}\right) \frac{(z^{s_n} - \tilde{A}_n(z))'}{z^{s_n} - \tilde{A}_n(z)}\, \mathrm{d}z\right\}. \tag{22}$$

Pollaczek-type integrals like (19)–(22) first occurred in the work of Pollaczek on the classical single-server queue (see [1,13,21] for historical accounts). These integrals are fairly straightforward to evaluate numerically and hence give rise to efficient algorithms for performance evaluation [1,9]. The integrals also proved useful in establishing heavy-traffic results by asymptotic evaluation of the integrals in various heavy-traffic regimes [8,13,22,27], and in this paper we follow that approach for a heavy-traffic regime that is suitable for overdispersion.

3 Heavy-traffic limits

In this section, we present the result on the convergence of the discrete process $\hat{Q}_n$ to a non-degenerate limiting process and of the associated stationary moments. The latter requires an interchange of limits. Using this asymptotic result, we derive two sets of approximations for $\mathbb{E}Q_n$, Var Q_n and $\mathbb{P}(Q_n = 0)$ that capture the limiting behavior of Q_n. The first set provides a rather crude estimation for the first cumulants of the

queue length process for any arrival process A_n satisfying Assumption 1. The second set, which is the subject of the next section, is derived for the specific case of a Gamma prior and is therefore expected to provide more accurate, robust approximations for the performance metrics.

We start by indicating how the asymptotic properties of the scaled arrival process give rise to a proper limiting random variable describing the stationary queue length. The asymptotic normality of $\hat{A}_n$ provides a link with the Gaussian random walk and nearly deterministic queues [36,37]. The main results in [36,37] were obtained under the assumption that $\rho_n \sim 1 - \beta/\sqrt{n}$, in which case it follows from [37, Thm. 3] that the rescaled stationary waiting time process converges to a reflected Gaussian random walk.

We shall also identify the Gaussian random walk as the appropriate scaling limit for our stationary system. However, since the normalized natural fluctuations of our system are given by μ_n/σ_n instead of $\sqrt{n}$, we assume that the load grows like $\rho_n \sim 1 - \frac{\beta}{\mu_n/\sigma_n}$. Hence, in contrast to [36,37], our systems' characteristics display larger natural fluctuations, due to the mixing factor that drives the arrival process. Yet, by matching this overdispersed demand with the appropriate hedge against variability, we again obtain Gaussian limiting behavior. This is not surprising, since we saw in Lemma 1 that the increments start resembling Gaussian behavior for $n \to \infty$. The following result summarizes this.

Theorem 1 *Let A_n be a nonnegative random variable such that $\hat{A}_n = (A_n - \mu_n)/\sigma_n$ is asymptotically standard normal, with μ_n and σ_n as defined in (9), and $\mathbb{E}[(\max\{\hat{A}_n, 0\})^4]$ is bounded in n. Then, under Assumption 1, for $n \to \infty$,*

(i) $\hat{Q}_n \overset{d}{\Rightarrow} M_\beta$,
(ii) $\mathbb{P}(Q_n = 0) \to \mathbb{P}(M_\beta = 0)$,
(iii) $\mathbb{E}\hat{Q}_n \to \mathbb{E}M_\beta$,
(iv) $\text{Var}\ \hat{Q}_n \to \text{Var}\ M_\beta$,

where M_β is the all-time maximum of a random walk with i.i.d. normal increments with mean $-\beta$ and unit variance.

The proof of Theorem 1 is given in Appendix A. We remark that for convergence of the mean scaled queue length, only $\mathbb{E}[(\max\{\hat{A}_n, 0\})^3] < \infty$ is needed. The following result shows that Theorem 1 also applies to Gamma mixtures, which is a direct consequence of Corollary 1.

Corollary 2 *Let $\Lambda_n \sim Gamma(a_n, b_n)$. Then, under Assumption 2 the four convergence results of Theorem 1 hold true.*

It follows from Theorem 1 that the scaled stationary queueing process converges under (4) to a reflected Gaussian random walk. Hence, the performance measures of the original system should be well approximated by the performance measures of the reflected Gaussian random walk, yielding heavy-traffic approximations.

Like our original system, the Gaussian random walk falls in the classical setting of the reflected one-dimensional random walk, whose behavior is characterized by both Spitzer's identity and Pollaczek's formula. In particular, Pollaczek's formula

gives rise to contour integral expressions for performance measures that are easy to evaluate numerically, also in heavy-traffic conditions. The numerical evaluation of such integrals is considered in [1]. For $\mathbb{E}M_\beta$, such an integral is as follows:

$$\mathbb{E}M_\beta = -\frac{1}{\pi}\int_0^\infty \mathrm{Re}\left[\frac{1-\phi(-z)}{z^2}\right]\mathrm{d}y, \tag{23}$$

with $\phi(z) = \exp(-\beta z + \frac{1}{2}z^2)$, the Laplace transform of a normal random variable with mean $-\beta$ and unit variance, and $z = x + iy$ with an appropriately chosen real part x. Note that this integral involves complex-valued functions with complex arguments. Similar Pollaczek-type integrals exist for $\mathbb{P}(M_\beta = 0)$ and Var M_β; see [1]. The following result simply rewrites these integrals in terms of real integrals and uses the fact that the scaled queue length process mimics the maximum of the Gaussian random walk for large n.

Corollary 3 *Under Assumption 1, the leading order behavior of* $\mathbb{P}(Q_n = 0)$, $\mathbb{E}Q_n$ *and* Var Q_n *as* $n \to \infty$ *is given by, respectively,*

$$\exp\left[\frac{1}{\pi}\int_0^\infty \frac{\beta/\sqrt{2}}{\frac{1}{2}\beta^2+t^2}\ln\left(1-e^{-\frac{1}{2}\beta^2-t^2}\right)\mathrm{d}t\right], \tag{24}$$

$$\frac{\sqrt{2}\sigma_n}{\pi}\int_0^\infty \frac{t^2}{\frac{1}{2}\beta^2+t^2}\frac{\exp(-\frac{1}{2}\beta^2-t^2)}{1-\exp(-\frac{1}{2}\beta^2-t^2)}\mathrm{d}t, \tag{25}$$

$$\frac{\sqrt{2}\beta\sigma_n^2}{\pi}\int_0^\infty \frac{t^2}{(\frac{1}{2}\beta^2+t^2)^2}\frac{\exp(-\frac{1}{2}\beta^2-t^2)}{1-\exp(-\frac{1}{2}\beta^2-t^2)}\mathrm{d}t. \tag{26}$$

Proof According to [1, Eq. (15)],

$$-\ln[\mathbb{P}(M_\beta = 0)] = c_0, \quad \mathbb{E}M_\beta = c_1, \quad \text{Var } M_\beta = c_2,$$

where

$$c_n = \frac{(-1)^n n!}{\pi}\mathrm{Re}\left[\int_0^\infty \frac{\ln(1-\exp(\beta z+\frac{1}{2}z^2))}{z^{n+1}}\mathrm{d}y\right],$$

in which $z = -x + i\,y$, $y \geq 0$, and x is any fixed number between 0 and 2β. Take $x = \beta$, so that

$$\beta z + \tfrac{1}{2}z^2 = -\tfrac{1}{2}\beta^2 - \tfrac{1}{2}y^2 \leq 0, \quad y \geq 0.$$

For $n = 0$, this gives

$$c_0 = \frac{1}{\pi}\mathrm{Re}\left[\int_0^\infty \frac{\ln(1-\exp(-\frac{1}{2}\beta^2-\frac{1}{2}y^2))}{-\beta+i\,y}\mathrm{d}y\right]$$

$$= -\frac{1}{\pi}\int_0^\infty \frac{\beta}{\beta^2+y^2}\ln\left(1-\exp(-\tfrac{1}{2}\beta^2-\tfrac{1}{2}y^2)\right)\mathrm{d}y$$
$$= -\frac{1}{\pi}\int_0^\infty \frac{\beta/\sqrt{2}}{\frac{1}{2}\beta^2+t^2}\ln\left(1-\exp(-\tfrac{1}{2}\beta^2-t^2)\right)\mathrm{d}t,$$

where we used that

$$\mathrm{Re}\left[\frac{1}{-\beta+i\,y}\right]=\frac{-\beta}{\beta^2+y^2},$$

together with the substitution $y=t\sqrt{2}$. For $n=1,2,\ldots,$ partial integration gives

$$c_n = \frac{(-1)^n n!}{\pi}\,\mathrm{Re}\left[\int_0^\infty \frac{\ln(1-\exp(-\frac{1}{2}\beta^2-\frac{1}{2}y^2))}{(-\beta+i\,y)^{n+1}}\mathrm{d}y\right]$$
$$= \frac{(-1)^{n-1}(n-1)!}{\pi}\,\mathrm{Im}\left[\int_0^\infty \ln(1-\exp(-\tfrac{1}{2}\beta^2-\tfrac{1}{2}y^2))\mathrm{d}\left(\frac{1}{(-\beta+i\,y)^n}\right)\right]$$
$$= -\frac{(-1)^{n-1}(n-1)!}{\pi}\mathrm{Im}\left[\int_0^\infty \frac{y}{(-\beta+i\,y)^n}\,\frac{\exp(-\frac{1}{2}\beta^2-\frac{1}{2}y^2)}{1-\exp(-\frac{1}{2}\beta^2-\frac{1}{2}y^2)}\mathrm{d}y\right],$$

where we have used that

$$\mathrm{Im}\left[\frac{\ln(1-\exp(-\frac{1}{2}\beta^2-\frac{1}{2}y^2))}{(-\beta+i\,y)^n}\right]\Big|_0^\infty = 0.$$

Using

$$\frac{1}{(-\beta+i\,y)^n}=(-1)^n\,\frac{(\beta+i\,y)^n}{(\beta^2+y^2)^n},$$

we then get

$$c_n = \frac{(n-1)!}{\pi}\,\mathrm{Im}\left[\int_0^\infty \frac{y(\beta+i\,y)^n}{(\beta^2+y^2)^n}\,\frac{\exp(-\frac{1}{2}\beta^2-\frac{1}{2}y^2)}{1-\exp(-\frac{1}{2}\beta^2-\frac{1}{2}y^2)}\mathrm{d}y\right],$$

which, after substitution of $y=t\sqrt{2}$ gives (25) and (26). □

4 Robust heavy-traffic approximations

We shall now establish robust heavy-traffic approximations for the canonical case of Gamma–Poisson mixtures; see (11).

