Submitted to *Operations Research*
manuscript (Please, provide the manuscript number!)

Robust heavy-traffic approximations for service systems facing overdispersed demand

B.W.J. Mathijsen, A.J.E.M. Janssen, J.S.H. van Leeuwaarden, A.P. Zwart
Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB, Eindhoven,
{b.w.j.mathijsen,a.j.e.m.janssen,j.s.h.v.leeuwaarden,a.p.zwart}@tue.nl

A. Mandelbaum
Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, 32000 Haifa, Israel,
avim@ie.technion.ac.il

Arrival processes to service systems are prevalently assumed non-homogeneous Poisson. Though mathematically convenient, arrival processes are often more volatile, a phenomenon that is referred to as overdispersion. Motivated by this, we analyze a class of stochastic models for which we develop performance approximations that are scalable in the system size, under a heavy traffic condition. 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. To illustrate the value of our approach, we apply it to actual arrival data of an emergency department of a hospital.

1. Introduction

In service systems, a central question is how to match capacity and demand. By taking into account the natural fluctuations of demand and capacity, stochastic models that describe congestion over time have proved instrumental in quantifying performance and discovering near-optimal capacity sizing rules. The bulk of the literature assumes perfect knowledge about the model primitives, including the mean demand per time period. For large-scale service systems, like health care systems, communication systems or call centers, the dominant assumption is that demand arrives according to a non-homogeneous Poisson process, which in practice translates into the assumption that arrival rates are known for each basic time period (second, hour or day).

Although natural and convenient from a mathematical viewpoint, the Poisson assumption often fails to be confirmed in practice. A deterministic arrival rate implies that the demand over any given period is a Poisson random variable, whose variance equals its expectation. A growing number of

empirical studies shows that the variance of demand typically deviates from the mean significantly. Recent work of Kim et al. (2015a,b) reports variance being strictly less than the mean in health care settings employing a appointment booking system. 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 typically dominates the mean, see Jongbloed and Koole (2001), Chen and Henderson (2001), Avramidis et al. (2004), Brown et al. (2005), Maman (2009), Bassamboo and Zeevi (2009), Steckley et al. (2009), Gurvich et al. (2010), Robbins et al. (2010), Bassamboo et al. (2010), Mehrotra et al. (2010), Gans et al. (2012), Zan (2012) and Kim and Whitt (2014). The feature that variability is higher than one expects from the Poisson assumption is referred to as *overdispersion*. The latter concept will be the center of our attention.

Stochastic models with the Poisson assumption have been widely applied to optimize capacity levels in service systems. The goal is to minimize operating costs while providing sufficiently high Quality-of-Service in terms of performance measures such as mean delay or excess delay. 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.

To deal with overdispersion new models are needed, scaling rules must be adapted, and existing capacity sizing rules need to be modified in order to incorporate a correct hedge against (increased) variability. Within the realm of Poisson processes, overdispersion can be modeled by viewing the arrival rate itself as being stochastic. The resulting doubly stochastic Poisson process, also known as Cox process (first presented in Cox (1955)), gives rise to demand in a given interval that follows a mixed Poisson distribution. In this paper, we consider a queueing model that has a doubly stochastic Poisson process as input, and we identify the heavy-traffic regime in which it displays Quality-and-Efficiency Driven (QED) behavior, first explored in the classical work of Halfin and Whitt (1981). By this, one roughly means that for systems with large demand and capacity, in heavy-traffic, the empty-queue probability is a controllable number strictly between 0 and 1, and that the mean delay is negligible. The key idea is to approximate the behavior of the stochastic model for a service system with that of a limiting process. The limiting process arises from a specific relationship between the arrival rate and the capacity level as both grow large without bound. Of the aforementioned papers, our work best relates to Maman (2009), in the sense that our model applies to situations where the arrival rate is stochastic. We therefore expand the paradigm of the QED regime, in order to have it accomodate for overdispersed demand that follows from a doubly stochastic Poisson process.

We divide time into periods of equal length, whereas demand in each period is then generated in two steps. First, the *rate* Λ of the Poisson variable is drawn from some distribution on $(0,\infty)$ and then a Poisson variable with that realized rate is generated. Hence, the arrival process is as if it is influenced by an external factor Λ. The mean demand is then given by $\mathbb{E}\Lambda$, while the variance of the demand is $\operatorname{Var}\Lambda + \mathbb{E}\Lambda$. By selecting the distribution of the mixing factor Λ, the variance can be made arbitrarily large, and only a deterministic Λ leads to a true Poisson process. The uncertain arrival rate can also be viewed as forecast errors. Indeed, in many cases, the uncertainty in forecasting the arrival rate can be large relative to the fluctuations naturally expected in Poisson processes, and should then be taken into account.

Without parameter uncertainty, a popular rule to choose the number of servers s in a service system, if the mean service time equals one, is $s = \mathbb{E}\Lambda + \beta\sqrt{\mathbb{E}\Lambda}$, for some tuning parameter $\beta > 0$. This is the well-known square-root safety staffing rule, that underpins much of the literature on systems in the QED regime (cf. Borst et al. (2004) and references therein). Whitt (1999) is among the first to call for Poisson models with an uncertain arrival rate in view of forecast errors, and to choose capacity levels accordingly. Using infinite server approximations, Whitt (1999) illustrates that this choice of s is naive, and that parameter uncertainty related to overdispersion can be accounted for by considering $s = \mathbb{E}\Lambda + \beta\sqrt{\operatorname{Var}\Lambda + \mathbb{E}\Lambda}$. In a similar vein, Maman (2009) considers the $M/M/s+G$ system, assuming the arrival rate to be mixed Poisson with a Gamma distributed Λ, which leads to an arrival rate with mean λ and a standard deviation of the form λ^c, where $0 < c \leq 1$. The natural capacity prescription would then become $s = \lambda + O(\lambda^c)$; without overdispersion, $c = 1/2$. We provide additional support for these staffing rules, and show how they can be updated to account for the fact that convergence towards heavy traffic can be slow. To illustrate our findings in more detail, we now provide a description of our model.

1.1. A discrete stochastic model with overdispersed input

In operations management, overdispersion has predominantly been modeled in a Poissonian setting, in which the arrival process is lifted from a Poisson process to a doubly stochastic Poisson process, in order to explain the overdispersion observed in datasets in terms of arrival rate uncertainty. Although the viewpoint of rate uncertainty is sensible and leads to a better fit with real data, the actual process that drives demand is most likely neither Poisson nor mixed Poisson. The papers on capacity sizing in service systems facing overdispersion, e.g. Maman (2009), Koçaga et al. (2014), Whitt (1999) and Whitt (2006), depart from the premise that the arrival process is of a Poissonian nature, and built on stochastic models for individual customer arrivals and departures under Markovian assumptions, that is, queues of the $M/M/s$ and $M/M/s+M$ type.

The classical birth-death processes that describe these systems are the main drivers, both for performance evaluation, and as input for cost minimization problems.

We propose a basic queueing model that differs from the setting of these birth-death models, but nonetheless captures the queueing facets of aforementioned service systems, while accounting for the overdispersed arrival stream in a natural way. We divide time into periods of equal length, and model the net input in period k as the difference between the incoming demand $A_n(k)$ and the capacity s_n, which is assumed to be fixed for all periods. Our model represents processes embedded at equidistant time points, driven by *arrival counts* in the periods between these time points. Since the random variable $A_n(k)$ leaves room for interpretations that do not rely on the Poisson assumption, our model has a wide scope of applications. Let us mention some possible interpretations:

(i) *Many-sources paradigm.* The canonical framework for large data-handling systems considers a buffer that receives messages from n i.i.d. information sources. Source i generates $X_i(k)$ data packets in slot k, so that in total $A_n(k) = \sum_{i=1}^{n} X_i(k)$ packets join the buffer in slot k. The buffer depletes through an output channel with a maximum transmission capacity of s_n packets per time slot. As such our model can be viewed as a discrete version of the Anick-Mitra-Sondhi model, see Anick et al. (1982), with the additional feature that sources can be correlated, which then leads to an overdispersed arrival process of packets.

(ii) *Data fitting.* 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 non-parametrically, as proposed in Jongbloed and Koole (2001); see also Maman (2009). A popular parametric family is the Gamma distribution, which gives rise to an effective data fitting procedure that makes use of the fact that a Gamma mixed Poisson random variable follows a negative binomial distribution. Hence, when process data for service systems is available, and in particular gives information on mean and variance of demand, the fitted mixed Poisson model can be fed into our stochastic model in order to evaluate the system's performance.

(iii) *Factor models.* Although having received little attention in queueing theory, factor models have a long history in the modeling of overdispersion in a wide variety of applications (see e.g. Johnson et al. (1993), Section 8.3.2). The mixed Poisson model is one of the base models in the rapidly expanding area of Credit Risk (e.g. Glasserman (2003)), and models the portfolio risk by imposing positive correlation (overdispersion) among individuals loans in the form of a common factor. The economical reasoning behind this can be transferred directly to service systems: There is a common factor Λ that influences the behavior of all customers. Conditional on the realization of Λ, customers all individually generate demand for the service systems. The mixed Poisson model is just one of many possible mixture models that are fit for describing correlation and overdispersion.

Since we have not made any further assumptions on $A_n(k)$, we can feed into our system all sorts of mixture models, like the mixed binomial model, or models with multiple factors.

(iv) *Panel sizing.* A matter of acute societal interest is accessing medical care in a timely manner. The average primary care physician's panel size is often too large for delivering consistently high quality care under the traditional practice model. Case studies suggest that a primary care physician providing all recommended acute, chronic, and preventive care for a panel of 2,500 patients receives about 24 appointment requests per day, whereas the physician is only able to serve 12 patients each day, thus creating huge backlogs (see Murray et al. (2003) and Green et al. (2007)). There thus seems to be a mismatch between workload and primary care physicians' capacity to deliver consistently high quality care. Zacharias and Armony (2014) model this panel sizing problem in terms of a queueing model for the appointment book of a clinic, with a panel of n patients and which can schedule a maximum of s_n patients per day. The realized schedule depends on the appointment queue at the beginning of the working day. This new demand for day k, added to the appointment queue, consists of new requests for appointments coming from the panel of n patients, and can be captured by a random variable $A_n(k)$ which is likely to be overdispersed.

(v) *Open access.* Our model is also suitable for describing service systems with open access scheduling, which means that the system serves customers on a first-come-first-serve basis without using an appointment book (Murray and Tantau (1999)). An example of this setting is given by Izady (2015), who considers appointment capacity planning in specialty clinics. Indeed, particularly in health care settings, open access gains popularity, because it holds the promise to strike the proper balance between utilization and quality of service. Moreover, in some health care settings like an emergency department, it is reasonable to assume that patients arrive without appointment and should be treated on demand. In this paper, we apply our model to an open access setting in a hospital in the far east, which we shall refer to as SKHospital. Here emergency patients require diagnostic tests at the radiology department of the hospital. We shall demonstrate that our model fits the data and that the capacity sizing rules that follow from this model can lead to significant performance improvement.

1.2. A non-standard saddle point method

The other advantage of our model is its tractability. It is amenable to powerful mathematical methods from complex and asymptotic analysis. For our heavy-traffic limits, we take an original approach that starts from Pollaczek's formula, which represents the transform of the stationary queue length distributions in terms of a contour integral. From this classical transform representation, contour integrals for the zero-queue probability and the mean queue length follow immediately. Contour integrals are often suitable for asymptotic evaluation (see e.g. Cohen (1982)),

particularly for obtaining classical heavy traffic asymptotics. While we also subject the contour integral representations to asymptotic evaluation, ours is not the classical heavy-traffic scaling. This asymptotic analysis requires a *non-standard* saddle point method, tailored to the specific form of the integral expressions that arise under overdispersed arrivals and QED-type capacity sizing rules. This leads to asymptotic expansions for performance measures, of which the limiting forms correspond to heavy-traffic limits, and pre-limit forms present refined approximations for pre-limit systems ($n < \infty$) in heavy traffic. Such refinements to heavy-traffic limits are commonly referred to as *corrected diffusion approximations*; see Siegmund (1978), Blanchet and Glynn (2006), Asmussen (2003).

Let us briefly explain why our saddle point method is *non-standard*. 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 non-standard saddle point method, originally proposed by de Bruijn (1981) and recently applied in Janssen et al. (2015), aims specifically to overcome this challenge. In Section 5 we elaborate on the technicalities of this method and explain why relying directly on the path of analysis of Janssen et al. (2015) is insufficient in the presence of overdispersion.

1.3. Contributions

The first set of results in this paper cover the mixed Poisson demand with general mixing factor Λ. We prove that our stochastic model with Pois(Λ) demand and capacity set according to the QED-type regime, converges to the Gaussian random walk. More specifically, denote the mean and variance of the demand $A_n(k)$ by μ_n and σ_n^2, and assume that the system load $\rho_n = \mu_n/s_n$ approaches one such that $(1-\rho_n)\frac{\mu_n}{\sigma_n} \to \gamma$, as $n \to \infty$. Under additional assumptions on the growth rates of μ_n and σ_n, the stationary queue length, normalized by σ_n, converges to the all-time maximum M_γ of a random walk with i.i.d. normal increments, having mean $-\gamma$ and unit variance. This Gaussian random walk is a sampled version of the Brownian motion, the properties of which are well understood; see Chang and Peres (1997) and Janssen and van Leeuwaarden (2006). Carrying over known results on M_γ yields heavy-traffic approximations for stationary performance measures. As a result, for large-scale critical service systems, using the QED-type rule $s_n = \mu_n + \gamma\sigma_n$ for matching capacity with demand, systems facing overdispersion can be dimensioned in such a way that the delay probability is strictly between 0 and 1, and has mean delays that are asymptotically negligible.

The other question, also of operational significance but at a more refined level, concerns improving and understanding overdispersion by assessing its consequences. We focus on the impact on stationary performance measures. Our heavy-traffic analysis and numerical examples show that overdispersion can have tremendous adverse effects, both in terms of performance measures and selection of capacity, the latter being a function of demand that provides a predetermined grade of service. Moreover, we show that overdispersion causes the system to converge more slowly to its limiting behavior. This slow convergence begs refinements to make our performance results more broadly applicable. As it happens, our analytic approach is well-suited for the challenge.

This leads to our second set of results, for the more specific (yet practically relevant) case in which Λ_n has a Gamma distribution with parameters a_n and $1/b_n$. While our first set of results yields the conventional heavy-traffic approximation $Q_n \approx \sigma_n M_\gamma$, for the invariant queue length Q_n, one of our robust refinements implies that the parameters γ and σ_n in this approximation should be replaced by

$$\gamma_n = \gamma \sqrt{1 - \frac{1}{\sigma_n^2/\mu_n + \sigma_n/\gamma}} \qquad \text{and} \qquad \tilde{\sigma}_n = \frac{\gamma_n}{\gamma}\sigma_n + \gamma_n\Big(\frac{\sigma_n^2}{\mu_n} - 1\Big). \tag{1.1}$$

Close inspection of the functions γ_n and $\tilde{\sigma}_n$ show that $\gamma_n \to \gamma$ and $\tilde{\sigma}_n/\sigma_n \to 1$, for large n; it follows that for large service systems, the difference between the classical and robust approximation should be negligible. More importantly, for small and moderate n, the difference between γ_n and $\tilde{\sigma}_n$ and their original counterparts is considerable, and the robust approximations are decisively more accurate, particularly in situations of overdispersion.

1.4. Connection with literature on staffing

While all results reveal a clear impact of overdispersion on system performance, implications on staffing are less pronounced. Whitt (2006) studies the $M/GI/s+GI$ queue via an approximating fluid model to show that overdispersion leads to severe performance degradation but, at the same time, the effect on staffing/capacity sizing is less significant. Whitt attributes this to the fact that the objective function over which performance is optimized, is relatively flat as a function of the servers. Koçaga et al. (2014) consider the $M/M/s+M$ queue with uncertain arrival rate and an outsourcing option, and determine the asymptotically optimal policy as the solution to a cost-minimization problem. Despite the presence of overdispersion, the results and numerical finding of Koçaga et al. (2014) are similar to Borst et al. (2004): square-root capacity sizing is near optimal and robust against many circumstances.

One way of understanding Koçaga et al. (2014) is that they consider a situation in which the uncertainty of the arrival rate is of the exact same order as the natural system uncertainty. The

present paper deals predominantly with higher levels of overdispersion that renders the square-root rule invalid, rather than refining it as explained above. This is consistent with the arguments of Ding and Koole (2014) who show that factor models used to forecast systems loads exhibit significant overdispersion; such models are used in practice when service levels need to be scheduled days or weeks in advance. In addition, Bassamboo et al. (2010) propose a capacity sizing rule for the $M/M/s+M$ queue with uncertain arrival rate. Exploiting a newsvendor approximation, the optimal capacity sizing rule is shown to consist of a base capacity μ_n and an additional capacity that is proportional to σ_n. When $\sigma_n^2/\mu_n \to \infty$, overdispersion is dominant, a situation which is called in Bassamboo et al. (2010) the uncertainty-dominated regime.

Our work shows that in cases of mild yet dominating forms of overdispersion, capacity sizing rules, based on hedging the natural fluctuations of the demand, lead to behavior that is favorable over conventional staffing rules, if slow convergence properties that play a minor role in conventional systems are taken into account appropriately. We expect that this slow convergence requires refinements, not only at the level of performance measures as in this paper, but also in cost minimization models, though we do not pursue this here.

1.5. Organisation

The remainder of this paper is structured as follows. Our model is introduced in Section 2. In Section 3 we present our main theoretical results, including classical and robust heavy-traffic approximations for the stationary queue length. In Section 4, we describe the numerical results and demonstrate the heavy-traffic approximation for a real data set coming from a SKHospital. Section 5 contains the proof of the results formulated in Subsection 3.1, as where Section 6 contains the technical details of those in Subsection 3.2.

2. Model description and preliminaries

We consider a discrete stochastic model in which time is divided into periods of equal length. At the beginning of each period $k=1,2,3,...$ new demand $A_n(k)$ arrives to the system. The demands per period $A_n(1), A_n(2), ...$ are assumed independent and equal in distribution to some non-negative integer-valued random variable A_n. The system has a service capacity $s_n \in \mathbb{N}$ per period, so that the recursion

$$Q_n(k+1) = \max\{Q_n(k) + A_n(k) - s_n, 0\}, \qquad k=0,1,2,..., \tag{2.1}$$

with $Q_n(0)=0$. For brevity, we define $\mu_n := \mathbb{E}A_n$ and $\sigma_n^2 = \operatorname{Var} A_n$. The duality principle shows that this expression is equivalent to

$$Q_n(k+1) \stackrel{d}{=} \max_{0\le j\le k}\Big\{\textstyle\sum_{i=1}^{j}(A_n(i)-s_n)\Big\}, \qquad k=0,1,2,..., \tag{2.2}$$

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

$$Q_n \stackrel{d}{=} \max_{k\geq 0}\Big\{\textstyle\sum_{i=1}^{k}(A_n(i)-s_n)\Big\}. \tag{2.3}$$

The behavior of $Q_n(k)$ 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. With this constraint in mind, we set $s_n = \mu_n + \gamma\sigma_n$, with $\gamma > 0$, for which we provided intuition in Section 1.

We further impose a heavy-traffic condition, $\rho_n = \mu_n/s_n \to 1$, which for our choice of s_n is equivalent to requiring

$$(1-\rho_n)\frac{\mu_n}{\sigma_n} \to \gamma, \qquad \text{as } n\to\infty. \tag{2.4}$$

Another condition we impose is that

$$\frac{\sigma_n^2}{\mu_n} \to \infty, \qquad n\to\infty, \tag{2.5}$$

which roughly says that the overdispersed nature of the arrival process is persistent when $n\to\infty$. Since we are mainly interested in the system in heavy traffic it is appropriate to look at the queue length process in a scaled form. Filling in s_n as well as dividing both sides of (2.3) by σ_n, gives

$$\frac{Q_n}{\sigma_n} = \max_{k\geq 0}\Big\{\textstyle\sum_{i=1}^{k}\Big(\dfrac{A_n(i)-\mu_n}{\sigma_n} - \gamma\Big)\Big\}. \tag{2.6}$$

By defining $\hat{Q}_n := Q_n/\sigma_n$ and $\hat{A}_n(i) := (A_n(i)-\mu_n)/\sigma_n$, 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 $\hat{A}_n - \gamma$. Besides $\mathbb{E}\hat{A}_n = 0$ and $\operatorname{Var}\hat{A}_n = 1$, the scaled and centered arrival counts $\hat{A}_n$ has a few other nice properties which we turn to later in this section.

The model in (2.1) 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 (2.5) does not hold then. We will deviate too much from this setting. Instead, we assume A_n to be Poisson distributed with uncertain arrival rate rendered by the non-negative 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 Grandell (1997). The probability generating function (pgf) of A_n can be written in terms of the moment generating function (mgf) of Λ_n, namely,

$$A_n(z) = \mathbb{E}[\mathbb{E}[z^{A_n}|\Lambda_n]] = \mathbb{E}[\exp(\Lambda_n(z-1))] = M_n(z-1), \tag{2.7}$$

where $M_n(t)$ is the mgf of Λ. From (2.7), we get

$$\mu_n = \mathbb{E}A_n = \mathbb{E}\Lambda_n, \qquad \sigma_n^2 = \operatorname{Var}A_n = \operatorname{Var}\Lambda_n + \mathbb{E}\Lambda_n, \tag{2.8}$$

so that $\mu_n < \sigma_n^2$ if Λ_n is non-deterministic. The condition in (2.5) hence translates to $\operatorname{Var}\Lambda_n/\mathbb{E}\Lambda_n \to \infty$ for $n \to \infty$. The next result relates the converging 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*

$$\hat{\Lambda}_n := \frac{\Lambda_n - \mu_n}{\sigma_n} \overset{d}{\Rightarrow} \mathcal{N}(0,1), \qquad \text{for } n \to \infty, \tag{2.9}$$

then $\hat{A}_n$ converges weakly to a standard normal variable as $n \to \infty$.

The proof can be found in Section 7.

The prevalent choice for Λ_n is the Gamma distribution. The Gamma-Poisson mixture turns out to provide a very good fit to arrival counts experienced by service systems, as was observed by Jongbloed and Koole (2001). Assuming Λ_n to be of Gamma type with scale and rate parameters a_n and $1/b_n$, respectively, we get

$$A_n(z) = \Big(\frac{1}{1 + b_n(1-z)}\Big)^{a_n}, \tag{2.10}$$

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

$$\mu_n = a_n b_n, \qquad \sigma_n^2 = a_n b_n (b_n + 1). \tag{2.11}$$

Hence, requiring $b_n \to \infty$ as $n \to \infty$, gives the desired persistent overdispersion. An important implication of Λ_n being a Gamma random variable is the following.

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 $\hat{\Lambda}_n \Rightarrow \mathcal{N}(0,1)$ for this particular choice of Λ_n. We do this by proving the pointwise convergence of the 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,

$$\begin{aligned}
\varphi_{\hat{\Lambda}_n}(t) &= \mathrm{e}^{-i\mu_n t/\sigma_n}\, \varphi_{\Lambda_n}(t/\sigma_n) = \mathrm{e}^{-i\mu_n t/\sigma_n} \Big(1 - \frac{ib_n t}{\sigma_n}\Big)^{-a_n} \\
&= \exp\Big[-\frac{i\mu_n t}{\sigma_n} - a_n \ln\Big(1 - \frac{ib_n t}{\sigma_n}\Big)\Big] \\
&= \exp\Big[-\frac{i\mu_n t}{\sigma_n} - a_n\Big(-\frac{i\, b_n t}{\sigma_n} + \frac{b_n^2 t^2}{2\sigma_n^2} + O(b_n^3/\sigma_n^3)\Big)\Big] \\
&= \exp\Big[-\frac{b_n\, t^2}{2(b_n+1)} + O\left(1/\sqrt{a_n}\right)\Big] \to \exp\left(-t^2/2\right),
\end{aligned} \tag{2.12}$$

for $n \to \infty$. By Lévy's continuity theorem this implies $\hat{\Lambda}_n$ is indeed asymptotically standard normal. □

The characterization of the arrival process as a Gamma-Poisson mixture be 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 $P(Q_n = i) = \lim_{k\to\infty} \mathbb{P}(Q_n(k) = i)$. Denote the pgf of Q_n by

$$\tilde{Q}_n(w) = \sum_{i=0}^{\infty} P(Q_n = i) w^i. \tag{2.13}$$

We next recall two characterizations of $\tilde{Q}_n(w)$ that play prominent roles in the remainder of our analysis. Throughout we assume that the pgf of A_n, denoted by $A_n(w)$, exists within $|z| < r_0$, for some $r_0 > 1$, so that all moments of A_n are finite.