Theorem 2 *Let a_n, b_n and s_n be as in Assumption 2. Then, the leading order behavior of $\mathbb{E}Q_n$ is given by*

$$\frac{\sqrt{2}\,\beta_n}{\pi}\left(\frac{b_n+\rho_n}{1-\rho_n}\right)\int_0^\infty \frac{t^2}{\frac{1}{2}\beta_n^2+t^2}\,\frac{\exp(-\frac{1}{2}\beta_n^2-t^2)}{1-\exp(-\frac{1}{2}\beta_n^2-t^2)}\mathrm{d}t\,(1+o(1)), \tag{27}$$

where

$$\beta_n^2 = s_n\left(\frac{1-\rho_n}{b_n+1}\right)^2\left(1+\frac{b_n}{\rho_n}\right). \tag{28}$$

Furthermore, the leading order behavior of $\mathbb{P}(Q_n=0)$ and $\mathrm{Var}\,Q_n$ is given by

$$\exp\left[\frac{1}{\pi}\,\frac{b_n+\rho_n}{b_n+1}\int_0^\infty \frac{\beta_n/\sqrt{2}}{\frac{1}{2}\beta_n^2+t^2}\ln\left(1-\mathrm{e}^{-\frac{1}{2}\beta_n^2-t^2}\right)\mathrm{d}t\right],$$

and

$$\frac{\beta_n^3/\sqrt{2}}{\pi}\left(\frac{b_n+\rho_n}{1-\rho_n}\right)^2\left(\frac{b_n+1}{b_n+\rho_n}+1\right)\int_0^\infty \frac{t^2}{(\frac{1}{2}\beta_n+t^2)^2}\,\frac{\exp(-\frac{1}{2}\beta_n-t^2)}{1-\exp(-\frac{1}{2}\beta_n^2-t^2)}\mathrm{d}t, \tag{29}$$

respectively.

The proof of Theorem 2 requires asymptotic evaluation of the Pollaczek-type integrals (20)–(22), for which we shall use a *nonstandard* saddle point method. The saddle point method in its standard form is typically suitable for large deviation regimes, for instance excess probabilities, and it cannot be applied to asymptotically characterize other stationary measures such as the mean or mass at zero. Indeed, in the presence of overdispersion, the saddle point converges to one (as $n\to\infty$), which is a singular point of the integrand, and renders the standard saddle point method useless. Our nonstandard saddle point method, originally proposed by [15] and also applied in [22], aims specifically to overcome this challenge. Subsequently, we apply the nonstandard saddle point method to turn these contour integrals into practical approximations. In contrast to the setting of [22], the analyticity radius tends to one in the setting with overdispersion, which is a singular point of the integrand. For the proof of Theorem 2, we therefore modify the special saddle point method developed in [22] to account for this circumstance.

Proof Our starting point is the probability generating function of the number of arrivals per time slot, given in (11), which is analytic for $|z|<1+1/b_n=:r_0$. Under Assumption 2, we consider $\mathbb{E}Q_n$ as given in (20). We set

$$g(z) = -\ln z + \frac{1}{s_n}\ln\big[\tilde{A}_n(z)\big] = -\ln z - \frac{a_n}{s_n}\ln\left(1+(1-z)b_n\right), \tag{30}$$

to be considered in the entire complex plane with branch cuts $(-\infty, 0]$ and $[r, \infty)$. The relevant saddle point $z_{\rm sp}$ is the unique zero z of $g'(z)$ with $z \in (1, r_0)$. Since

$$g'(z) = -\frac{1}{z} + \frac{\rho_n}{1 + (1-z)b_n}, \tag{31}$$

this yields

$$1 + (1 - z_{\rm sp})b_n = \rho_n z_{\rm sp}, \quad \text{i.e.,} \quad z_{\rm sp} = 1 + \frac{1-\rho_n}{\rho_n + b_n}. \tag{32}$$

We then find

$$\mathbb{E}Q_n = \frac{s_n}{2\pi i} \int_{|z|=1+\varepsilon} \frac{g'(z)}{z-1} \frac{\exp(s_n\, g(z))}{1 - \exp(s_n\, g(z))} \mathrm{d}z, \tag{33}$$

and take $1 + \varepsilon = z_{\rm sp}$. There are no problems with the branch cuts since we consider $\exp(s_n g(z))$ with integer s_n.

We continue as in [22] and thus we intend to substitute $z = z(v)$ in the integral in (33), where $z(v)$ satisfies

$$g(z(v)) = g(z_{\rm sp}) - \tfrac{1}{2}\, v^2\, g''(z_{\rm sp}) =: q(v)$$

in the range $-\frac{1}{2}\delta_n \le v \le \frac{1}{2}\delta_n$ with $\delta_n \to 0$ as $n \to \infty$. Note that this range depends on n, whereas these bounds $\pm\frac{1}{2}\delta_n$ remained bounded away from zero in [22]. This severely complicates the present analysis. We consider the approximate representation

$$\frac{-s_n\, g''(z_{\rm sp})}{2\pi i} \int_{-\frac{1}{2}\delta_n}^{\frac{1}{2}\delta_n} \frac{v}{z(v)-1} \frac{\exp(s_n\, q(v))}{1 - \exp(s_n\, q(v))} \mathrm{d}v \tag{34}$$

of $\mathbb{E}Q_n$. We have to operate here with additional care, since both the analyticity radius $r_0 = 1 + 1/b_n$ and the saddle point $z_{\rm sp}$ outside the unit circle tend to 1 as $n \to \infty$. Specifically, proceeding under the assumptions that $(1-\rho_n)^2 a_n$ is bounded while $a_n \to \infty$ and $b_n \ge 1$, see Assumption 2, we have from (32) that

$$z_{\rm sp} - 1 = \frac{1-\rho_n}{b_n + \rho_n} = \frac{1-\rho_n}{b_n} + O\left(\frac{1-\rho_n}{b_n^2}\right), \tag{35}$$

where the O-term is small compared to $(1-\rho_n)/b_n$ when $b_n \to \infty$. Next, we approximate r_0, using that $r_0 > 1$ satisfies

$$-\ln r_0 - \frac{\rho_n}{b_n} \ln\left(1 + (1-r_0)b_n\right) = 0.$$

Write $r_0 = 1 + u/b_n$, so that we get the equation

$$\begin{aligned} 0 &= -\ln\left(1+\frac{u}{b_n}\right) - \frac{\rho_n}{b_n}\ln(1-u) \\ &= -\frac{u}{b_n}\left(1-\rho_n - \tfrac{1}{2}\left(\frac{1}{b_n}+\rho_n\right)u - \tfrac{1}{3}\left(\frac{-1}{b_n^2}+\rho_n\right)u^2 + \cdots\right), \end{aligned}$$

where we have used the Taylor expansion of $\ln(1+x)$ at $x = 0$. Thus, we find

$$u = \frac{2(1-\rho_n)}{\rho_n + 1/b_n} + O(u^2) = 2(1-\rho_n) + O((1-\rho_n)^2) + O\left(\frac{1-\rho_n}{b_n}\right),$$

and so,

$$r_0 = 1 + 2\,\frac{1-\rho_n}{b_n} + O\left(\frac{(1-\rho_n)^2}{b_n}\right) + O\left(\frac{1-\rho_n}{b_n^2}\right).$$

In (34) we choose δ_n so large that the integral has converged within exponentially small error using $\pm\,\delta_n$ as integration limits and, at the same time, so small that there is a convergent power series

$$z(v) = z_{\rm sp} + i\,v + \sum_{k=2}^{\infty} c_k (i\,v)^k, \quad \text{for } |v| \le \tfrac{1}{2}\delta_n. \tag{36}$$

To achieve these goals, we supplement the information on $g(z)$, as given by (30)–(32), by

$$g''(z) = \frac{1}{z^2} + \frac{\rho_n b_n}{(1+(1-z)b_n)^2}, \quad g''(1) = 1 + \rho_n b_n, \quad g''(z_{\rm sp}) = \frac{1}{z_{\rm sp}^2}\left(1 + \frac{b_n}{\rho_n}\right). \tag{37}$$

Now,

$$\exp(s_n\, q(v)) = \exp(s_n\, g(z_{\rm sp}))\exp(-\tfrac{1}{2}\, s_n\, g''(z_{\rm sp})\, v^2),$$

and

$$s_n\, g''(z_{\rm sp}) v^2 = s_n\, b_n v^2 (1 + o(1)) = a_n (b_n\, v)^2 (1 + o(1)).$$

Therefore, (34) approximates $\mathbb{E}\,Q_n$ with exponentially small error when we take $\frac{1}{2}\delta_n$ of the order $1/b_n$.

We next aim at showing that we have a power series for $z(v)$ as in (36) that converges for $|v| \le \frac{1}{2}\delta_n$, with $\frac{1}{2}\delta_n$ of the order $1/b_n$.

Lemma 2 *Let*

$$r_n := \frac{1}{2\,b_n} - (z_{\rm sp} - 1), \quad m_n := \tfrac{2}{3}\rho_n r_n \sqrt{\frac{b_n + \rho_n^{-1}}{b_n + \rho_n}},$$

where we assume $r_n > 0$. *Then,* (36) *holds with real coefficients* c_k *satisfying*

$$|c_k| \le \frac{r_n}{m_n^k}, \quad k = 2, 3, \ldots. \tag{38}$$

Proof We let

$$G(z) := \frac{2(g(z) - g(z_{\rm sp}))}{g''(z_{\rm sp})(z - z_{\rm sp})^2}. \tag{39}$$

Then $G(z_{\rm sp}) = 1$ and so we can write (4) as

$$F(z) := (z - z_{\rm sp})\sqrt{G(z)} = i\,v \tag{40}$$

when $|z - z_{\rm sp}|$ is sufficiently small. Since $F(z_{\rm sp}) = 0$, $F'(z_{\rm sp}) = 1$, the Bürmann–Lagrange inversion theorem implies validity of a power series as in (36), with real c_k since $G(z)$ is positive and real for real z close to $z_{\rm sp}$. We therefore just need to estimate the convergence radius of this series from below.