The first characterization of $\tilde{Q}_n(w)$ originates from a random walk perspective. As we see from (2.3), 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 $A_n - \gamma$ (or $\hat{A}_n - \gamma$ in the scaled setting). Spitzer's identity, see e.g. (Asmussen 2003, Theorem VIII4.2), then gives

$$\tilde{Q}_n(w) = \exp\Big\{\sum_{k=1}^{\infty} \frac{1}{k}(\mathbb{E}[w^{\left(\sum_{i=1}^{k}\{A_n(i)-s_n\}\right)^+}] - 1)\Big\}, \tag{2.14}$$

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

$$\mathbb{E}Q_n = \tilde{Q}_n'(1) = \sum_{k=1}^{\infty} \frac{1}{k} \mathbb{E}\Big[\sum_{i=1}^{k}(A_n(i) - s_n)\Big]^+, \tag{2.15}$$

$$\operatorname{Var} Q_n = \tilde{Q}_n''(1) + Q_n'(1) - \Big(\tilde{Q}_n'(1)\Big)^2 = \sum_{k=1}^{\infty} \frac{1}{k} \mathbb{E}\Big[\left(\sum_{i=1}^{k}(A_n(i) - s_n)\right)^+\Big]^2, \tag{2.16}$$

$$P(Q_n = 0) = \tilde{Q}_n(0) = \exp\Big\{-\sum_{k=1}^{\infty} \frac{1}{k} P\Big(\textstyle\sum_{i=1}^{k}(A_n(i) - s_n) > 0\Big)\Big\}. \tag{2.17}$$

A second characterization follows from Pollaczek's formula, see Abate et al. (1993), Janssen et al. (2015):

$$\tilde{Q}_n(w) = \exp\Big\{\frac{1}{2\pi i}\int_{|z|=1+\varepsilon} \ln\Big(\frac{w-z}{1-z}\Big) \frac{(z^{s_n} - A_n(z))'}{z^{s_n} - A_n(z)}\, dz\Big\}, \tag{2.18}$$

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

$$\mathbb{E}Q_n = \frac{1}{2\pi i}\int_{|z|=1+\varepsilon} \frac{1}{1-z} \frac{(z^{s_n} - A_n(z))'}{z^{s_n} - A_n(z)}\, dz, \tag{2.19}$$

$$\operatorname{Var} Q_n = \frac{1}{2\pi i}\int_{|z|=1+\varepsilon} \frac{-z}{(1-z)^2} \frac{(z^{s_n} - A_n(z))'}{z^{s_n} - A_n(z)}\, dz, \tag{2.20}$$

$$P(Q_n = 0) = \exp\Big\{\frac{1}{2\pi i}\int_{|z|=1+\varepsilon} \ln\Big(\frac{z}{z-1}\Big) \frac{(z^{s_n} - A_n(z))'}{z^{s_n} - A_n(z)}\, dz\Big\}. \tag{2.21}$$

These two sets of expressions for the characteristics of the queue reappear several times in the next sections.

3. Main results on robust approximations

3.1. Process-level convergence and stationary moments

Observe that (2.1) is in fact Lindley's recursion for the waiting time in a $D/G/1$ system. Bearing in mind the many-sources interpretation, for large n and s_n, this recursion starts resembling that of a $D/D/1$ system, which suggests that our system becomes *nearly deterministic*, and only due to the traffic intensity increasing as in (2.4) the system displays interesting limiting behavior. A more generic class of nearly deterministic queueing systems was introduced in Sigman and Whitt (2011a,b), in terms of the $G_n/G_n/1$ system, where G_n indicates *cyclic thinning* of order n, indicating that some point process is thinned to contain only every nth point. As $n\to\infty$, the $G_n/G_n/1$ systems thus approaches the deterministic $D/D/1$ system. For $G_n/G_n/1$ systems, Sigman and Whitt (2011a) establishes stochastic-process limits, and Sigman and Whitt (2011b) derives heavy-traffic limits for stationary waiting times. In the framework of Sigman and Whitt (2011a,b), the recursion (2.1) corresponds to a $D/G_n/1$ queue, where the sequence of service times $(A_n(k))_{k\geq 1}$ follows from a cyclically thinned sequence of i.i.d. random variables. For the $G_n/D/1$ queue, which describes the waiting-time process in a $G/D/n$ queue, a similar result was obtained in Jelenkovic et al. (2004). The main results in Jelenkovic et al. (2004), Sigman and Whitt (2011a,b) were obtained under the assumption that $\rho_n\sim 1-\gamma/\sqrt{n}$, in which case it follows from (Sigman and Whitt 2011b, Theorem 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 $\rho_n\sim 1-\frac{\gamma}{\mu_n/\sigma_n}$. Hence, in contrast to Jelenkovic et al. (2004) and Sigman and Whitt (2011a,b), our systems' characteristics display larger natural fluctuations, due to the mixing factor that renders the arrivals. Yet, by matching this overdispersed demand with the appropriate hedge against variability, we again obtain Gaussian limiting behavior. Note that 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 Λ_n be a non-negative random variable such that $(\Lambda_n-\mu_n)/\sigma_n$ is asymptotically standard normal with μ_n and σ_n^2 as defined in* (2.8)*. Assume both $\mu_n,\sigma_n^2\to\infty$ and $\sigma_n^2/\mu_n\to\infty$ as $n\to\infty$. Then, for $n\to\infty$,*

(i) $\hat{Q}_n\overset{d}{\Rightarrow} M_\gamma$,

(ii) $\mathbb{P}(Q_n=0)\to\mathbb{P}(M_\gamma=0)$,

(iii) $\mathbb{E}[\hat{Q}_n]\to\mathbb{E}M_\gamma$,

(iv) $\text{Var}\ Q_n\to\text{Var}\ M_\gamma$

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

The proof of Theorem 1 is given in Section 7. The following result shows that Theorem 1 also applies to Gamma mixtures, which is a direct consequence of Corollary 1.

COROLLARY 2. *Let Λ_n be Gamma distributed with scale and rate parameters a_n and $1/b_n$, respectively, such that $a_n, b_n \to \infty$. Then the four convergence results of Theorem 1 hold true.*

It follows from Theorem 1 that the scaled stationary queueing process converges under (2.4) 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, giving rise to 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 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. Abate et al. (1993) have considered the numerical evaluation of such integrals. For $\mathbb{E}M_\gamma$ such an integral is as follows

$$\mathbb{E}M_\gamma = -\frac{1}{\pi}\int_0^\infty \mathrm{Re}\Big[\frac{1-\varphi(-z)}{z^2}\Big]\,dy, \tag{3.1}$$

with $\varphi(z) = \exp(-\gamma\, z + \frac{1}{2}\, z^2)$, the Laplace transform of a normal random variable with mean $-\gamma$ and unit variance, and $z = x + iy$ with an appropriately chosen real part x. Note that this integral involves complex-valued numbers. Similar expressions appear for $\mathbb{P}(M_\gamma = 0)$ and $\mathrm{Var}\, M_\gamma$. The following result simply rewrites these integrals in (3.1) in terms of a real integral (the derivation is given in the e-companion) and uses the fact that the scaled queue length process mimics the maximum of the Gaussian random walk for large n.

COROLLARY 3. *Let $\mu_n, \sigma_n \to \infty$ and $\sigma_n^2/\mu_n \to \infty$. Then the leading order behavior of $\mathbb{P}(Q_n = 0)$, $\mathbb{E}Q_n$ and $\mathrm{Var}\, Q_n$ is characterized by*

$$\mathbb{P}(Q_n = 0) \approx \exp\Big[\frac{1}{\pi}\int_0^\infty \frac{\gamma/\sqrt{2}}{\frac{1}{2}\gamma^2 + t^2}\ln\Big(1 - e^{-\frac{1}{2}\gamma^2 - t^2}\Big)\,dt\Big], \tag{3.2}$$

$$\mathbb{E}Q_n \approx \frac{\sqrt{2}\sigma_n}{\pi}\int_0^\infty \frac{t^2}{\frac{1}{2}\gamma^2 + t^2}\,\frac{\exp(-\frac{1}{2}\gamma^2 - t^2)}{1 - \exp(-\frac{1}{2}\gamma^2 - t^2)}\,dt, \tag{3.3}$$

$$\mathrm{Var}\, Q_n \approx \frac{\sqrt{2}\gamma\sigma_n^2}{\pi}\int_0^\infty \frac{t^2}{(\frac{1}{2}\gamma^2 + t^2)^2}\,\frac{\exp(-\frac{1}{2}\gamma^2 - t^2)}{1 - \exp(-\frac{1}{2}\gamma^2 - t^2)}\,dt. \tag{3.4}$$

3.2. Robust heavy-traffic approximations

To obtain more accurate approximations for $\mathbb{E}Q_n$, $\operatorname{Var} Q_n$ and $\mathbb{P}(Q_n = 0)$, using the Pollaczek's formula given in (2.18), we need to be more specific about the arrival process A_n and its pgf $A_n(w)$. In the remainder of this paper we work with the Gamma-Poisson mixture with parameters a_n and b_n, so that

$$A_n(w) = \Big(\frac{1}{1+(1-z)b_n}\Big)^{a_n}. \tag{3.5}$$

As mentioned earlier, Gamma mixing yields the negative binomial distribution, with pgf as in (3.5), which allows us to establish the detailed asymptotic results in the next theorem.

THEOREM 2. *Let a_n, b_n and s_n be such that*

$$(1-\rho_n)\sqrt{a_n} \to \gamma \tag{3.6}$$

for some $\gamma > 0$, as $n \to \infty$. Then the leading order behavior of $\mathbb{E}Q_n$ is given by

$$\mathbb{E}Q_n = \frac{\sqrt{2}\,\gamma_n}{\pi}\Big(\frac{b_n+\rho_n}{1-\rho_n}\Big)\int_0^\infty \frac{t^2}{\frac{1}{2}\gamma_n^2+t^2}\,\frac{\exp(-\frac{1}{2}\gamma_n^2-t^2)}{1-\exp(-\frac{1}{2}\gamma_n^2-t^2)}dt\,(1+o(1)), \tag{3.7}$$

where

$$\gamma_n^2 = s_n\Big(\frac{1-\rho_n}{b_n+1}\Big)^2\Big(1+\frac{b_n}{\rho_n}\Big). \tag{3.8}$$

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

$$\exp\Big[\frac{1}{\pi}\,\frac{b_n+\rho_n}{b_n+1}\int_0^\infty \frac{\gamma_n/\sqrt{2}}{\frac{1}{2}\gamma_n^2+t^2}\ln\Big(1-\mathrm{e}^{-\frac{1}{2}\gamma_n^2-t^2}\Big)\,dt\Big], \tag{3.9}$$

and

$$\frac{\gamma_n^3/\sqrt{2}}{\pi}\Big(\frac{b_n+\rho_n}{1-\rho_n}\Big)^2\Big(\frac{b_n+1}{b_n+\rho_n}+1\Big)\int_0^\infty \frac{t^2}{(\frac{1}{2}\gamma_n+t^2)^2}\,\frac{\exp(-\frac{1}{2}\gamma_n-t^2)}{1-\exp(-\frac{1}{2}\gamma_n^2-t^2)}\,dt, \tag{3.10}$$

respectively.

Note that we can write (3.7) as

$$\mathbb{E}Q_n = \tilde{\sigma}_n\,\mathbb{E}M_{\gamma_n} \quad \text{and} \quad \operatorname{Var} Q_n \approx \tilde{\sigma}_n^2 \operatorname{Var} M_{\gamma_n} \tag{3.11}$$

with

$$\tilde{\sigma}_n = \gamma_n\Big(\frac{b_n+\rho_n}{1-\rho_n}\Big). \tag{3.12}$$

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 (2.7). The approximation (3.7) suggests that it is better to normalize with $\tilde{\sigma}_n$. Although $\tilde{\sigma}_n \to \sigma_n$ 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 γ by the robust hedge γ_n. This thus means that the original system for finite n is approximated by a Gaussian random walk with drift $-\gamma_n$. Apart form this approximation being asymptotically correct for $n \to \infty$, it is also expected to approximate the behavior better for finite n.

4. Numerical and empirical studies

4.1. Convergence of the robust hedge

We next examine the accuracy of the heavy-traffic approximations for $\mathbb{E}Q_n$ and $\operatorname{Var} Q_n$, which follow from Corollary 3 and Theorem 2. We expect the robust approximation to be considerably better than the classical approximation when γ_n and $\tilde{\sigma}_n$ differ substantially from their limiting counterparts. To further substantiate the convergence of γ_n to γ and $\tilde{\sigma}_n$ to σ_n we present the results below.

PROPOSITION 1. *For $b_n, s_n \to \infty$ and $b_n \leq s_n$,*

$$\gamma_n^2 = \gamma^2 \Big(1 - \frac{1}{1 + b_n + \sigma_n/\gamma}\Big). \tag{4.1}$$

Proof From (3.8), we have

$$\begin{aligned} \gamma_n^2 &= s_n \Big(\frac{1-\rho_n}{b_n+1}\Big)^2 \Big(1 + \frac{b_n}{\rho_n}\Big) = \frac{1}{s_n}\Big(\frac{s_n - a_n b_n}{b_n+1}\Big)^2 \Big(1 + \frac{s_n}{a_n}\Big) \\ &= \frac{1}{s_n} \frac{\gamma^2 a_n b_n (b_n+1)}{(b_n+1)^2} \Big(1 + \frac{s_n}{a_n}\Big) = \gamma^2 \frac{b_n}{b_n+1} \Big(1 + \frac{a_n}{s_n}\Big) =: \gamma^2 \bar{F}_n. \end{aligned} \tag{4.2}$$

Now consider the factor $\bar{F}_n$.

$$\begin{aligned} \bar{F}_n &= \frac{b_n}{b_n+1}\Big(1 + \frac{a_n}{s_n}\Big) = \frac{b_n}{b_n+1} + \frac{1}{b_n+1}\frac{a_n b_n}{s_n} \\ &= 1 - \frac{1}{b_n+1}\Big(1 - \frac{a_n b_n}{s_n}\Big) = 1 - \frac{1}{b_n+1}\frac{\gamma \sigma_n}{s_n} \\ &= 1 - \frac{1}{b_n+1}\frac{1}{1 + \frac{\mu_n}{\gamma \sigma_n}} = 1 - \frac{1}{b_n + 1 + \frac{1}{\gamma}\sqrt{a_n b_n (b_n+1)}}, \end{aligned} \tag{4.3}$$

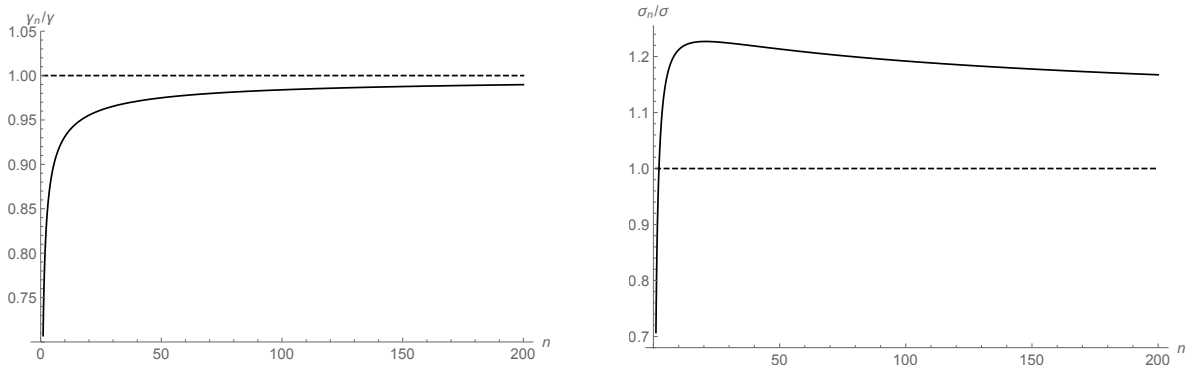
which together with $\sigma_n^2 = a_n b_n (b_n+1)$ proves the proposition.

□

Note that γ_n always approaches γ from below. Also, (4.1) shows that b_n is the dominant factor in determining the rate of convergence of γ_n.

PROPOSITION 2. *Let $\tilde{\sigma}_n$ as in* (3.12)*. Then*

$$\tilde{\sigma}_n = \sigma_n \left(1 + O(1/\sqrt{a_n b_n})\right) + b_n \gamma_n. \tag{4.4}$$



(a) Convergence of γ_n.

(b) Convergence of $\tilde{\sigma}_n$.

Figure 1

Proof Straightforward calculations give

$$\tilde{\sigma}_n = \gamma_n \left(\frac{s_n b_n + a_n b_n}{s_n - a_n b_n} \right) = \frac{\gamma_n}{\gamma} \frac{b_n}{\sigma_n} (s_n + a_n) = \frac{\gamma_n}{\gamma} \sqrt{\frac{b_n}{a_n(b_n+1)}} \left(a_n(b+1) + \gamma \sqrt{a_n b_n (b_n+1)} \right)$$
$$= \frac{\gamma_n}{\gamma} \left(\sqrt{a_n b_n (b_n+1)} + \gamma b_n \right) = \frac{\gamma_n}{\gamma} \sigma_n + \gamma_n b_n. \tag{4.5}$$

Applying Proposition 1 together with the observation

$$\sqrt{1 - \frac{1}{1 + b_n + \sigma_n/\gamma}} = 1 + O(1/\sqrt{a_n} b_n) \tag{4.6}$$

yields the result.

In Figure 1, we visualize the convergence speed of both parameters in case $\mu_n = n$, $\sigma_n = n^\delta$ with $\delta = 0.7$ and $\gamma = 1$. This implies $a_n = n/(n^{2\delta} - 1)$ and $b_n = n^{2\delta} - 1$.

We observe that γ_n starts resembling γ 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 $\operatorname{Var} Q_n$ are approximated by $\tilde{\sigma_n}$ 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.

4.2. Comparison between heavy-traffic approximations

We set, so that $\mu_n = n$ and $\sigma_n^2 = n^{2\delta}$ with $\delta > \frac{1}{2}$, so that $s_n = n + \gamma n^\delta$, and $a_n = n/(n^{2\delta-1} - 1)$ and $b_n = n^{2\delta-1} - 1$.

Tables 1 to 4 present numerical results for various parameter values. The exact values are calculated using the expression in Appendix A.

Several conclusions are drawn from these tables. First observe that the heavy-traffic approximations based on the Gaussian random walk, (3.3) and (3.4), capture the right order of magnitude

s_n	ρ_n	$\mathbb{E}Q_n$	(3.3)	(3.7)	$\sqrt{\mathrm{Var}\, Q_n}$	(3.4)	(3.10)
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

Table 1 Numerical results for the Gamma-Poisson case with $\gamma = 1$ and $\delta = 0.6$.

s_n	ρ_n	$\mathbb{E}Q_n$	(3.3)	(3.7)	$\sqrt{\mathrm{Var}\, Q_n}$	(3.4)	(3.10)
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	3.0248	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 2 Numerical results for the Gamma-Poisson case with $\gamma = 1$ and $\delta = 0.8$.

s_n	ρ_n	$\mathbb{E}Q_n$	(3.3)	(3.7)	$\sqrt{\mathrm{Var}\, Q_n}$	(3.4)	(3.10)
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 3 Numerical results for the Gamma-Poisson case with $\gamma = 0.1$ and $\delta = 0.6$.

s_n	ρ_n	$\mathbb{E}Q_n$	(3.3)	(3.7)	$\sqrt{\mathrm{Var}\, Q_n}$	(3.4)	(3.10)
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

Table 4 Numerical results for the Gamma-Poisson case with $\gamma = 0.1$ and $\delta = 0.8$.

for both $\mathbb{E}Q_n$ and $\sqrt{\mathrm{Var}\, Q_n}$. However, the values are off, in particular for small s_n and 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, (3.7) and (3.10) 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 $\sqrt{\mathrm{Var}\, Q_n}$, 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.

4.3. Capacity allocation in health care

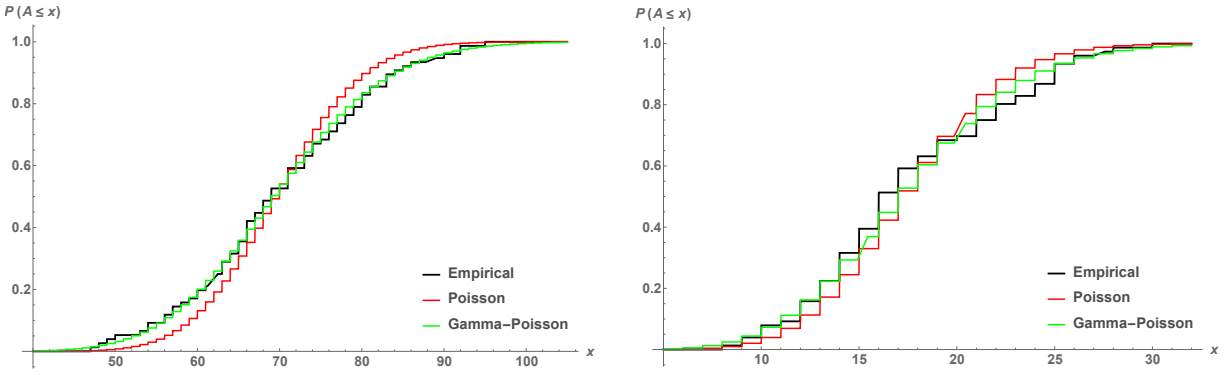
We next apply our model and robust approximations to real-life patient arrivals. We consider emergency patients who require diagnostic tests at the radiology department of a hospital. Green

(2004) points out that patients at the radiology department can be roughly categorized into three groups: Inpatients, outpatients and emergency patients. Inpatient and outpatient arrivals are relatively predictable as these are usually by appointments. Emergency patients, on the other hand, are inherently unpredictable: They typically require urgent care and therefore timely access to testing facilities, as well as a quick assessment of the test results. This translates into emergency patients getting priority over the other two groups in such settings, so that they do not experience any delay caused by the groups of lower priority. However, patients from the same top-priority group can still cause considerable congestion. A careful evaluation of capacity allocation is hence required, bearing in mind that additional sophisticated pieces of medical equipment are very costly.

In the setting we study, capacity is defined by the number of time slots available to perform radiology tests on emergency patients in a given time period, which we set at 24 hours. As radiology tests are commonly performed in appointment slots of fixed length, the number of slots available per day is also indirectly fixed. In terms of our model parameters, see Section 2, we have s as the number of slots per day allocated to emergency patients, and $A(k)$ the number of test requests received by the department on day k. We omit the subscript n in this section due to the absence of limits. Then $\mathbb{E}Q$ can be interpreted as the expected number of patients in stationarity whose test is carried over to the next day. A more natural performance measure in this setting is the expected waiting time, namely the time between the physician's request and the actual start of the test. However, Little's law implies that there is a linear relation between the two, hence we choose to only evaluate $\mathbb{E}Q$.

The data set on which our empirical study is based originates from the emergency department of SKHospital, monitored over a period of 76 days from September to November 2012. We extracted information of ED patients referred to the radiology department by the ED physicians, which includes the time the test request was made and the exact test type performed. The two test types, X-ray and CT scans, are performed on different equipment and hence it makes sense to analyze their queueing processes separately.

We refer to test requests as arrivals. The empirical cumulative distribution function of the number of arrivals per day, for each type, are depicted by the black lines in Figure 2 and 2. The sample means equal 69.81 and 17.47, for the X-ray and CT scans respectively, whereas the sample variances are 121.8 and 26.12. These values suggest that fitting a Poisson distribution is inappropriate, which is visually backed up by the fitted Poisson cdf, depicted in Figure 2 by the red line. To strengthen this claim, we tested both samples for the Poisson assumption using the *dispersion test*, see Appendix B, and obtained p-values equal $7.01 \cdot 10^{-3}$ and $3.57 \cdot 10^{-3}$ respectively, which allow us to safely reject the Poisson hypothesis in both cases.



(a) X-ray (b) CT scan

Figure 2 Empirical, fitted Poisson and fitted Gamma-Poisson cumulative distribution functions of the number of arrivals.

In search for a better distributional fit with the arrivals count, we resort to Gamma-Poisson mixtures. Here we employ the procedure in Jongbloed and Koole (2001), which is basically a maximum loglikelihood method, to obtain estimates for the parameters a and b. This yields

$$\hat{a}_{X-ray} = 95.68, \quad \hat{b}_{X-ray} = 0.7297, \quad \hat{a}_{CT} = 37.19, \quad \hat{b}_{CT} = 0.4698. \tag{4.7}$$

Applying the bootstrapping method to the data and the fitted model, also described in the appendix of Jongbloed and Koole (2001), returns p-values that equal 0.7354 and 0.2120 for X-ray and CT scans, respectively. Therefore, the null hypothesis, stating that the data originated from a Gamma-Poisson mixture, cannot be rejected. The cdfs of the fitted Gamma-Poisson distributions, plotted in green in Figure 2, give visual confirmation of this claim as well. Naturally, we also compared the estimated densities to the empirical pdf of the data. However, these fail to give a convincing visual fit due to the relatively small sample size and are therefore omitted here.