To this end, we start by showing that

$$\mathrm{Re}[g''(z)] > \frac{4}{9}\,\rho_n^2 \frac{b_n + \rho_n^{-1}}{b_n + \rho_n}, \quad |z - z_{\rm sp}| \le r_n. \tag{41}$$

For this, we consider the representation

$$G(z) = 2\int_0^1 \int_0^1 \frac{g''(z_{\rm sp} + s\,t\,(z - z_{\rm sp}))}{g''(z_{\rm sp})}\, t\,\mathrm{d}s\mathrm{d}t. \tag{42}$$

We have, for $\zeta \in \mathbb{C}$ and $|\zeta - 1| \le 1/2b_n \le 1/2$, from (37) that

$$\mathrm{Re}[g''(\zeta)] = \mathrm{Re}(1/\zeta^2) + \rho_n b_n\,\mathrm{Re}\left[\left(\frac{1}{1 + (1 - \zeta)b_n}\right)^2\right] \ge \tfrac{4}{9}(1 + \rho_n b_n). \tag{43}$$

To show the inequality in (43), it suffices to show that

$$\min_{|\xi - 1| \le 1/2} \mathrm{Re}\left(\frac{1}{\xi^2}\right) = \frac{4}{9}. \tag{44}$$

The minimum in (44) is assumed at the boundary $|\xi - 1| = 1/2$, and for a boundary point ξ we write

$$\xi = 1 + \tfrac{1}{2}\cos\theta + \tfrac{1}{2}i\sin\theta, \quad 0 \le \theta \le 2\pi,$$

so that

$$\mathrm{Re}\left(\frac{1}{\xi^2}\right) = \frac{1+\cos\theta+\frac{1}{4}\cos 2\theta}{\left(\frac{5}{4}+\cos\theta\right)^2}.$$

Now

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\left[\frac{1+\cos\theta+\frac{1}{4}\cos 2\theta}{\left(\frac{5}{4}+\cos\theta\right)^2}\right] = \frac{\sin\theta\,(1-\cos\theta)}{4\left(\frac{5}{4}+\cos\theta\right)^3}$$

vanishes for $\theta = 0, \pi, 2\pi$, where $\mathrm{Re}(1/\xi^2)$ assumes the values 4/9, 4, 4/9, respectively. This shows (44).

We use (44) with $\xi = \zeta$ and $\xi = 1 + (1-\zeta)b_n$, with

$$\zeta = \zeta(s,t) = z_{\mathrm{sp}} + s\,t\,(z - z_{\mathrm{sp}}), \quad 0 \le s,\ t \le 1, \tag{45}$$

where we take ζ such that $|\zeta - 1| \le 1/2b_n$. It is easy to see from $1 < z_{\mathrm{sp}} < 1 + 1/2b_n$ that $|\zeta - 1| \le 1/2b_n$ holds when $|z - z_{\mathrm{sp}}| \le r_n = 1/2b_n - (z_{\mathrm{sp}} - 1)$. We have, furthermore, from (32) that $0 < g''(z_{\mathrm{sp}}) \le 1 + b_n/\rho_n$. Using this, together with (43) where ζ is as in (45), yields

$$\mathrm{Re}[G(z)] \le \frac{4}{9}\,\frac{1+\rho_n b_n}{1+b_n/\rho_n}\,2\int_0^1\int_0^1 t\,\mathrm{d}s\,\mathrm{d}t = \tfrac{4}{9}\,\rho_n^2\,\frac{b_n+\rho_n^{-1}}{b_n+\rho_n}$$

when $|z - z_{\mathrm{sp}}| \le r_n$, and this is (41). We therefore have from (40) that

$$|F(z)| > r_n \cdot \frac{2}{3}\rho_n\sqrt{\frac{b_n+\rho_n^{-1}}{b_n+\rho_n}} = m_n, \quad |z - z_{\mathrm{sp}}| = r_n.$$

Hence, for any v with $|v| \le m_n$, there is exactly one solution $z = z(v)$ of the equation $F(z) - iv = 0$ in $|z - z_{\mathrm{sp}}| \le r_n$ by Rouché's theorem [2]. This $z(v)$ is given by

$$z(v) = \frac{1}{2\pi i}\int_{|z-z_{\mathrm{sp}}|=r_n} \frac{F'(z)\,z}{F(z) - iv}\mathrm{d}z,$$

and depends analytically on v, $|v| \le m_n$. From $|z(v) - z_{\mathrm{sp}}| \le r_n$, we can finally bound the power series coefficients c_k according to

$$|c_k| = \left|\frac{1}{2\pi i}\int_{|iv|=m_n} \frac{z(v) - z_{\mathrm{sp}}}{(iv)^{k+1}}\mathrm{d}(iv)\right| \le \frac{r_n}{m_n^k},$$

and this completes the proof of the lemma. □

Remark 1 We have $z_{\rm sp} - 1 = o(1/b_n)$, see (35), and so

$$r_n = \frac{1}{2b_n}(1 + o(1)), \quad m_n = \frac{1}{3b_n}(1 + o(1)),$$

implying that the radius of convergence of the series in (36) is indeed of order $1/b_n$ (since we have assumed $b_n \geq 1$).

We let $\delta_n = m_n$, and we write, for $0 \leq v \leq \frac{1}{2}\delta_n$,

$$\frac{v}{z(v) - 1} + \frac{-v}{z(-v) - 1} = \frac{-2iv\,\mathrm{Im}(z(v))}{|z(v) - 1|^2},$$

where we have used that all c_k are real, so that $z(-v) = z(v)^*$, where * denotes the complex conjugate. Now, from (38) and realness of the c_k, we have

$$\mathrm{Im}(z(v)) = v + \sum_{l=1}^{\infty} c_{2l+1}(-1)^l\, v^{2l+1} = v + O(v^3), \tag{46}$$

and in similar fashion

$$|z(v) - 1|^2 = (z_{\rm sp} - 1)^2 + v^2 + O((z_{\rm sp} - 1)^2 v^2) + O(v^4) \tag{47}$$

when $0 \leq v \leq \frac{1}{2}\delta_n$. The order terms in (46), (47) are negligible in leading order, and so we get for μ_{Q_n} via (34) the leading order expression

$$\frac{-s_n\, g''(z_{\rm sp})}{2\pi i} \int_0^{\frac{1}{2}\delta_n} \frac{-2iv^2}{(z_{\rm sp} - 1)^2 + v^2} \frac{\exp(s_n\, q(v))}{1 - \exp(s_n\, q(v))} \mathrm{d}v.$$

We finally approximate $q(v) = g(z_{\rm sp}) - \frac{1}{2}g''(z_{\rm sp})v^2$. There is a z_1, $1 \leq z_1 \leq z_{\rm sp}$, such that

$$g(z_{\rm sp}) = -\tfrac{1}{2}(z_{\rm sp} - 1)^2\, g''(z_1),$$

and, see (35) and (37),

$$g''(z_1) = g''(z_{\rm sp}) + O((1 - \rho_n)b_n).$$

Hence

$$\begin{aligned} s_n\, q(v) &= -\tfrac{1}{2}s_n\, g''(z_{\rm sp})\,[(z_{\rm sp} - 1)^2 + v^2] + O((1 - \rho_n)b_n s_n (z_{\rm sp} - 1)^2) \\ &= -\tfrac{1}{2}s_n\, g''(z_{\rm sp})[(z_{\rm sp} - 1)^2 + v^2] + O((1 - \rho_n)^2 a_n), \end{aligned} \tag{48}$$

where (35) has been used, and $a_n b_n = s_n(1 + o(1))$. Therefore, the O-term in (48) tends to 0 by our assumption that $(1 - \rho_n)^2 a_n$ is bounded. Thus, we get for μ_{Q_n} in leading order

$$\frac{s_n g''(z_{\rm sp})}{\pi} \int_0^{\frac{1}{2}\delta_n} \frac{v^2}{(z_{\rm sp}-1)^2 + v^2} \frac{\exp(-\frac{1}{2} g''(z_{\rm sp}) s_n((z_{\rm sp}-1)^2 + v^2))}{1 - \exp(-\frac{1}{2} g''(z_{\rm sp}) s_n((z_{\rm sp}-1)^2 + v^2))} {\rm d}v. \tag{49}$$

When we substitute $t = v\sqrt{s_n\, g''(z_{\rm sp})/2}$ and extend the integration in (49) to all $t \geq 0$ (at the expense of an exponentially small error), we get for μ_{Q_n} in leading order

$$= \frac{1}{\pi} \sqrt{2\, s_n\, g''(z_{\rm sp})} \int_0^{\infty} \frac{t^2}{\frac{1}{2}\beta_n^2} \frac{\exp(-\frac{1}{2}\beta_n^2 - t^2)}{1 - \exp(-\frac{1}{2}\beta_n^2 - t^2)} {\rm d}t,$$

where

$$\beta_n^2 = s_n\, g''(z_{\rm sp})(z_{\rm sp} - 1)^2.$$

Now using (32) and (37), we get the result of Theorem 2. A separate analysis of β_n is provided in Sect. 5.1. □

5 Main insights and numerics

Through Theorem 2, we can write (27) as

$$\mathbb{E} Q_n \approx \tilde{\sigma}_n\, \mathbb{E}[M_{\beta_n}]$$

with

$$\tilde{\sigma}_n = \beta_n \left(\frac{b_n + \rho_n}{1 - \rho_n} \right). \tag{50}$$

This robust approximation for $\mathbb{E} Q_n$ is suggestive of the following two properties that extend beyond the mean system behavior, and hold at the level of approximating the queue by σ_n times the Gaussian random walk:

(i) At the process level, the space should be normalized with σ_n, as in (8). The approximation (27) suggests that it is better to normalize with $\tilde{\sigma}_n$. Although $\tilde{\sigma}_n/\sigma_n \to 1$ for $n \to \infty$, the $\tilde{\sigma}_n$ is expected to lead to sharper approximations for finite n.