We now have clear evidence that both the X-ray and CT scan facilities face an overdispersed arrival stream. In our final step of the empirical study we now evaluate the accuracy of our performance measure of interest $\mathbb{E}Q$, and the consequences of assessing system performance while ignoring the presence of overdispersion. We take the following approach: Trivially, we need to choose $s > \mathbb{E}A$ in order for the system to be stable. Hence, we vary s from 70 to 80 for X-rays and from 18 to 24 for CT scans and simulate the queue length process by sampling the number of requests per day from the actual arrival counts. The number of iterations performed is 10^8 for each configuration, excluding a warm-up interval of length 10^7 (days). Around the mean of Q obtained from this simulation, we create a 95% confidence interval. Next, we approximate the expected stationary queue length under two scaling rules. Assuming Poisson arrivals, the appropriate capacity allocation rule would be $s = \hat{\mu} + \gamma\sqrt{\hat{\mu}}$, for some $\gamma > 0$. Our novel capacity sizing rule prescribes $s = \hat{\mu} + \gamma\hat{\sigma} = \hat{a}\hat{b} + \gamma\sqrt{\hat{a}\hat{b}(\hat{b}+1)}$. We compute the first approximation based on square-root safety

s	ρ	$\mathbb{E}Q$ ($\pm$ conf. iv.)	$\mathbb{E}Q^{\text{srs}}$	(3.3)	(3.7)	rel. error
70	0.997	$328.313 \pm 6.6 \cdot 10^{-2}$	186.664	324.231	325.221	$9.6 \cdot 10^{-3}$
71	0.983	$45.073 \pm 1.0 \cdot 10^{-2}$	24.946	45.290	45.308	$5.4 \cdot 10^{-3}$
72	0.970	$21.988 \pm 5.4 \cdot 10^{-3}$	11.650	21.982	22.129	$6.6 \cdot 10^{-3}$
73	0.956	$13.546 \pm 3.6 \cdot 10^{-3}$	6.847	13.455	13.649	$7.8 \cdot 10^{-3}$
74	0.943	$9.230 \pm 2.7 \cdot 10^{-3}$	4.438	9.106	9.319	$1.0 \cdot 10^{-2}$
75	0.931	$6.655 \pm 2.1 \cdot 10^{-3}$	3.031	6.513	6.731	$1.2 \cdot 10^{-2}$
76	0.919	$4.949 \pm 1.7 \cdot 10^{-3}$	2.136	4.821	5.037	$1.8 \cdot 10^{-2}$
77	0.907	$3.755 \pm 1.4 \cdot 10^{-3}$	1.534	3.650	3.861	$2.8 \cdot 10^{-2}$
78	0.895	$2.884 \pm 1.1 \cdot 10^{-3}$	1.115	2.807	3.009	$4.4 \cdot 10^{-2}$
79	0.884	$2.230 \pm 1.0 \cdot 10^{-3}$	0.816	2.183	2.374	$6.5 \cdot 10^{-2}$
80	0.873	$1.734 \pm 8.5 \cdot 10^{-4}$	0.600	1.710	1.890	$9.1 \cdot 10^{-2}$

Table 5 Computational results for X-ray.

s	ρ	$\mathbb{E}Q$ ($\pm$ conf.iv.)	$\mathbb{E}Q^{\text{srs}}$	(3.3)	(3.7)	rel. error
18	0.970	$22.116 \pm 4.9 \cdot 10^{-3}$	14.235	21.965	21.724	$1.8 \cdot 10^{-2}$
19	0.919	$6.289 \pm 1.7 \cdot 10^{-3}$	3.640	5.941	6.040	$4.0 \cdot 10^{-2}$
20	0.873	$3.101 \pm 1.0 \cdot 10^{-3}$	1.589	2.772	2.917	$6.0 \cdot 10^{-2}$
21	0.832	$1.767 \pm 6.6 \cdot 10^{-4}$	0.800	1.507	1.658	$6.1 \cdot 10^{-2}$
22	0.794	$1.066 \pm 4.6 \cdot 10^{-4}$	0.425	0.874	1.016	$4.7 \cdot 10^{-2}$
23	0.760	$0.653 \pm 3.3 \cdot 10^{-4}$	0.230	0.522	0.649	$7.1 \cdot 10^{-3}$
24	0.728	$0.377 \pm 2.3 \cdot 10^{-4}$	0.124	0.315	0.424	$1.2 \cdot 10^{-1}$

Table 6 Computational results for CT scan.

capacity sizing by deducing γ for each s, which we substitute in $\mathbb{E}Q^{\text{srs}} \approx \sqrt{\hat{\mu}}\,\mathbb{E}M_\gamma$. Similarly, we obtain γ from the new rule, and plug in this value, together with the fitted parameters $\hat{a}$ and $\hat{b}$, into (3.7). The results are given in Tables 5 and 6. The last column shows the 95% relative error bound of the second approximation.

Based on these figures, we make several remarks. First, assuming the conventional regime (neglecting overdispersion) the approximation severely overestimates system performance in both queues. Because the square-root safety margin underestimates the stochastic fluctuations in the arrival process, the safety parameter γ is overestimated, which leads to a smaller magnitude of the approximated queue length process. This clearly illustrates the distorted view estimated performance characteristics can give under the wrong scaling. Secondly, it is worth noticing the very good proximity of (3.7) to the values obtained through simulation. As we expected, the quality of the approximation deteriorates with increasing values of s. This makes sense, because we assumed the system to be in heavy traffic in the derivation of the formulas. What is surprising, on the other hand, is the fact that the approximation performs very well, even though the parameter b is very small for these particular data sets, while the analysis of Theorem 2 assumes that a and b become large. This demonstrates that the approximation scheme is remarkably robust and is able to capture the pre-limit behavior of these types of queues very well.

5. Proof of robust approximations (Theorem 2)

For the proof of Theorem 2, we modify the special saddle point method developed in Janssen et al. (2015) to account for the circumstance, caused by overdispersion, that the relevant saddle point and the analyticity radius tend to 1, as $n \to \infty$. Our starting point is the probability generating function of the number of arrivals per time slot, given in (3.5), which is analytic for $|z| < 1 + 1/b_n =: r$. Assuming the same choices of s_n and thereby ρ_n as in Section 2, we consider

$$\mathbb{E}Q_n = \frac{1}{2\pi i} \int_{|z|=1+\varepsilon} \frac{1}{1-z} \, \frac{(z^{s_n} - A_n(z))'}{z^{s_n} - A_n(z)} dz, \tag{5.1}$$

where $0 < \varepsilon < r_0 - 1 < 1/b_n = r - 1$, with r_0 the zero of $z^{s_n} - A_n(z)$ outside $|z| \le 1$ of smallest modulus. We set

$$g(z) = -\ln z + \frac{1}{s_n} \ln A_n(z) = -\ln z - \frac{a_n}{s_n} \ln\left(1 + (1-z)b_n\right), \tag{5.2}$$

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

$$g'(z) = -\frac{1}{z} + \frac{\rho_n}{1 + (1-z)b_n}, \tag{5.3}$$

this yields,

$$1 + (1 - z_{\rm sp})b_n = \rho_n z_{\rm sp}, \quad \text{i.e.,} \quad z_{\rm sp} = 1 + \frac{1-\rho_n}{\rho_n + b_n}. \tag{5.4}$$

We then find

$$\mathbb{E}Q_n = \frac{s_n}{2\pi i} \int_{|z|=1+\varepsilon} \frac{g'(z)}{z-1} \, \frac{\exp(s_n\, g(z))}{1 - \exp(s_n\, g(z))} \, dz, \tag{5.5}$$

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

We continue as in Janssen et al. (2015), Sec. 3, and thus we intend to substitute $z = z(v)$ in the integral in (5.5), where $z(v)$ satisfies

$$g(z(v)) = g(z_{\rm sp}) - \tfrac{1}{2}\, v^2\, g''(z_{\rm sp}) =: q(v) \tag{5.6}$$

on a range $-\frac{1}{2}\delta_n \le v \le \frac{1}{2}\delta_n$. Thus, we consider the approximate representation

$$\frac{-s_n\, g''(z_{\rm sp})}{2\pi i} \int_{-\frac{1}{2}\delta_n}^{\frac{1}{2}\delta_n} \frac{v}{z(v) - 1} \, \frac{\exp(s_n\, q(v))}{1 - \exp(s_n\, q(v))} dv \tag{5.7}$$

of $\mathbb{E}Q_n$. We have to operate here with additional care, since in the present case, the analyticity radius $r = 1 + 1/b_n$, the saddle point $z_{\rm sp}$ as well as the outside zero r_0 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 $b_n \geq 1$, we have from (5.4) that

$$z_{\rm sp} - 1 = \frac{1-\rho_n}{b_n+\rho_n} = \frac{1-\rho_n}{b_n} + O\Big(\frac{1-\rho_n}{b_n^2}\Big), \tag{5.8}$$

where the O-term is small compared to the first term of the right-hand side of (5.8) when $b_n \to \infty$. Next, we approximate r_0, using that $r_0 > 1$ satisfies

$$-\ln r_0 - \frac{\rho_n}{b_n}\ln\left(1+(1-r_0)b_n\right) = 0. \tag{5.9}$$

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

$$\begin{aligned} 0 &= -\ln\left(1+\frac{u}{b_n}\right) - \frac{\rho_n}{b_n}\ln(1-u) \\ &= -\frac{u}{b_n}\Big(1-\rho_n - \tfrac{1}{2}\Big(\frac{1}{b_n}+\rho_n\Big)u - \tfrac{1}{3}\Big(\frac{-1}{b_n^2}+\rho_n\Big)u^2 + \cdots\Big), \end{aligned} \tag{5.10}$$

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

$$u = \frac{2(1-\rho_n)}{\rho_n + 1/b_n} + O(u^2) = 2(1-\rho_n) + O((1-\rho_n)^2) + O\Big(\frac{1-\rho_n}{b_n}\Big), \tag{5.11}$$

and so,

$$r_0 = 1 + 2\,\frac{1-\rho_n}{b_n} + O\Big(\frac{(1-\rho_n)^2}{b_n}\Big) + O\Big(\frac{1-\rho_n}{b_n^2}\Big). \tag{5.12}$$

In (5.7) 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 convergence power series

$$z(v) = z_{\rm sp} + iv + \sum_{k=2}^{\infty} c_k (iv)^k, \qquad \text{for } |v| \leq \tfrac{1}{2}\delta_n. \tag{5.13}$$

To achieve these goals, we supplement the information on $g(z)$, as given by (5.2) – (5.4), by

$$g''(z) = \frac{1}{z^2} + \frac{\rho_n b_n}{(1+(1-z)b_n)^2}, \quad g''(1) = 1+\rho_n b_n, \quad g''(z_{\rm sp}) = \frac{1}{z_{\rm sp}^2}\Big(1+\frac{b_n}{\rho_n}\Big). \tag{5.14}$$

Now

$$\exp(s_n\, q(v)) = \exp(s_n\, g(z_{\rm sp}))\exp(-\tfrac{1}{2}\, s_n\, g''(z_{\rm sp})\, v^2), \tag{5.15}$$

and

$$s_n\, g''(z_{\rm sp}) v^2 = s_n\, b_n v^2 (1+o(1)) = a_n (b_n\, v)^2 (1+o(1)). \tag{5.16}$$

Therefore, (5.7) approximates $\mathbb{E}Q_n$ with exponentially small error where 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 (5.13) that converges for $|v| \leq \frac{1}{2}\delta_n$ with $\frac{1}{2}\delta_n$ of the order $1/b_n$.

LEMMA 2. *Let*

$$r_n := \frac{1}{2\,b_n} - (z_{\rm sp} - 1), \quad m_n := \tfrac{2}{3}\rho_n r_n \sqrt{\frac{b_n + \rho_n^{-1}}{b_n + \rho_n}}, \tag{5.17}$$

where we assume $r_n > 0$, *see* (5.8). *Then* (5.13) *holds with real coefficients* c_k *satisfying*

$$|c_k| \le \frac{r_n}{m_n^k}, \quad k = 2, 3, \ldots. \tag{5.18}$$

Proof We let

$$G(z) := \frac{2(g(z) - g(z_{\rm sp})}{g''(z_{\rm sp})(z - z_{\rm sp})^2}. \tag{5.19}$$

Then $G(z_{\rm sp}) = 1$ and so we can write (5.6) as

$$F(z) := (z - z_{\rm sp})\sqrt{G(z)} = iv \tag{5.20}$$

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 (5.19), 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

$$\mathrm{Re}[g''(z)] > \tfrac{4}{9}\,\rho_n^2 \frac{b_n + \rho_n^{-1}}{b_n + \rho_n}, \quad |z - z_{\rm sp}| \le r_n. \tag{5.21}$$

For this, we consider the representation

$$G(z) = 2\int_0^1 \int_0^1 \frac{g''(z_{\rm sp} + s\,t(z - z_{\rm sp}))}{g''(z_{\rm sp})}\, t\, ds\, dt. \tag{5.22}$$

We have for $\zeta \in \mathbb{C}$ and $|\zeta - 1| \le 1/2b_n \le 1/2$ from (5.14) that

$$\mathrm{Re}[g''(\zeta)] = \mathrm{Re}(1/\zeta^2) + \rho_n b_n\, \mathrm{Re}\Big[\Big(\frac{1}{1 + (1 - \zeta)b_n}\Big)^2\Big] \ge \tfrac{4}{9}(1 + \rho_n b_n). \tag{5.23}$$

To show the inequality in (5.23), it suffices to show that

$$\min_{|\xi - 1| \le 1/2} \mathrm{Re}\Big(\frac{1}{\xi^2}\Big) = \tfrac{4}{9}. \tag{5.24}$$

The minimum in (5.24) is assumed at the boundary $|\xi - 1| = 1/2$, and for a boundary point ξ, we write

$$\xi = 1 + \tfrac{1}{2}\cos\theta + \tfrac{1}{2} i \sin\theta, \quad 0 \le \theta \le 2\pi, \tag{5.25}$$

so that

$$\mathrm{Re}\Big(\frac{1}{\xi^2}\Big) = \frac{1 + \cos\theta + \frac{1}{4}\cos 2\theta}{(\frac{5}{4} + \cos\theta)^2}. \tag{5.26}$$

Now

$$\frac{d}{d\theta}\Big[\frac{1+\cos\theta+\frac{1}{4}\cos 2\theta}{(\frac{5}{4}+\cos\theta)^2}\Big] = \frac{\sin\theta\,(1-\cos\theta)}{4(\frac{5}{4}+\cos\theta)^3} \tag{5.27}$$

vanishes for $\theta = 0, \pi, 2\pi$, where $\mathrm{Re}(1/\xi^2)$ assumes the values 4/9, 4, 4/9, respectively. This shows (5.24).

We use (5.24) with $\zeta+\xi$ and with $\xi = 1+(1-\zeta)b_n$, with

$$\zeta = \zeta(s,t) = z_{\mathrm{sp}} + s\,t\,(z - z_{\mathrm{sp}}), \quad 0 \le s,\, t \le 1, \tag{5.28}$$

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 (5.4) that $0 < g''(z_{\mathrm{sp}}) \le 1 + b_n/\rho_n$. Using this, together with (5.23) where ζ is as in (5.28), yields

$$\mathrm{Re}[G(z)] \le \frac{4}{9}\,\frac{1+\rho_n b_n}{1+b_n/\rho_n}\,2\int_0^1\int_0^1 t\,ds\,dt = \tfrac{4}{9}\,\rho_n^2\,\frac{b_n+\rho_n^{-1}}{b_n+\rho_n} \tag{5.29}$$

when $|z - z_{\mathrm{sp}}| \le r_n$, and this is (5.21).

We therefore have from (5.20) that

$$|F(z)| > r_n \cdot \frac{2}{3}\rho_n\sqrt{\frac{b_n+\rho_n^{-1}}{b_n+\rho_n}} = m_n, \quad |z - z_{\mathrm{sp}}| = r_n. \tag{5.30}$$

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. This $z(v)$ is given by

$$z(v) = \frac{1}{2\pi i}\int_{|z-z_{\mathrm{sp}}|=r_n} \frac{F'(z)\,z}{F(z)-iv}\,dz, \tag{5.31}$$

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

$$|c_k| = \Big|\frac{1}{2\pi i}\int_{|iv|=m_n} \frac{z(v) - z_{\mathrm{sp}}}{(iv)^{k+1}}\,d(iv)\Big| \le \frac{r_n}{m_n^k}, \tag{5.32}$$

and this completes the proof of the lemma.

□

Remark 1. We have $z_{\mathrm{sp}} - 1 = o(1/b_n)$, see (5.8), and so

$$r_n = \frac{1}{2b_n}(1+o(1)), \quad m_n = \frac{1}{3b_n}(1+o(1)), \tag{5.33}$$

implying that the radius of convergence of the series in (5.13) is indeed of order $1/b_n$ (since we have assumed $b_n \ge 1$).

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

$$\frac{v}{z(v)-1} + \frac{-v}{z(-v)-1} = \frac{-2iv\,\mathrm{Im}(z(v))}{|z(v)-1|^2}, \tag{5.34}$$

where we have used that all c_k are real, so that $z(-v) = z(v)^*$. Now from (5.18) and realness of the c_k, we have

$$\mathrm{Im}(z(v)) = v + \sum_{l=1}^{\infty} c_{2l+1}(-1)^l\, v^{2l+1} = v + O(v^3), \tag{5.35}$$

and in similar fashion

$$|z(v)-1|^2 = (z_{\mathrm{sp}}-1)^2 + v^2 + O((z_{rmsp}-1)^2 v^2) + O(v^4) \tag{5.36}$$

when $0 \le v \le \frac{1}{2}\delta_n$. The order terms in (5.35)-(5.36) are negligible in leading order, and so we get for μ_{Q_n} via (5.7) the leading order expression

$$\frac{-s_n\, g''(z_{\mathrm{sp}})}{2\pi i} \int_0^{\frac{1}{2}\delta_n} \frac{-2iv^2}{(z_{\mathrm{sp}}-1)^2+v^2}\, \frac{\exp(s_n\, q(v))}{1-\exp(s_n\, q(v))}\, dv. \tag{5.37}$$

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

$$g(z_{\mathrm{sp}}) = -\tfrac{1}{2}(z_{\mathrm{sp}}-1)^2\, g''(z_1), \tag{5.38}$$

and, see (5.8) and (5.14),

$$g''(z_1) = g''(z_{\mathrm{sp}}) + O((1-\rho_n)b_n). \tag{5.39}$$

Hence

$$\begin{aligned} s_n\, q(v) &= -\tfrac{1}{2}s_n\, g''(z_{\mathrm{sp}})\,[(z_{\mathrm{sp}}-1)^2+v^2] + O((1-\rho_n)b_n s_n (z_{\mathrm{sp}}-1)^2), \\ &= -\tfrac{1}{2}s_n\, g''(z_{\mathrm{sp}})[(z_{\mathrm{sp}}-1)^2+v^2] + O((1-\rho_n)^2 a_n), \end{aligned} \tag{5.40}$$

where (5.8) has been used and $a_n b_n = s_n(1+o(1))$ Therefore, the O-term in (5.40) 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

$$\frac{s_n g''(z_{\mathrm{sp}})}{\pi} \int_0^{\frac{1}{2}\delta_n} \frac{v^2}{(z_{\mathrm{sp}}-1)^2+v^2}\, \frac{\exp(-\frac{1}{2}g''(z_{\mathrm{sp}})s_n((z_{\mathrm{sp}}-1)^2+v^2))}{1-\exp(-\frac{1}{2}g''(z_{\mathrm{sp}})s_n((z_{\mathrm{sp}}-1)^2+v^2))} dv, \tag{5.41}$$

When we substitute $t = v\sqrt{s_n\, g''(z_{\mathrm{sp}})/2}$ and extend the integration in (5.41) to all $t \ge 0$ (at the expense of an exponentially small error), we get for μ_{Q_n} in leading order

$$= \frac{1}{\pi}\sqrt{2\, s_n\, g''(z_{\mathrm{sp}})} \int_0^{\infty} \frac{t^2}{\frac{1}{2}\gamma_n^2}\, \frac{\exp(-\frac{1}{2}\gamma_n^2 - t^2)}{1-\exp(-\frac{1}{2}\gamma_n^2 - t^2)} dt, \tag{5.42}$$

where

$$\gamma_n^2 = s_n\, g''(z_{\mathrm{sp}})(z_{\mathrm{sp}}-1)^2. \tag{5.43}$$

How using (5.4) and (5.14), we get the result of Theorem 2. A separate analysis of γ_n is provided in Subsection 4.1.

A similar analysis, modeled after the one given in (Janssen et al. 2015, Subsecs. 5.2, 5.3) gives under assumption (3.6) the leading-order expression

$$\frac{1}{z_{\rm sp}\pi}\int_0^\infty \frac{\gamma_n/\sqrt{2}}{\frac{1}{2}\gamma_n^2+t^2}\ln(1-e^{-\frac{1}{2}\gamma_n^2-t^2})\,dt \tag{5.44}$$

for $\ln\mathbb{P}(Q_n=0)$. Observe that the quantity in (5.44) is negative, but bounded away from $-\infty$ when γ_n is bounded away from 0. Furthermore, we find for the variance of Q_n the approximation

$$\frac{\gamma_n^3/\sqrt{2}}{\pi}\frac{z_{\rm sp}+1}{(z_{\rm sp}-1)^2}\int_0^\infty \frac{t^2}{(\frac{1}{2}\gamma_n+t^2)^2}\,\frac{\exp(-\frac{1}{2}\gamma_n-t^2)}{1-\exp(-\frac{1}{2}\gamma_n^2-t^2)}\,dt. \tag{5.45}$$

6. Proof of Gaussian approximations (Corollary 3)

According to (Abate et al. 1993, (15)) we have for the maximum M_γ of a Gaussian random walk with drift parameter $-\gamma$ and unit variance

$$-\ln\left[\mathbb{P}(M_\gamma=0)\right]=c_0,\quad \mathbb{E}M_\gamma=c_1,\quad \text{Var } M_\gamma=c_2, \tag{6.1}$$

where

$$c_n=\frac{(-1)^n n!}{\pi}\,\mathrm{Re}\Big[\int_0^\infty \frac{\ln\left(1-\exp(\gamma\,z+\frac{1}{2}z^2)\right)}{z^{n+1}}dy\Big], \tag{6.2}$$

in which $z=-x+i\,y$, $y\geq 0$, and x is any fixed number between 0 and 2γ. We take $x=\gamma$, so that

$$\gamma z+\tfrac{1}{2}z^2=-\tfrac{1}{2}\gamma^2-\tfrac{1}{2}y^2\leq 0,\quad y\geq 0. \tag{6.3}$$

For $n=0$, we then have

$$\begin{aligned}
c_0&=\frac{1}{\pi}\,\mathrm{Re}\Big[\int_0^\infty \frac{\ln\left(1-\exp(-\frac{1}{2}\gamma^2-\frac{1}{2}y^2)\right)}{-\gamma+i\,y}dy\\
&=-\frac{1}{\pi}\int_0^\infty \frac{\gamma}{\gamma^2+y^2}\ln\left(1-\exp(-\tfrac{1}{2}\gamma^2-\tfrac{1}{2}y^2)\right)dy,\\
&=-\frac{1}{\pi}\int_0^\infty \frac{\gamma/\sqrt{2}}{\frac{1}{2}\gamma^2+t^2}\ln\left(1-\exp(-\tfrac{1}{2}\gamma^2-t^2)\right)dt,
\end{aligned} \tag{6.4}$$

where we used that

$$\mathrm{Re}\Big[\frac{1}{-\gamma+i\,y}\Big]=\frac{-\gamma}{\gamma^2+y^2}, \tag{6.5}$$

together with the substitution $y=t\sqrt{2}$. For $n=1,2,\cdots$, we have by partial integration

$$\begin{aligned}
c_n&=\frac{(-1)^n n!}{\pi}\,\mathrm{Re}\Big[\int_0^\infty \frac{\ln(1-\exp(-\frac{1}{2}\gamma^2-\frac{1}{2}y^2))}{(-\gamma+i\,y)^{n+1}}dy\\
&=\frac{(-1)^{n-1}(n-1)!}{\pi}\,\mathrm{Im}\Big[\int_0^\infty \ln(1-\exp(-\tfrac{1}{2}\gamma^2-\tfrac{1}{2}y^2))\,d\Big(\frac{1}{(-\gamma+i\,y)^n}\Big)\Big]\\
&=-\frac{(-1)^{n-1}(n-1)!}{\pi}\mathrm{Im}\Big[\int_0^\infty \frac{y}{(-\gamma+i\,y)^n}\,\frac{\exp(-\frac{1}{2}\gamma^2-\frac{1}{2}y^2)}{1-\exp(-\frac{1}{2}\gamma^2-\frac{1}{2}y^2)}dy\Big],
\end{aligned} \tag{6.6}$$

where we have used that

$$\mathrm{Im}\Big[\frac{\ln(1-\exp(-\frac{1}{2}\gamma^2-\frac{1}{2}y^2))}{(-\gamma+i\,y)^n}\Big]\Big|_0^\infty = 0. \tag{6.7}$$

Using

$$\frac{1}{(-\gamma+i\,y)^n} = (-1)^n\,\frac{(\gamma+i\,y)^n}{(\gamma^2+y^2)^n}, \tag{6.8}$$

we then get

$$c_n = \frac{(n-1)!}{\pi}\,\mathrm{Im}\Big[\int_0^\infty \frac{y(\gamma+i\,y)^n}{(\gamma^2+y^2)^n}\,\frac{\exp(-\frac{1}{2}\gamma^2-\frac{1}{2}y^2)}{1-\exp(-\frac{1}{2}\gamma^2-\frac{1}{2}y^2)}dy\Big]. \tag{6.9}$$