(ii) Again at the process level, it seems better to replace the original hedge β by the robust hedge β_n. This thus means that the original system for finite n is approximated by a Gaussian random walk with drift $-\beta_n$. Apart from this approximation being asymptotically correct for $n \to \infty$, it is also expected to approximate the behavior better for finite n.

5.1 Convergence of the robust hedge

We next examine the accuracy of the heavy-traffic approximations for $\mathbb{E}Q_n$ and σ_Q^2, following Corollary 3 and Theorem 2. We expect the robust approximation to be considerably better than the classical approximation when β_n and $\tilde{\sigma}_n$ differ substantially from their limiting counterparts. Before substantiating this claim numerically, we present a result on the convergence rates of β_n to β and $\tilde{\sigma}_n$ to σ_n.

Proposition 1 *Let a_n, b_n and s_n be as in Assumption 2. Then*

$$\beta_n^2 = \beta^2 \left(1 - \frac{1}{1 + b_n + \sigma_n/\beta}\right). \tag{51}$$

Proof From (28), we have

$$\begin{aligned}\beta_n^2 &= s_n \left(\frac{1-\rho_n}{b_n+1}\right)^2 \left(1 + \frac{b_n}{\rho_n}\right) = \frac{1}{s_n}\left(\frac{s_n - a_n b_n}{b_n+1}\right)^2 \left(1 + \frac{s_n}{a_n}\right) \\ &= \frac{1}{s_n} \frac{\beta^2 a_n b_n (b_n+1)}{(b_n+1)^2} \left(1 + \frac{s_n}{a_n}\right) = \beta^2 \frac{b_n}{b_n+1} \left(1 + \frac{a_n}{s_n}\right) =: \beta^2 \bar{F}_n.\end{aligned}$$

Now,

$$\begin{aligned}\bar{F}_n &= \frac{b_n}{b_n+1}\left(1 + \frac{a_n}{s_n}\right) = \frac{b_n}{b_n+1} + \frac{1}{b_n+1}\frac{a_n b_n}{s_n} \\ &= 1 - \frac{1}{b_n+1}\left(1 - \frac{a_n b_n}{s_n}\right) = 1 - \frac{1}{b_n+1}\frac{\beta\,\sigma_n}{s_n} \\ &= 1 - \frac{1}{b_n+1}\frac{1}{1 + \frac{\mu_n}{\beta\sigma_n}} = 1 - \frac{1}{b_n + 1 + \frac{1}{\beta}\sqrt{a_n b_n (b_n+1)}},\end{aligned}$$

which, together with $\sigma_n^2 = a_n b_n (b_n+1)$, proves the proposition. □

Note that β_n always approaches β from below. Also, (51) shows that b_n is the dominant factor in determining the rate of convergence of β_n.

Proposition 2 *Let $\tilde{\sigma}_n$ as in (50). Then*

$$\tilde{\sigma}_n = \sigma_n + b_n \beta_n + O(1).$$

Proof Straightforward calculations give

$$\begin{aligned}\tilde{\sigma}_n &= \beta_n \left(\frac{s_n b_n + a_n b_n}{s_n - a_n b_n}\right) \\ &= \frac{\beta_n}{\beta}\frac{b_n}{\sigma_n}(s_n + a_n) = \frac{\beta_n}{\beta}\sqrt{\frac{b_n}{a_n(b_n+1)}}\left(a_n(b_n+1) + \beta\sqrt{a_n b_n (b_n+1)}\right) \\ &= \frac{\beta_n}{\beta}\left(\sqrt{a_n b_n (b_n+1)} + \beta b_n\right) = \frac{\beta_n}{\beta}\sigma_n + \beta_n b_n.\end{aligned}$$

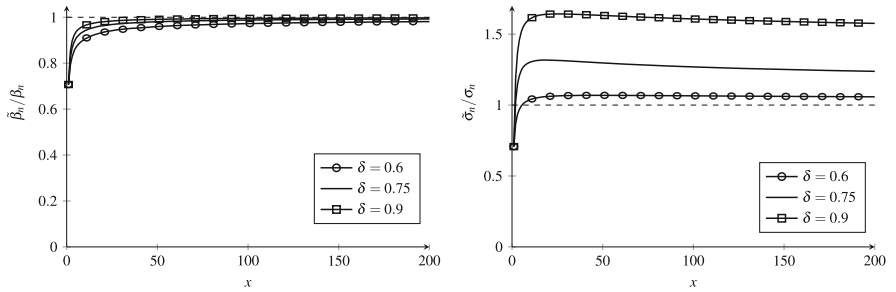

Fig. 1 Convergence of the robust hedge. **a** Convergence of β_n, **b** convergence of $\tilde{\sigma}_n$

Table 1 Numerical results for the Gamma–Poisson case with $\beta = 1$ and $\delta = 0.6$

s_n	ρ_n	$\mathbb{E}Q_n$	(25)	(27)	$\sqrt{\text{Var } Q_n}$	(26)	(29)
5	0.609	0.343	0.246	0.363	1.002	0.835	0.978
10	0.683	0.535	0.400	0.551	1.239	1.063	1.216
50	0.815	1.405	1.168	1.405	1.995	1.817	1.971
100	0.855	2.113	1.824	2.105	2.445	2.270	2.420
500	0.920	5.446	5.006	5.412	3.923	3.762	3.899

Applying Proposition 1 together with the observation

$$\sigma_n \sqrt{1 - \frac{1}{1 + b_n + \sigma_n/\beta}} = \sigma_n(1 + O(1/\sqrt{a_n b_n})) = \sigma_n + O(1)$$

yields the result. □

In Fig. 1, we visualize the convergence speed of both parameters in the case $\mu_n = n$, $\sigma_n = n^\delta$ with $\delta = 0.7$ and $\beta = 1$. This implies $a_n = n/(n^{2\delta} - 1)$ and $b_n = n^{2\delta} - 1$. We observe that β_n starts resembling β fairly quickly, as predicted by Proposition 1; $\tilde{\sigma}_n$, on the other hand, converges extremely slowly to its limiting counterpart. Since $\mathbb{E}Q_n$ and Var Q_n are approximated by $\tilde{\beta}_n$ and $\tilde{\sigma}_n$, multiplied by a term that remains almost constant as n grows, the substitution of σ_n by $\tilde{\sigma}_n$ is essential for obtaining accurate approximations, as we illustrate further in the next subsection.

5.2 Comparison between heavy-traffic approximations

We set $\mu_n = n$ and $\sigma_n^2 = n^{2\delta}$ with $\delta > \frac{1}{2}$, so that $s_n = n + \beta n^\delta$, and $a_n = n/(n^{2\delta-1} - 1)$ and $b_n = n^{2\delta-1} - 1$. Tables 1, 2, 3 and 4 present numerical results for various parameter values. The exact values are calculated using the method in Appendix B. Several conclusions are drawn from these tables. Observe that the heavy-traffic approximations based on the Gaussian random walk, (25) and (26), capture the

Table 2 Numerical results for the Gamma–Poisson case with $\beta = 1$ and $\delta = 0.8$

s_n	ρ_n	$\mathbb{E}Q_n$	(25)	(27)	$\sqrt{\text{Var } Q_n}$	(26)	(29)
5	0.550	0.462	0.284	0.479	1.162	0.896	1.130
10	0.587	0.852	0.521	0.855	1.570	1.213	1.528
50	0.668	3.197	2.093	3.106	3.025	2.433	2.947
100	0.700	5.561	3.784	5.377	3.983	3.270	3.887
500	0.766	19.887	14.741	19.202	7.514	6.455	7.361

Table 3 Numerical results for the Gamma–Poisson case with $\beta = 0.1$ and $\delta = 0.6$

s_n	ρ_n	$\mathbb{E}Q_n$	(25)	(27)	$\sqrt{\text{Var } Q_n}$	(26)	(29)
5	0.949	11.532	11.306	11.495	3.634	3.559	3.602
10	0.961	17.565	17.268	17.548	4.474	4.398	4.444
50	0.979	46.368	45.869	46.418	7.241	7.168	7.218
100	0.984	70.340	69.735	70.430	8.910	8.839	8.888
500	0.991	184.900	183.989	185.108	14.422	14.357	14.404

Table 4 Numerical results for the Gamma–Poisson case with $\beta = 0.1$ and $\delta = 0.8$

s_n	ρ_n	$\mathbb{E}Q_n$	(25)	(27)	$\sqrt{\text{Var } Q_n}$	(26)	(29)
5	0.931	15.730	15.209	15.909	4.276	4.127	4.233
10	0.939	27.561	26.672	27.958	5.652	5.466	5.605
50	0.955	100.660	97.967	102.070	10.760	10.476	10.698
100	0.961	175.591	171.360	177.818	14.189	13.855	14.117
500	0.971	638.097	626.346	644.105	26.963	26.490	26.864

right order of magnitude for both $\mathbb{E}Q_n$ and Var Q_n. However, the values are off, in particular for small s_n and relatively low $\rho_n := \mathbb{E}[A_n]/s_n$. The inaccuracy also increases with the level of overdispersion. In contrast, the approximations that follow from Theorem 2, (27) and (29), are remarkably accurate. Even for small systems with $s_n = 5$ or 10, the approximations for $\mathbb{E}Q_n$ are within 6% of the exact value for small ρ_n and within 2% for ρ_n close to 1. For σ_Q^2, these percentages even reduce to 3% and 1%, respectively. For larger values of s_n these relative errors naturally reduce further. Overall, we observe that the approximations improve for heavily loaded systems, and the corrected approximations are particularly useful for systems with increased overdispersion.

Acknowledgements The authors are grateful to Avi Mandelbaum for many inspiring discussions and comments. The research of BM was supported by NWO Free Competition Grant No. 613.001.213. The research of JvL is supported by an ERC Starting Grant and by NWO Gravitation Networks Grant No. 024.002.003. The research of BZ is supported by NWO VICI Grant No. 639.033.413.

A Proofs of convergence results

This section presents the details of the proofs of Lemma 1 and Theorem 1, using the random walk perspective of the process $\{Q_{k,n}\}_{k=0}^{\infty}$. This section is structured as follows. The next two lemmata are necessary for proving the first assertion of Theorem 1, concerning the weak convergence of the scaled process to the maximum of the Gaussian random walk, which is summarized in Proposition 4. The two remaining propositions of this section show convergence of $\hat{Q}_n$ at the process level as well as in terms of the three characteristics.

Let us first fix some notation:

$$Y_{k,n} := \hat{A}_{k,n} - \beta, \quad S_{k,n} = \sum_{i=1}^{k} Y_{i,n}, \tag{52}$$

with $S_{0,n} = 0$ and $k = 1, 2, \ldots$. Then (6) can be rewritten as

$$\hat{Q}_n \stackrel{d}{=} \max_{0 \le k} \left\{ \sum_{i=1}^{k} Y_{i,n} \right\} =: M_{\beta,n}. \tag{53}$$

Last, we introduce the sequence of independent normal random variables $Z_1, Z_2, \ldots$ with mean $-\beta$ and unit variance 1, and

$$M_\beta \stackrel{d}{=} \max_{k \ge 0} \left\{ \sum_{i=1}^{k} Z_i \right\}.$$

A.1 Proof of Lemma 1

Proof We show weak convergence of the random variable $\hat{A}_n$, as defined as the common distribution of (7), to a standard normal random variable. Since $\hat{\Lambda}_n$ is asymptotically standard normal, its characteristic function converges pointwise to the corresponding limiting characteristic function, i.e.,

$$\lim_{n\to\infty} \phi_{\hat{\Lambda}_n}(t) = \lim_{n\to\infty} \mathrm{e}^{-i\mu_n t/\sigma_n}\, \phi_{\Lambda_n}(t/\sigma_n) = \mathrm{e}^{-t^2/2}, \quad \forall t \in \mathbb{R}. \tag{54}$$

Furthermore, by the definition of A_n,

$$\phi_{A_n}(t) = \mathbb{E}\left[\exp(\Lambda_n(\mathrm{e}^{it} - 1))\right] = \phi_{\Lambda_n}\left(-i(\mathrm{e}^{it} - 1)\right),$$

so that

$$\phi_{\hat{A}_n}(t) = \mathrm{e}^{-i\mu_n t/\sigma_n}\,\phi_{A_{k,n}}(t/\sigma_n) = \mathrm{e}^{-i\mu_n t/\sigma_n}\phi_{\Lambda_n}\left(-i(\mathrm{e}^{it/\sigma_n}-1)\right). \tag{55}$$

Now fix $t \in \mathbb{R}$. By using

$$-i(\mathrm{e}^{it/\sigma_n}-1) = \frac{t}{\sigma_n} + O\left(t^2/\sigma_n^2\right),$$

we expand the last term in (55),

$$\phi_{\Lambda_n}(t/\sigma_n) + O\left(t^2/\sigma_n^2\right)\phi'_{\Lambda_n}(\zeta)$$

for some ζ such that $|\zeta - t/\sigma_n| < |i(1-\mathrm{e}^{it/\sigma_n}) - t/\sigma_n|$. Also,

$$\begin{aligned}|\phi'_{\Lambda_n}(u)| &= \left|\frac{\delta}{\mathrm{d}u}\int_{-\infty}^{\infty}\mathrm{e}^{iux}\mathrm{d}F_{\Lambda_n}(x)\right| = \left|\int_0^{\infty} ix\,\mathrm{e}^{iux}\mathrm{d}F_{\Lambda_n}(x)\right| \\ &\le \int_{-\infty}^{\infty}|ix\,\mathrm{e}^{iux}|\,\mathrm{d}F_{\Lambda_n}(x) = \int_0^{\infty} x\mathrm{d}F_{\Lambda_n}(x) = \mu_n \end{aligned} \tag{56}$$

for all $u \in \mathbb{R}$. This implies

$$\phi_{\hat{A}_{k,n}}(t) = \phi_{\Lambda_n}(t/\sigma_n) + O\left(t^2\mu_n/\sigma_n^2\right),$$

in which the order term tends to zero as $n \to \infty$ by our assumption that $\mu_n/\sigma_n^2 \to 0$. Combining this with (54), we find that $\phi_{\hat{A}_n}(t)$ converges to $\mathrm{e}^{-t^2/2}$ for all $t \in \mathbb{R}$, so that we can conclude by Lévy's continuity theorem that $\hat{A}_{k,n} \overset{d}{\Rightarrow} N(0,1)$. □

A.2 Proof of Theorem 1

To secure convergence in distribution of $\hat{Q}_n$ to M_β, i.e., the maximum of a Gaussian random walk with negative drift, the following property of the sequence $\{Y_{k,n}\}_{n\in\mathbb{N}}$ needs to hold. Because the sequence $\{Y_{k,n}\}_{k\in\mathbb{N}}$ is i.i.d. for all n, we omit the index k in this result and its proof.

Lemma 3 *Let Y_n be defined as in (52) with $\mu_n, \sigma_n^2 < \infty$ for all $n \in \mathbb{N}$. Then, the sequence $\{(Y_n)^+\}_{n\in\mathbb{N}}$ is uniformly integrable.*

Proof Note that the sequence $\{Y_n\}_{n\in\mathbb{N}}$ has constant finite mean and variance equal to 0 and 1, respectively, which implies it is bounded in L^2. Since $(Y_n^+)^2 \le Y_n^2$, the sequence $\{Y_n^+\}_{n\in\mathbb{N}}$ is bounded in L^2 as well. It readily follows from elementary probability theory that any sequence bounded in L^2 is uniformly integrable; see, for example, [7]. □

By combining the properties proved in Lemmas 1 and 3 with Assumption 1, the next result follows directly by [3, Thm. X6.1].

Proposition 3 *Let* $\hat{Q}_n$ *be as in* (53)*. Then,*

$$\hat{Q}_n \overset{d}{\Rightarrow} M_\beta, \quad \text{as } n \to \infty.$$

Although Proposition 3 tells us that the properly scaled Q_n converges to a non-degenerate limiting random variable, it does not cover the convergence of its mean, variance and the empty-queue probability. In order to secure convergence of these performance measures as well, we follow an approach similar [37], using Assumptions 1 and 3.

Proposition 4 *Let* $\hat{Q}_n$ *be as in* (53)*,* $\mu_n, \sigma_n^2 \to \infty$ *such that both* $\sigma_n^2/\mu_n \to \infty$ *and* $\mathbb{E}[(\max\{\hat{A}_n, 0\})^m]$ *is bounded in* n *for* $m = 3, 4$*. Then*

$$\begin{aligned} \mathbb{P}(\hat{Q}_n = 0) &\to \mathbb{P}(M_\beta = 0), \\ \mathbb{E}[\hat{Q}_n] &\to \mathbb{E}[M_\beta], \\ \text{Var } \hat{Q}_n &\to \text{Var } M_\beta, \end{aligned}$$

as $n \to \infty$*.*

Proof First, we recall that $\hat{Q}_n \overset{d}{=} M_{\beta,n}$ for all $n \in \mathbb{N}$, so that $\mathbb{P}(\hat{Q}_n = 0) = \mathbb{P}(M_{\beta,n} = 0)$, $\mathbb{E}[\hat{Q}_n] = \mathbb{E}[M_{\beta,n}]$ and Var $\hat{Q}_n$ = Var $M_{\beta,n}$ as defined in (52). Our starting point is Spitzer's identity, see [3, p. 230],

$$\mathbb{E}[\mathrm{e}^{itM_{\beta,n}}] = \exp\left(\sum_{k=1}^{\infty} \frac{1}{k}(\mathbb{E}[\mathrm{e}^{itS_{k,n}^+}] - 1)\right), \tag{57}$$

with $S_{k,n}$ as in (52), and $M_{\beta,n}$ the all-time maximum of the associated random walk. Simple manipulations of (57) give

$$\ln \mathbb{P}(M_{\beta,n} = 0) = -\sum_{k=1}^{\infty} \frac{1}{k} \mathbb{P}(S_{k,n} > 0), \tag{58}$$

$$\mathbb{E}[M_{\beta,n}] = \sum_{k=1}^{\infty} \frac{1}{k} \mathbb{E}[S_{k,n}^+] = \sum_{k=1}^{\infty} \frac{1}{k} \int_0^{\infty} \mathbb{P}(S_{k,n} > x) \mathrm{d}x, \tag{59}$$

$$\text{Var } M_{\beta,n} = \sum_{k=1}^{\infty} \frac{1}{k} \mathbb{E}[(S_{k,n}^+)^2] = \sum_{k=1}^{\infty} \frac{1}{k} \int_0^{\infty} \mathbb{P}(S_{k,n} > \sqrt{x}) \mathrm{d}x. \tag{60}$$

By Lemma 1, we know

$$\mathbb{P}(S_{k,n} > y) = \mathbb{P}\left(\sum_{i=1}^{k} Y_{i,n} > y\right) \to \mathbb{P}\left(\sum_{i=1}^{k} Z_i > y\right),$$

for $n \to \infty$, where the Z_i are independent and identically normally distributed with mean $-\beta$ and variance 1. Because equivalent expressions to (58)–(60) apply to the limiting Gaussian random walk, it is sufficient to show that the sums converge uniformly in n, so that we can apply dominated convergence to prove the result.