Hence for $n=1,2$, we finally get by the substitution $y=t\sqrt{2}$

$$\begin{aligned} c_1 &= \frac{1}{\pi}\int_0^\infty \frac{y^2}{\gamma^2+y^2}\,\frac{\exp(-\frac{1}{2}\gamma^2-\frac{1}{2}y^2)}{1-\exp(-\frac{1}{2}\gamma^2-\frac{1}{2}y^2)}\,dy \\ &= \frac{\sqrt{2}}{\pi}\int_0^\infty \frac{t^2}{\frac{1}{2}\gamma^2+t^2}\,\frac{\exp(-\frac{1}{2}\gamma^2-t^2)}{1-\exp(-\frac{1}{2}\gamma^2-t^2)}\,dt, \end{aligned} \tag{6.10}$$

$$\begin{aligned} c_2 &= \frac{2\gamma}{\pi}\int_0^\infty \frac{y^2}{(\gamma^2+y^2)^2}\,\frac{\exp(-\frac{1}{2}\gamma^2-\frac{1}{2}y^2)}{1-\exp(-\frac{1}{2}\gamma^2-\frac{1}{2}y^2)}\,dy \\ &= \frac{\gamma\sqrt{2}}{\pi}\int_0^\infty \frac{t^2}{(\frac{1}{2}\gamma^2+t^2)^2}\,\frac{\exp(-\frac{1}{2}\gamma^2-t^2)}{1-\exp(-\frac{1}{2}\gamma^2-t^2)}\,dt, \end{aligned} \tag{6.11}$$

7. Proofs of convergence results (Theorem 1)

This section presents the details of the proof of Lemma 1 and Theorem 1, using the random walk perspective of the process $\{Q_n(k)\}_{k=0}^\infty$. 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:

$$Y_n(k) := \hat{A}_n(k)-\gamma, \quad S_n(k) = \sum_{i=1}^k Y_n(i), \tag{7.1}$$

with $S_0^n = 0$ and $k=1,2,....$ Then (2.6) can be rewritten as

$$\hat{Q}_n \overset{d}{=} \max_{0\le i\le k}\Big\{\textstyle\sum_{i=1}^k Y_n(i)\Big\} =: M_{\gamma,n}, \tag{7.2}$$

Last, we introduce the sequence of independent normal random variables $Z(1), Z(2), ...$ with mean *gamma* and unit variance 1, and

$$M_\gamma \overset{d}{=} \max_{k\ge 0}\{\textstyle\sum_{i=1}^k Z(i)\} \tag{7.3}$$

7.1. Proof of Lemma 1

Proof We show weak convergence of the random variable $\hat{A}_n$, as defined in (7.1), 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.

$$\lim_{n\to\infty} \varphi_{\hat{\Lambda}_n}(t) = \lim_{n\to\infty} \mathrm{e}^{-i\mu_n t/\sigma_n}\, \varphi_{\Lambda_n}(t/\sigma_n) = \mathrm{e}^{-t^2/2}, \qquad \forall t\in\mathbb{R}. \tag{7.4}$$

Furthermore, by definition of A_k^n,

$$\varphi_{A_k^n}(t) = \mathbb{E}\left[\exp(\Lambda_n(\mathrm{e}^{it}-1))\right] = \varphi_{\Lambda_n}\left(-i(\mathrm{e}^{it}-1)\right), \tag{7.5}$$

so that

$$\varphi_{\hat{A}_k^n}(t) = \mathrm{e}^{-i\mu_n t/\sigma_n}\, \varphi_{A_k^n}(t/\sigma_n) = \mathrm{e}^{-i\mu_n t/\sigma_n} \varphi_{\Lambda_n}\left(-i(\mathrm{e}^{it/\sigma_n}-1)\right). \tag{7.6}$$

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

$$-i(\mathrm{e}^{it/\sigma_n}-1) = \frac{t}{\sigma_n} - \frac{it^2}{2\sigma_n^2} + O\left(t^3/\sigma_n^3\right), \tag{7.7}$$

we expand the last term in (7.6),

$$\varphi_{\Lambda_n}(t/\sigma_n) + \Big(-\frac{i\,t^2}{2\sigma_n^2} + O\left(t^3/\sigma_n^3\right)\Big)\varphi'_{\Lambda_n}(t/\sigma_n) + O\Big(\Big(-\frac{i\,t^2}{2\sigma_n^2} + O\left(t^3/\sigma_n^3\right)\Big)^2 \varphi''_{\Lambda_n}(t/\sigma_n)\Big) \tag{7.8}$$

$$= \varphi_{\Lambda_n}(t/\sigma_n) - \Big(\frac{i\,t^2}{2\sigma_n^2} + O\left(t^3/\sigma_n^3\right)\Big)\varphi'_{\Lambda_n}(\zeta) \tag{7.9}$$

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

$$\begin{aligned} |\varphi'_{\Lambda_n}(u)| &= \left|\frac{d}{du}\int_{-\infty}^{\infty} \mathrm{e}^{iux}\, dF_{\Lambda_n}(x)\right| = \left|\int_0^{\infty} ix\,\mathrm{e}^{iux}\, dF_{\Lambda_n}(x)\right| \\ &\le \int_{-\infty}^{\infty} |ix\,\mathrm{e}^{iux}|\, dF_{\Lambda_n}(x) = \int_0^{\infty} x\, dF_{\Lambda_n}(x) = \mu_n \end{aligned} \tag{7.10}$$

for all $u\in\mathbb{R}$. Hence, by substituting (7.6),

$$\begin{aligned} \left|\varphi_{\hat{A}_k^n}(t) - \mathrm{e}^{-i\mu_n t/\sigma_n}\varphi_{\Lambda_n}(t/\sigma_n)\right| &= \left|\mathrm{e}^{-i\mu_n t/\sigma_n}\, \left(\frac{i\,t^2}{2\sigma_n^2} + O(t^3/\sigma_n^3)\right)\, \varphi'_{\Lambda_n}(\zeta)\right| \\ &\le \left(\frac{t^2}{2\sigma_n^2} + O(t^3/\sigma_n^3)\right) |\varphi'_{\Lambda_n}(\zeta)| \\ &= \frac{\mu_n t^2}{\sigma_n^2} + O\left(\frac{\mu_n t^3}{\sigma_n^3}\right), \end{aligned} \tag{7.11}$$

which tends to zero as $n\to\infty$ by our assumption that $\mu_n/\sigma_n^2 \to 0$. Finally,

$$\left|\varphi_{\hat{A}_k^n}(t) - \mathrm{e}^{-\frac{1}{2}t^2}\right| \le \left|\varphi_{\hat{A}_k^n}(t) - \mathrm{e}^{-i\mu_n t/\sigma_n}\varphi_{\Lambda_n}(t/\sigma_n)\right| + \left|\mathrm{e}^{-i\mu_n t/\sigma_n}\varphi_{\Lambda_n}(t/\sigma_n) - \mathrm{e}^{-\frac{1}{2}t^2}\right|, \tag{7.12}$$

in which both terms go to zero for $n\to\infty$, by (7.4) and (7.11). Hence $\varphi_{\hat{A}_{n(k)}}(t)$ converges to $\mathrm{e}^{-t^2/2}$ for all $t\in\mathbb{R}$, so that we can conclude by Lévy's continuity theorem that $\hat{A}_k^n \overset{d}{\Rightarrow} \mathcal{N}(0,1)$.

□

7.2. Proof of Theorem 1

To secure convergence in distribution of $\hat{Q}^n$ to M_γ, i.e. the maximum of a Gaussian random walk with negative drift, the first assertion of Theorem 1. the following property of the sequence $\{Y_k^n\}_{n\in\mathbb{N}}$ needs to hold.

LEMMA 3. *Let $Y_n(k)$ be defined as in (7.1) with $\mu_n, \sigma_n^2 < \infty$ for all $n \in \mathbb{N}$. Then the sequence $\{(Y_k^n)^+\}_{n\in\mathbb{N}}$ is uniform integrable, i.e.*

$$\lim_{K\to\infty} \sup_n \mathbb{E}[Y_n(k)^+ | 1_{\{|Y_n(k)^+|\geq K\}}] = 0. \tag{7.13}$$

Proof Because the sequence $\{Y_n(k)\}_{k\in\mathbb{N}}$ is i.i.d. for all n, we omit the index k in this proof. First, fix $K > 0$ and note that

$$\mathbb{E}[|Y_n^+|1_{\{|Y_n^+|\geq K\}}] = \mathbb{E}[Y_n^+ 1_{\{Y_n^+\geq K\}}] = \mathbb{E}[Y_n 1_{\{Y_n\geq K\}}]. \tag{7.14}$$

This last expression can be bounded from above using the Cauchy-Schwarz inequality, so that

$$\mathbb{E}[Y_n 1_{\{Y_n\geq K\}}] \leq \mathbb{E}[Y_n^2]^{1/2}\, \mathbb{P}(Y_n \geq K)^{1/2}. \tag{7.15}$$

By the definition of Y_n, we know $\mathbb{E}Y_n = -\gamma$ and $\operatorname{Var} Y_n = \operatorname{Var} A_n/\sigma_n^2 = 1$. Using this information, we find

$$\mathbb{E}[Y_n^2] = \operatorname{Var} Y_n + (\mathbb{E}Y_n)^2 = 1 + \gamma^2 \tag{7.16}$$

and

$$\begin{aligned} \mathbb{P}(Y_n \geq K) &= \mathbb{P}(Y_n + \gamma \geq K + \gamma) \leq \mathbb{P}(|Y_n + \gamma| \geq K + \gamma) \\ &\leq \frac{\operatorname{Var} Y_n}{(K+\gamma)^2} = \frac{1}{(K+\gamma)^2}, \end{aligned} \tag{7.17}$$

where we used Chebyshev's inequality for the last upper bound. Therefore,

$$\begin{aligned} \lim_{K\to\infty} \sup_n \mathbb{E}[|Y_n^+|1_{\{|Y_n^+|\geq K\}}] &= \lim_{K\to\infty} \sup_n \mathbb{E}[Y_n 1_{\{Y_n\geq K\}}] \\ &\leq \lim_{K\to\infty} \sup_n \mathbb{E}[Y_n^2]^{1/2}\, \mathbb{P}(Y_n \geq K)^{1/2} \\ &\leq \lim_{K\to\infty} \frac{\sqrt{1+\gamma^2}}{K+\gamma} = 0. \end{aligned} \tag{7.18}$$

□

By combining the properties proved in Lemma 1 and 3, the next result follows directly by (Asmussen 2003, Thm.X6.1).

PROPOSITION 3. *Let $\hat{Q}_n$ as in (7.2), and $\mu_n, \sigma_n^2 \to \infty$ such that $\sigma_n^2/\mu_n \to \infty$. Then*

$$\hat{Q}_n \overset{d}{\Rightarrow} M_\gamma, \qquad \text{as } n \to \infty. \tag{7.19}$$

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 the approach similar Sigman and Whitt (2011b).

PROPOSITION 4. *Let $\hat{Q}_n$ as in (7.2), $\mu_n, \sigma_n^2 \to \infty$ such that $\sigma_n^2/\mu_n \to \infty$ and $\mathbb{E}\hat{A}_n^3 < \infty$. Then*

$$\mathbb{P}(\hat{Q}_n = 0) \to \mathbb{P}(M_\gamma = 0), \tag{7.20}$$

$$\mathbb{E}\hat{Q}_n \to \mathbb{E}M_\gamma, \tag{7.21}$$

$$\text{Var } \hat{Q}_n \to \text{Var } M_\gamma, \tag{7.22}$$

as $n \to \infty$.

*F*irst, we recall that $\hat{Q}_n \overset{d}{=} M_{\gamma,n}$ for all $n \in \mathbb{N}$, so that $\mathbb{P}(\hat{Q}_n = 0) = \mathbb{P}(M_{\gamma,n} = 0)$, $\mathbb{E}\hat{Q}_n = \mathbb{E}M_{\gamma,n}$ and $\text{Var } \hat{Q}_n = \text{Var } M_{\gamma,n}$ as defined in (7.1). Our starting point is Spitzer's identity, see (Asmussen 2003, p. 230),

$$\mathbb{E}[\mathrm{e}^{itM_{\gamma,n}}] = \exp\Big(\sum_{k=1}^{\infty} \frac{1}{k}(\mathbb{E}[\mathrm{e}^{it(S_n(k))^+}] - 1)\Big), \tag{7.23}$$

with $S_n(k)$ as in (7.1), and $M_{\gamma,n}$ the all-time maximum of the associated random walk. Simple manipulations of (7.23) give

$$\ln \mathbb{P}(M_{\gamma,n} = 0) = -\sum_{k=1}^{\infty} \frac{1}{k}\, \mathbb{P}(S_n(k) > 0), \tag{7.24}$$

$$\mathbb{E}M_{\gamma,n} = \sum_{k=1}^{\infty} \frac{1}{k}\mathbb{E}[S_n(k)^+] = \sum_{k=1}^{\infty} \frac{1}{k}\int_0^{\infty} \mathbb{P}(S_n(k) > x)\, dx, \tag{7.25}$$

$$\text{Var}\, M_{\gamma,n} = \sum_{k=1}^{\infty} \frac{1}{k}\mathbb{E}[(S_n(k)^+)^2] = \sum_{k=1}^{\infty} \frac{1}{k}\int_0^{\infty} \mathbb{P}(S_n(k) > \sqrt{x})\, dx. \tag{7.26}$$