We start with the empty-queue probability. To justify interchangeability of the infinite sum and limit, note

$$\mathbb{P}(S_{k,n} > 0) \le \mathbb{P}(|S_{k,n} + k\beta| > k\beta) \le \frac{k}{\beta^2 k^2} = \frac{1}{\beta^2 k},$$

where we used that $\mathbb{E}[S_{k,n}] = k\mathbb{E}[Y_{1,n}] = -k\beta$ and Var $S_{k,n} = k$, and apply Chebyshev's inequality, so that

$$\sum_{k=1}^{\infty} \frac{1}{k} \mathbb{P}(S_{k,n} > 0) \le \sum_{k=1}^{\infty} \frac{1}{\beta^2 k^2} < \infty, \quad \forall n \in \mathbb{N}.$$

Hence,

$$\begin{aligned} \lim_{n\to\infty} \ln \mathbb{P}(\hat{Q}_n = 0) &= \lim_{n\to\infty} -\sum_{k=1}^{\infty} \frac{1}{k} \mathbb{P}(S_{k,n} > 0) = -\sum_{k=1}^{\infty} \frac{1}{k} \lim_{n\to\infty} \mathbb{P}(S_{k,n} > 0) \\ &= -\sum_{k=1}^{\infty} \frac{1}{k} \mathbb{P}\left(\sum_{i=1}^{k} Z_i > 0\right) = \ln \mathbb{P}(M_\beta = 0). \end{aligned}$$

Finding a suitable upper bound on $\frac{1}{k}\int_0^\infty \mathbb{P}(\hat{Q}_n > x)\mathrm{d}x$ and $\frac{1}{k}\int_0^\infty \mathbb{P}(\hat{Q}_n > \sqrt{x})\mathrm{d}x$ requires a bit more work. We initially focus on the former, and the latter follows easily. The following inequality from [32] proves to be very useful:

$$\mathbb{P}(\bar{S}_k > y) \le C_r \left(\frac{k\sigma^2}{y^2}\right)^r + k\,\mathbb{P}(X > y/r), \tag{61}$$

where $\bar{S}_k$ is the sum of k i.i.d. random variables distributed as X, with $\mathbb{E}[X] = 0$ and Var $X = \sigma^2$, $y > 0$, $r > 0$ and C_r a constant only depending on r. We take $r = 3$ for brevity in the remainder of the proof, although any $r > 2$ will suffice. We have, from (61) with $X = \hat{A}_n$, so that $\mathbb{E}[X] = 0$, Var $X = 1$, and $r = 3$, $y = x + k\beta$,

$$\begin{aligned} \mathbb{P}(S_{k,n} > x) &= \mathbb{P}\left(\sum_{i=1}^{k} \hat{A}_{i,n} > x + k\beta\right) \\ &\le C_3 \left(\frac{k}{(x+k\beta)^2}\right)^3 + k\,\mathbb{P}\left(\hat{A}_{1,n} > \frac{x+k\beta}{3}\right). \end{aligned} \tag{62}$$

The quantity $(k/(x+k\beta)^2)^3$ is independent of n, and we have

$$\sum_{k=1}^{\infty} \frac{1}{k} \int_0^{\infty} \left(\frac{k}{(x+k\beta)^2} \right)^3 \mathrm{d}x = \sum_{k=1}^{\infty} k^2 \int_0^{\infty} \frac{\mathrm{d}x}{(x+k\beta)^6} = \sum_{k=1}^{\infty} \frac{k^2}{5(k\beta)^5} = \frac{1}{5\beta^5} \sum_{k=1}^{\infty} \frac{1}{k^3} < \infty. \tag{63}$$

Next, by assumption, there is an $M_3 > 0$ such that

$$\mathbb{E}[(\max\{\hat{A}_{1,n}, 0\})^3] = \int_0^{\infty} t^3 \, \mathrm{d}P_{\hat{A}_{1,n}}(t) \le M_3 \tag{64}$$

for all $n = 1, 2, \ldots$. It follows that for all $x > 0$, $k = 1, 2, \ldots$, and all $n = 1, 2, \ldots$,

$$\mathbb{P}\left(\hat{A}_{1,n} > \frac{x+k\beta}{3} \right) = \int_{\frac{x+k\beta}{3}}^{\infty} \mathrm{d}P_{\hat{A}_{1,n}}(t) \le \int_{\frac{x+k\beta}{3}}^{\infty} \frac{t^3}{\left(\frac{x+k\beta}{3}\right)^3} \mathrm{d}P_{\hat{A}_{1,n}}(t) \le \frac{27}{(x+k\beta)^3} \mathbb{E}[(\max\{\hat{A}_{1,n}, 0\})^3] \le \frac{27M_3}{(x+k\beta)^3}. \tag{65}$$

The quantity $1/(x+k\beta)^3$ is independent of n, and we have

$$\sum_{k=1}^{\infty} \int_0^{\infty} \frac{\mathrm{d}x}{(x+k\beta)^3} = \sum_{k=1}^{\infty} \frac{1}{2k^2\beta^2} < \infty. \tag{66}$$

Thus we see that for all $x > 0$, $k = 1, 2, \ldots$, and all $n = 1, 2, \ldots$,

$$\frac{1}{k} \mathbb{P}(S_{k,n} > x) \le \frac{C_3 k^2}{(x+k\beta)^6} + \frac{27M_3}{(x+k\beta)^3}, \tag{67}$$

with

$$\sum_{k=1}^{\infty} \int_0^{\infty} \left(\frac{C_3 k^2}{(x+k\beta)^6} + \frac{27M_3}{(x+k\beta)^3} \right) \mathrm{d}x < \infty. \tag{68}$$

By Lebesgue's theorem on dominated convergence, it then follows that

$$\lim_{n\to\infty} \mathbb{E}[\hat{Q}_n] = \lim_{n\to\infty} \sum_{k=1}^{\infty} \frac{1}{k} \int_0^{\infty} \mathbb{P}(S_{k,n} > x) \, \mathrm{d}x = \sum_{k=1}^{\infty} \int_0^{\infty} \mathbb{P}\left(\sum_{i=1}^{k} Z_i > x \right) \mathrm{d}x = \mathbb{E}[M_\beta]. \tag{69}$$

In a similar fashion, by using (61) with $y = \sqrt{x} + k\beta$, we have

$$\mathbb{P}(S_{k,n} > \sqrt{x}) \leq C_3 \left(\frac{k}{(\sqrt{x} + k\beta)^2} \right)^3 + k\, \mathbb{P}\left(\hat{A}_{1,n} > \frac{\sqrt{x} + k\beta}{3} \right). \tag{70}$$

Now we have

$$\sum_{k=1}^{\infty} \frac{1}{k} \int_0^{\infty} \left(\frac{k}{(\sqrt{x} + k\beta)^2} \right)^3 \mathrm{d}x = \frac{1}{10\beta^4} \sum_{k=1}^{\infty} \frac{1}{k^2} < \infty, \tag{71}$$

where we have used that, by partial integration,

$$\int_0^{\infty} \frac{\mathrm{d}x}{(\sqrt{x} + k\beta)^6} = \int_0^{\infty} \frac{2u\, \mathrm{d}u}{(u + k\beta)^6} = \frac{1}{10(k\beta)^4}. \tag{72}$$

Next, by assumption, there in an $M_4 > 0$ such that

$$\mathbb{E}[(\max\{\hat{A}_{1,n}, 0\})^4] = \int_0^{\infty} t^4\, \mathrm{d}P_{\hat{A}_{1,n}}(t) \leq M_4 \tag{73}$$

for all $n = 1, 2, \ldots$. Then, as in (65), we get that for all $x > 0$, $k = 1, 2, \ldots$, and all $n = 1, 2, \ldots$,

$$\mathbb{P}\left(\hat{A}_{1,n} > \frac{\sqrt{x} + k\beta}{3} \right) \leq \frac{81 M_4}{(\sqrt{x} + k\beta)^4}, \tag{74}$$

while, as in (66) and (72),

$$\sum_{k=1}^{\infty} \int_0^{\infty} \frac{\mathrm{d}x}{(\sqrt{x} + k\beta)^4} = \sum_{k=1}^{\infty} \frac{1}{3k^2\beta^2} < \infty. \tag{75}$$

Therefore, for all $x > 0$, $k = 1, 2, \ldots$, and all $n = 1, 2, \ldots$,

$$\frac{1}{k}\, \mathbb{P}\left(S_{k,n} > \sqrt{x} \right) \leq \frac{C_3 k^2}{(\sqrt{x} + k\beta)^6} + \frac{81 M_4}{(\sqrt{x} + k\beta)^4}, \tag{76}$$

with

$$\sum_{k=1}^{\infty} \int_0^{\infty} \left(\frac{C_3 k^2}{(\sqrt{x} + k\beta)^6} + \frac{81 M_4}{(\sqrt{x} + k\beta)^4} \right) \mathrm{d}x < \infty. \tag{77}$$

Hence, we conclude, as in (69), that $\lim_{n\to\infty} \mathrm{Var}\, \hat{Q}_n = \mathrm{Var}\, M_\beta$. This completes the proof. □

We shall now indicate that the condition in Proposition 4 on boundedness of $\mathbb{E}[(\max\{\hat{A}_n, 0\})^m]$ in n for $m = 3, 4$ is not overly restrictive by showing that it is satisfied by the Poisson mixture A_n considered in Sect. 2. With $p_{n,k} = \mathbb{P}(A_n = k)$, we have that the pgf $\tilde{A}_n$ of A is given by

$$\tilde{A}_n(z) = \left(\frac{1}{1 + b_n - b_n z}\right)^{a_n} = \sum_{k=0}^{\infty} p_{n,k} z^k. \tag{78}$$

For any $m = 1, 2, \ldots$, we have

$$\mathbb{E}[(\max\{\hat{A}_n, 0\})^m] = \mathbb{E}\left[\left(\max\left\{\frac{A_n - \mu_n}{\sigma_n}, 0\right\}\right)^m\right] = \frac{1}{\sigma_n^m} \sum_{k \geq \mu_n} (k - \mu_n)^m p_{n,k}. \tag{79}$$

We shall conduct a somewhat heuristical saddle point analysis for the Cauchy integral representation

$$p_{k-1} = \frac{1}{2\pi i} \oint \frac{1}{z^k} \left(\frac{1}{1 + b - bz}\right)^a \mathrm{d}z, \quad k \geq \mu, \tag{80}$$

with integration along a circle with center 0 and radius less than $(1 + b)/b$. For convenience, we have temporarily omitted the index n in $p_{n,k-1}$, a_n, b_n and μ_n. The saddle point z_{sp} lies on the positive real axis between 0 and $(1 + b)/b$ and is obtained by setting $f'(z) = 0$, where

$$f(z) = -a \ln(1 + b - bz) - k \ln z. \tag{81}$$

When $k \geq \mu = ab$, it is found that

$$z_{\mathrm{sp}} = 1 + \frac{k - ab}{(k + a)b}. \tag{82}$$

We have, moreover,

$$\begin{aligned} f(z_{\mathrm{sp}}) &= -a \ln\left(1 - \frac{k - ab}{k + a}\right) - k \ln\left(1 + \frac{k - ab}{(k + a)b}\right) \\ &= -\frac{(k - ab)^2}{(k + a)b}\left(\tfrac{1}{2} + O\left(\frac{k - ab}{k + a}\right)\right) \end{aligned} \tag{83}$$

and

$$f''(z_{\mathrm{sp}}) = \frac{1}{z_{\mathrm{sp}}^2} \frac{k^2}{a} + \frac{k}{z_{\mathrm{sp}}^2} = ab^2 \left(1 + O\left(\frac{k - ab}{ab}\right) + O\left(\frac{1}{b}\right)\right). \tag{84}$$

Thus, we find the saddle point approximation

$$p_{k-1} \approx \frac{1}{2\pi}\sqrt{\frac{2\pi}{f''(z_{\rm sp})}}\,\exp\big(f(z_{\rm sp})\big) \approx \frac{1}{\sqrt{2\pi ab^2}}\,\exp\left(-\frac{(k-ab)^2}{2ab^2}\right), \tag{85}$$

with k restricted to the range $0 \le k - ab = O(ab)$. This range of k is sufficiently large for the series and the integral in (86) below to have converged. We now use this approximation in (79). Thus we have, using $\sigma_n^2 = a_n b_n(b_n+1) \approx a_n b_n^2$ and $\mu_n = a_n b_n$,

$$\begin{aligned}
\mathbb{E}[(\max\{\hat{A}_n, 0\})^m] &= \frac{1}{\sigma_n^m}\sum_{k\ge\mu_n}(k-\mu_n)^m p_{n,k} \\
&\approx \frac{1}{\sqrt{2\pi}}\left(\frac{1}{a_n b_n^2}\right)^{\frac{m+1}{2}}\sum_{k\ge\mu_n}(k-\mu_n)^m\exp\left(-\frac{(k-\mu_n)^2}{2a_n b_n^2}\right) \\
&\approx \frac{1}{\sqrt{2\pi}}\left(\frac{1}{a_n b_n^2}\right)^{\frac{m+1}{2}}\int_0^\infty t^m\exp\left(-\frac{t^2}{2a_n b_n^2}\right)\mathrm{d}t \\
&= \frac{2^{\frac{m-1}{2}}}{\sqrt{2\pi}}\,\Gamma\left(\frac{m+1}{2}\right).
\end{aligned} \tag{86}$$

The final member of (86) is independent of n, and this provides evidence that $\mathbb{E}[(\max\{\hat{A}_n, 0\})^m]$ is bounded.

B Numerical procedures

An alternative characterization of the stationary distribution is based on the analysis in [10] and considers a factorization in terms of (complex) roots:

$$Q_n(w) = \frac{(s_n - \mathbb{E}[A_n])(w-1)}{w^{s_n} - \tilde{A}_n(w)}\prod_{k=1}^{s_n-1}\frac{w - z_k^n}{1 - z_k^n},$$

where $z_1^n, z_2^n \ldots, z_{s_n-1}^n$ are the $s_n - 1$ zeros of $z^{s_n} - \tilde{A}_n(z)$ in $|z| < 1$, yielding

$$\begin{aligned}
\mathbb{E}Q_n &= \frac{\sigma_n^2}{2(s_n-\mu_n)} - \frac{s_n - 1 + \mu_n}{2} + \sum_{k-1}^{s_n-1}\frac{1}{1-z_k^n}, \\
\mathbb{P}(Q_n = 0) &= \frac{s_n - \mu_A}{\tilde{A}_n(0)}\prod_{k=1}^{s-1}\frac{z_k^n}{z_k^n - 1},
\end{aligned}$$

which for our choice of $\tilde{A}_n(z)$ becomes

$$\mathbb{E}Q_n = \frac{a_n b_n(b_n+1)}{2\beta\sqrt{a_n}b_n} - \frac{2a_n b_n + \beta\sqrt{a_n b_n(b_n+1)} - 1}{2} + \sum_{k=1}^{s_n-1} \frac{1}{1-z_k^n},$$

$$\mathbb{P}(Q_n = 0) = \beta\sqrt{a_n b_n(b_n+1)}(1+b_n)^{a_n} \prod_{k=1}^{s_n-1} \frac{z_k^n}{z_k^n - 1},$$

where $z_1, \ldots, z_{s_n-1}$ denote the zeros of $z^{s_n} - \tilde{A}_n(z)$ in $|z| < 1$. These zeros exist under the assumption $s_n > a_n b_n$; see [2]. A robust numerical procedure to obtain these zeros is essential for a base of comparison. We discuss two methods that fit these requirements. The first follows directly from [20].

Lemma 4 *Define the iteration scheme*

$$z_k^{n,l+1} = w_k^n [\tilde{A}_n(z_k^{n,l})]^{1/s_n}, \tag{87}$$

with $w_k^n = \mathrm{e}^{2\pi i k/s_n}$ *and* $z_k^{n,0} = 0$ *for* $k = 0, 1, \ldots, s_{n-1}$. *Then* $z_k^{n,l} \to z_k^n$ *for all* $k = 0, 1, \ldots, s_n - 1$ *for* $l \to \infty$.

Proof The successive substitution scheme given in (87) is the fixed point iteration scheme described in [20], applied to the pgf of our arrival process. The authors show that, under the assumption of $\tilde{A}_n(z)$ being zero-free within $|z| \le 1$, the zeros can be approximated arbitrarily closely, given that the function $[\tilde{A}_n(z)]^{1/s_n}$ is a contraction for $|z| \le 1$, i.e.,

$$\left| \frac{\mathrm{d}}{\mathrm{d}z} [\tilde{A}_n(z)]^{1/s_n} \right| < 1.$$

In our case,

$$\left| \frac{\mathrm{d}}{\mathrm{d}z} [\tilde{A}_n(z)]^{1/s_n} \right| = \left| \frac{\mathrm{d}}{\mathrm{d}z} (1 + (1-z)b_n)^{-a_n/s_n} \right| = \frac{a_n b_n}{s_n} |1 + (1-z)b_n|^{-a_n/s_n - 1}, \tag{88}$$

where $a_n b_n / s_n = \rho_n$ is close to, but less than, 1 and

$$|1 + (1-z)b_n| \ge |1 + b_n| - |z| b_n = 1 + (1 - |z|)b_n \ge 1,$$

when $|z| \le 1$. Hence, the expression in (88) is less than 1 for all $|z| \le 1$. Evidently, $\tilde{A}_n(z)$ is also zero-free in $|z| \le 1$. Thus, [20, Lemma 3.8] shows that $z_k^{n,l}$ as in (87) converges to the desired roots z_k^n for all k as l tends to infinity. □

Remark 2 The asymptotic convergence rate of the iteration in (87) equals $\frac{\mathrm{d}}{\mathrm{d}z}[\tilde{A}_n(z)]^{1/s_n}$ evaluated at $z = z_k^n$. Hence, convergence is slow for zeros near 1 and fast for zeros away from 1.

A different approach is based on the Bürmann–Lagrange inversion formula.

Lemma 5 *Let $w_k^n = e^{2\pi ik/s_n}$ and $\alpha_n = a_n/s_n$. Then, the zeros of $z^{s_n} - \tilde{A}_n(z)$ are given by*

$$z_k^n = \sum_{l=1}^{\infty} \frac{1}{l!} \frac{\beta(l\alpha_n + l - 1)}{\beta(l\alpha_n)} \frac{b_n + 1}{b_n} \left(\frac{b_n}{(b_n+1)^{\alpha_n+1}} \right)^l (w_k^n)^l,$$

for $k = 0, 1, \ldots, s_n - 1$.

Proof Note that we are looking for z's that solve

$$z\,[\tilde{A}_n(z)]^{-1/s_n} = z\,(1 + (1-z)b_n)^{a_n/s_n} = w,$$

where $w = w_k = \mathrm{e}^{2\pi ik/s_n}$. The Bürmann–Lagrange formula for $z = z(w)$, as can be found in [15, Sec. 2.2] for $z = z(w)$, is given by

$$\begin{aligned} z(w) &= \sum_{l=1}^{\infty} \frac{1}{l!} \left(\frac{\mathrm{d}}{\mathrm{d}z} \right)^{l-1} \left[\left(\frac{z}{z(1 + (1-z)b_n)^{a_n/s_n}} \right)^l \right]_{z=0} w^l \\ &= \sum_{l=1}^{\infty} \frac{1}{l!} \left(\frac{\mathrm{d}}{\mathrm{d}z} \right)^{l-1} \left[(1 + (1-z)b_n)^{-l\,a_n/s_n} \right]_{z=0} w^l. \end{aligned}$$

Set $\alpha_n = a_n/s_n$. We compute

$$\left(\frac{\mathrm{d}}{\mathrm{d}z} \right)^{l-1} \left[(1 + (1-z)b_n)^{-l\alpha_n} \right]_{z=0} = \frac{\beta(l\alpha_n + l - 1)}{\beta(l\alpha_n)} \frac{1 + b_n}{b_n} \left(\frac{b_n}{(1+b_n)^{\alpha_n+1}} \right)^l.$$

With $c_n = b_n/(1+b_n)^{\alpha_n+1}$ and $d_n = (1+b_n)/b_n$, we thus have

$$z(w) = d_n \sum_{l=1}^{\infty} \frac{\beta(l\alpha_n + l - 1)}{\beta(l+1)\beta(l\alpha_n)} c_n^l\, w^l.$$

By Stirling's formula

$$\frac{\beta(l\alpha_n + l - 1)}{\beta(l+1)\beta(l\alpha_n)} = \frac{D}{l\sqrt{l}} \left(\frac{(\alpha_n+1)^{\alpha_n+1}}{\alpha_n^{\alpha_n}} \right)^l,$$

where $D = \alpha_n^{1/2}(\alpha_n+1)^{-3/2}(2\pi)^{-1/2}$. Now,

$$\frac{(\alpha_n+1)^{\alpha_n+1}}{\alpha_n^{\alpha_n}} c_n = \frac{(\alpha_n+1)^{\alpha_n+1}}{\alpha_n^{\alpha_n}} \cdot \frac{b_n}{(1+b_n)^{\alpha_n+1}} = \left(\frac{b_n + \rho_n}{b_n + 1} \right)^{\rho_n/b_n+1} \left(\frac{1}{\rho_n} \right)^{\rho_n/b_n}.$$

This determines the radius of convergence r_{BL} of the above series for $z(w)$:

$$\frac{1}{r_{\mathrm{BL}}} := \left(\frac{b_n+\rho_n}{b_n+1}\right)^{\rho_n/b_n+1}\left(\frac{1}{\rho_n}\right)^{\rho_n/b_n}. \tag{89}$$

The derivative with respect to ρ_n of the quantity

$$\left(1+\frac{\rho_n}{b_n}\right)\ln\left(\frac{b_n+\rho_n}{b_n+1}\right)+\frac{\rho_n}{b_n}\ln\left(\frac{1}{\rho_n}\right) \tag{90}$$

is given by

$$\frac{1}{b_n}\ln\left(\frac{b_n+\rho_n}{b_n\rho_n+\rho_n}\right) > 0$$

for $0 < \rho_n < 1$ and $b_n > 0$. Furthermore, the quantity in (90) vanishes at $\rho_n = 1$ and is therefore negative for $0 < \rho_n < 1$ and $b_n > 0$. □

Remark 3 The formula for the radius of convergence in (89) clearly shows the decremental effect of both having a large b_n and/or having ρ_n close to 1. The quantities $\beta(l\alpha+l-1)/(\beta(l+1)\beta(l\alpha))$ in the power series for $z(w)$ are not very convenient for recursive computation, although normally $\alpha = a_n/s_n$ is a rational number.

References

1. Abate, J., Choudhury, G.L., Whitt, W.: Calculation of the $GI/G/1$ waiting-time distribution and its cumulants from Pollaczek's formulas. Archiv fur Elektronik und Ubertragungstechnik (Int. J. Electron. Commun.) **47**(5/6), 311–321 (1993)
2. Adan, I.J.B.F., van Leeuwaarden, J.S.H., Winands, E.M.M.: On the application of Rouché's theorem in queueing theory. Oper. Res. Lett. **34**(3), 355–360 (2006)
3. Asmussen, S.: Applied Probability and Queues, 2nd edn. Springer, New York (2003)
4. Avramidis, A.N., Deslauriers, A., L'Ecuyer, P.: Rate-based daily arrival process models with application to call centers. Manag. Sci. **50**(7), 893–908 (2004)
5. Bassamboo, A., Randhawa, R.S., Zeevi, A.: Capacity sizing under parammeter uncertainty: safety staffing principles revisited. Manag. Sci. **56**(10), 1668–1686 (2010)
6. Bassamboo, A., Zeevi, A.: On a data-driven method for staffing large call centers. Oper. Res. **57**(3), 714–726 (2009)
7. Billingsley, P.: Probability and Measure, 3rd edn. Wiley, Hoboken (1995)
8. Boon, M.A.A., Janssen, A.J.E.M., van Leeuwaarden, J.S.H.: Heavy-traffic limits for dimensioning fixed-cycle intersections. Working paper (2017)
9. Boon, M.A.A., Janssen, A.J.E.M., van Leeuwaarden, J.S.H.: Pollaczek contour integrals for the fixed-cycle traffic-light queue. arXiv:1701.02872 (preprint) (2017)
10. Boudreau, P.E., Griffin Jr., J.S., Kac, M.: An elementary queueing problem. Am. Math. Mon. **69**(8), 713–724 (1962)
11. Brown, L.D., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., Zhao, L.: Statistical analysis of a telephone call center: a queueing-science perspective. J. Am. Stat. Assoc. **100**(469), 36–50 (2005)
12. Chen, B.P.K., Henderson, S.G.: Two issues in setting call center staffing levels. Ann. Oper. Res. **108**(1), 175–192 (2001)
13. Cohen, J.W.: The Single Server Queue, Volume 8 of North-Holland Series in Applied Mathematics and Mechanics, 2nd edn. North-Holland Publishing Co., Amsterdam (1982)

14. Cox, D.R.: Some statistical models connected with series of events. J. R. Stat. Soc. **17**(2), 129–164 (1955)
15. de Bruijn, N.G.: Asymptotic Methods in Analysis, 3rd edn. Dover Publications Inc., New York (1981)
16. Fristedt, B.E., Gray, L.F.: A Modern Approach to Probability Theory. Birkhuser, Boston (1996)
17. Gans, N., Koole, G., Mandelbaum, A.: Telephone call centers: tutorial, review, and research prospects. Manuf. Serv. Oper. Manag. **5**(2), 79–141 (2003)
18. Grandell, J.: Mixed Poisson Processes. Monographs on Statistics and Applied Probability. Chapman & Hall, Boca Raton (1997)
19. Gurvich, I., Luedtke, J., Tezcan, T.: Staffing call-centers with uncertain demand forecasts: a chance-constrained optimization approach. Manag. Sci. **56**(7), 1093–1115 (2010)
20. Janssen, A.J.E.M., van Leeuwaarden, J.S.H.: Analytic computation schemes for the discrete-time bulk service queue. Queueing Syst. **50**(2), 141–163 (2005)
21. Janssen, A.J.E.M., van Leeuwaarden, J.S.H.: Back to the roots of the $M/D/s$ queue and the works of Erlang, Crommelin, and Pollaczek. Stat. Neerl. **62**(3), 299–313 (2008)
22. Janssen, A.J.E.M., van Leeuwaarden, J.S.H., Mathijsen, B.W.J.: Novel heavy-traffic regimes for large-scale service systems. SIAM J. Appl. Math. **75**(2), 787–812 (2015)
23. Jongbloed, G., Koole, G.: Managing uncertainty in call centres using Poisson mixtures. Appl. Stoch. Models Bus. Ind. **17**(4), 307–318 (2001)
24. Kim, S.-H., Vel, P., Whitt, W., Cha, W.C.: Poisson and non-Poisson properties in appointment-generated arrival processes: the case of an endocrinology clinic. Oper. Res. Lett. **43**(3), 247–253 (2015)
25. Kim, S.-H., Whitt, W.: Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? Manuf. Serv. Oper. Manag. **16**(3), 464–480 (2014)
26. Kim, S.-H., Whitt, W., Cha, W.C.: A data-driven model of an appointment-generated arrival process at an outpatient clinic. INFORMS J. Comput. **30**(1), 181–199 (2018)
27. Kingman, J.F.C.: On queues in heavy traffic. J. R. Stat. Soc. B **24**(2), 383–392 (1962)
28. Koçaga, Y.L., Armony, M., Ward, A.R.: Staffing call centers with uncertain arrival rate and co-sourcing. Prod. Oper. Manag. **24**(7), 1101–1117 (2015)
29. Latouche, G., Ramaswami, V.: Introduction to Matrix Analytic Methods in Stochastic Modeling. Statistics & Applied Probability Series. SIAM, Philadelphia (1999)
30. Maman, S.: Uncertainty in the demand for service: the case of call centers and emergency departments. Master's thesis, Technion—Israel Institute of Technology (2009)
31. Mehrotra, V., Ozlük, O., Saltzmann, R.: Intelligent procedures for intra-day updating of call center agent schedules. Prod. Oper. Manag. **19**(3), 353–367 (2010)
32. Nagaev, S.V.: Large deviations of sums of independent random variables. Ann. Probab. **7**(5), 745–789 (1979)
33. Neuts, M.F.: Matrix-Geometric Solutions in Stochastic Models. The Johns Hopkins University Press, Baltimore (1981)
34. Robbins, T.R., Medeiros, D.J., Harrison, T.P.: Does the Erlang C model fit in real call centers? In: Proceedings of the 2010 Winter Simulation Conference (2010)
35. Ross, S.M.: Stochastic Processes. Wiley, Hoboken (1996)
36. Sigman, K., Whitt, W.: Heavy-traffic limits for nearly deterministic queues. J. Appl. Probab. **48**(3), 657–678 (2011)
37. Sigman, K., Whitt, W.: Heavy-traffic limits for nearly deterministic queues: stationary distributions. Queueing Syst. **69**, 145–173 (2011)
38. Steckley, S.G., Henderson, S.G., Mehrotra, V.: Forecast errors in service systems. Probab. Eng. Inf. Sci. **23**(2), 305–332 (2009)
39. Whitt, W.: Dynamic staffing in a telephone call center aiming to immediately answer all calls. Oper. Res. Lett. **24**(5), 205–212 (1999)
40. Whitt, W.: Staffing a call center with uncertain arrival rate and absenteeism. Prod. Oper. Manag. **15**(1), 88–102 (2006)
41. Zan, J.: Staffing service centers under arrival-rate uncertainty. PhD thesis, University of Texas (2012)