By Lemma 1, we know

$$\mathbb{P}(S_n(k) > y) = \mathbb{P}\left(\textstyle\sum_{i=1}^{k} Y_n(i) > y\right) \to \mathbb{P}\left(\textstyle\sum_{i=1}^{k} Z(i) > y\right), \tag{7.27}$$

for $n \to \infty$, where the $Z(i)$'s are independent and identically normally distributed with mean $-\gamma$ and variance 1. Because equivalent expressions to (7.24)-(7.26) 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

$$\mathbb{P}(S_n(k) > 0) \le \mathbb{P}(|S_n(k) + k\gamma| > k\gamma) \le \frac{k}{\gamma^2 k^2} = \frac{1}{\gamma^2 k}, \tag{7.28}$$

where we used that $\mathbb{E}S_n(k) = k\mathbb{E}Y_n(1) = -k\gamma$ and $\operatorname{Var} S_n(k) = k$ and apply Chebychev's inequality, so that

$$\sum_{k=1}^{\infty} \frac{1}{k}\mathbb{P}(S_n(k) > 0) \leq \sum_{k=1}^{\infty} \frac{1}{\gamma^2 k^2} < \infty, \qquad \forall n \in \mathbb{N}. \tag{7.29}$$

Hence,

$$\begin{aligned}\lim_{n\to\infty} \ln \mathbb{P}(\hat{Q}_n = 0) &= \lim_{n\to\infty} -\sum_{k=1}^{\infty} \frac{1}{k}\mathbb{P}(S_n(k) > 0) = -\sum_{k=1}^{\infty} \frac{1}{k} \lim_{n\to\infty} \mathbb{P}(S_n(k) > 0) \\ &= -\sum_{k=1}^{\infty} \frac{1}{k}\mathbb{P}(\textstyle\sum_{i=1}^{k} Z(i) > 0) = \ln \mathbb{P}(M_\gamma = 0),\end{aligned} \tag{7.30}$$

Finding a suitable upper bound on $\frac{1}{k}\int_0^\infty \mathbb{P}(\hat{Q}_n > x)dx$ and $\frac{1}{k}\int_0^\infty \mathbb{P}(\hat{Q}_n > \sqrt{x})dx$ requires a bit more work. We initially focus on the former, the latter follows easily. The following inequality from Nagaev (1979) proves to be very useful:

$$\mathbb{P}(\bar{S}(k) > y) \leq C_r \left(\frac{k\,\sigma^2}{y^2}\right)^2 + k\,\mathbb{P}(X > y/r), \tag{7.31}$$

where $\bar{S}(k)$ is the sum of k i.i.d. random variables distributed as X, with $\mathbb{E}X = 0$ and $\operatorname{Var} X = \sigma^2$, $y > 0$, $r > 0$ 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 analyze the integral in two parts, one for the interval $(0, k)$ and one for $[k, \infty)$. For the first part, we have

$$\begin{aligned}\int_0^k \mathbb{P}(S_n(k) > x)dx &= \int_0^k \mathbb{P}(\textstyle\sum_{i=1}^{\infty} \hat{A}_n(i) > x + k\gamma)dx \leq \int_0^k \mathbb{P}(\sum_{i=1}^{\infty} \hat{A}_n(i) > k\gamma)dx \\ &= k\,\mathbb{P}(\textstyle\sum_{i=1}^{k} \hat{A}_n(i) > k\gamma) \leq \frac{C_3}{k^2\gamma^6} + k^2\mathbb{P}(\hat{A}_n(1) > \tfrac{1}{3}k),\end{aligned} \tag{7.32}$$

where we used (7.31) in the last inequality. Hence,

$$\sum_{k=1}^{\infty} \frac{1}{k} \int_0^k \mathbb{P}(S_n(k) > x)dx \leq \frac{C_3}{\gamma^6} \sum_{k=1}^{\infty} k^{-3} + \sum_{k=1}^{\infty} k\,\mathbb{P}(\hat{A}_n(1) > \tfrac{1}{3}k) \leq C_1^* + \sum_{k=1}^{\infty} k\,\mathbb{P}(\hat{A}_n(1) > \tfrac{1}{3}k). \tag{7.33}$$

With the help of the inequality (see Sigman and Whitt (2011b)),

$$(b-a)a\,\mathbb{P}(X > b) \leq \int_a^b x\,\mathbb{P}(X > x)dx \qquad \forall 0 < a < b, \tag{7.34}$$

we get by taking $a = (k-1)/3$ and $b = k/3$,

$$k\,\mathbb{P}(\hat{A}_n(1) > \tfrac{1}{3}k) \leq \frac{9\,k}{k-1} \int_{(k-1)/3}^{k/3} x\,\mathbb{P}(\hat{A}_n(1) > x)dx \leq 18 \int_{(k-1)/3}^{k/3} x\,\mathbb{P}(\hat{A}_n(1) > x)dx, \tag{7.35}$$

for $k \geq 2$. Since the tail probability for $k = 1$ is obviously bounded by 1, this yields

$$\begin{aligned}\sum_{k=1}^{\infty} k\,\mathbb{P}(\hat{A}_n(1) > \tfrac{1}{3}k) &\leq 1 + 18 \sum_{k=2}^{\infty} \int_{(k-1)/3}^{k/3} x\,\mathbb{P}(\hat{A}_n(1) > x)dx \\ &\leq 1 + \int_0^{\infty} x\,\mathbb{P}(\hat{A}_n(1) > x)dx \leq 1 + \mathbb{E}[\hat{A}_n(1)^2] < \infty,\end{aligned} \tag{7.36}$$

since $\hat{A}_n(1)$ has finite variance by assumption. This completes the integral over the first interval. For the second part, we use (7.31) again to find

$$\begin{aligned}\int_k^\infty \mathbb{P}(S_n(k)>x)dx &= \int_k^\infty \mathbb{P}(\textstyle\sum_{i=1}^\infty \hat{A}_n(i) > x+k\gamma)dx \leq \int_k^\infty \mathbb{P}(\sum_{i=1}^\infty \hat{A}_n(i)>x)dx \\ &\leq C_3\int_k^\infty \frac{k^2}{x^6}dx + k\int_k^\infty \mathbb{P}(\hat{A}_n(1)>\tfrac{1}{3}x)dx = \frac{5\,C_3}{k^3} + k\int_k^\infty \mathbb{P}(\hat{A}_n(i)>\tfrac{1}{3}x)dx.\end{aligned} \tag{7.37}$$

So,

$$\sum_{k=1}^\infty \frac{1}{k}\int_k^\infty \mathbb{P}(S_n(k)>x)dx \leq C_2^* + \sum_{k=1}^\infty \int_k^\infty \mathbb{P}(\hat{A}_n(i)>\tfrac{1}{3}x)dx, \tag{7.38}$$

for some constant C_2^*. Last, we are able to upper bound the second term in (7.38) by Tonelli's theorem:

$$\begin{aligned}\sum_{k=1}^\infty \int_k^\infty \mathbb{P}(\hat{A}_n(i)>\tfrac{1}{3}x)dx &\leq \int_1^\infty x\mathbb{P}(\hat{A}_n(1)>\tfrac{1}{3}x)dx \\ &\leq 9\int_0^\infty y\mathbb{P}(\hat{A}_n(1)>y)dy = 9\mathbb{E}[\hat{A}_n(1)^2] < \infty.\end{aligned} \tag{7.39}$$

Combining the results in (7.33), (7.36), (7.38) and (7.39), we find

$$\sum_{k=1}^\infty \frac{1}{k}\int_0^\infty \mathbb{P}(S_n(k)>x)dx < \infty, \tag{7.40}$$

and thus

$$\lim_{n\to\infty}\mathbb{E}\hat{Q}_n = \lim_{n\to\infty}\sum_{k=1}^\infty \frac{1}{k}\int_0^\infty \mathbb{P}(S_n(k)>x)dx = \sum_{k=1}^\infty \frac{1}{k}\int_0^\infty \mathbb{P}(\textstyle\sum_{i=1}^k Z(i)>x)dx = \mathbb{E}M_\gamma. \tag{7.41}$$

Finally, we show how the proof changes for the convergence of $\operatorname{Var}\hat{Q}_n$. The expressions for $\mathbb{E}\hat{Q}_n$ and $\operatorname{Var}\hat{Q}_n$ in (7.24) and (7.25) only differ in the term $\sqrt{x}$. Hence only minor modifications are needed to also prove convergence of the variance. Note that boundedness of the integral over the interval $(0,k)$ in (7.32)-(7.36) remains to hold when substituting $\sqrt{x}$ for x. (7.37) changes into

$$\begin{aligned}\int_k^\infty \mathbb{P}(S_n(k)>\sqrt{x})dx &= \int_k^\infty \mathbb{P}(\textstyle\sum_{i=1}^\infty \hat{A}_n(i) > \sqrt{x}+k\gamma)dx \\ &\leq C_3\int_k^\infty \frac{k^2}{(\sqrt{x}+k\gamma)^6}dx + k\int_k^\infty \mathbb{P}(\hat{A}_n(1)>\tfrac{1}{3}\sqrt{x})dx \\ &\leq \frac{C_4^*}{k} + k\int_k^\infty \mathbb{P}(\hat{A}_n(1)>\tfrac{1}{3}\sqrt{x})dx,\end{aligned} \tag{7.42}$$

for some constant C_4^*, so that

$$\sum_{k=1}^\infty \frac{1}{k}\int_k^\infty \mathbb{P}(S_n(k)>\sqrt{x})dx \leq C_4^* + \sum_{k=1}^\infty \int_k^\infty \mathbb{P}(\hat{A}_n(1)>\tfrac{1}{3}\sqrt{x})dx. \tag{7.43}$$

Lastly, we have

$$\sum_{k=1}^{\infty}\int_{k}^{\infty}\mathbb{P}(\hat{A}_n(1)>\tfrac{1}{3}\sqrt{x})dx\leq\int_{1}^{\infty}x\mathbb{P}(\hat{A}_n(1)>\tfrac{1}{3}\sqrt{x})dx$$
$$\leq 18\int_{0}^{\infty}y^2\mathbb{P}(\hat{A}_n(1)>y)dy=18\mathbb{E}[\hat{A}_n(1)^3]<\infty. \tag{7.44}$$

Therefore the sum describing the variance is also uniformly convergent in n, so that interchanging of infinite sum and limit is permitted and

$$\lim_{n\to\infty}\text{Var }\hat{Q}_n=\lim_{n\to\infty}\sum_{k=1}^{\infty}\frac{1}{k}\int_{0}^{\infty}\mathbb{P}(S_n(k)>\sqrt{x})dx=\sum_{k=1}^{\infty}\frac{1}{k}\int_{0}^{\infty}\mathbb{P}(\textstyle\sum_{i=1}^{k}Z(i)>\sqrt{x})dx=\text{Var }M_\gamma. \tag{7.45}$$

References

Abate, J., G.L. Choudhury, W. Whitt. 1993. Calculation of the GI/G/1 waiting-time distribution and its cumulants from Pollaczek's formulas. *Archiv fur Elektronik und Ubertragungstechnik (International Journal of Electronics and Communication)* **47** 311–321.

Anick, D., D. Mitra, M.M. Sondhi. 1982. Stochastic theory of a data-handling system with multiple sources. *The Bell System Technical Journal* **61**(8).

Asmussen, S. 2003. *Applied Probability and Queues (second edition)*. Springer-Verlag, New York.

Avramidis, A.N., A. Deslauriers, P. L'Ecuyer. 2004. Rate-based daily arrival process models with application to call centers. *Management Science* **50**(7) 893–908.

Bassamboo, A, R.S. Randhawa, A. Zeevi. 2010. Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Manag. Sci.* **56**(10) 1668–1686.

Bassamboo, A., A. Zeevi. 2009. On a data-driven method for staffing large call centers. *Oper. Res.* **57**(3) 714–726.

Blanchet, J., P.W. Glynn. 2006. Complete corrected diffusion approximations for the maximum of a random walk. *The Annals of Applied Probability* **16**(2) 951–983.

Borst, S., A. Mandelbaum, M.I. Reiman. 2004. Dimensioning large call centers. *Operations Research* **52**(1) 17–34.

Boudreau, P.E., J.S. Griffin Jr., M. Kac. 1962. An elementary queueing problem. *The American Mathematical Monthly* **69**(8) 713–724.

Brown, L.D., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: a queueing-science perspective. *Journal of the American Statistical Association* **100** 36–50.

Brown, L.D., L.H. Zhao. 2002. A test for the Poisson distribution. *The Indian Journal of Statistics* **64** 611–625.

Chang, J.T., Y. Peres. 1997. Ladder heights, Gaussian random walks and the Riemann zeta function. *Ann. Probab.* **25**(2) 787–802.

Chen, B.P.K., S. G. Henderson. 2001. Two issues in setting call center staffing levels. *Ann. Oper. Res.* **108** 175–192.

Cohen, J.W. 1982. *The Single Server Queue*, *North-Holland Series in Applied Mathematics and Mechanics*, vol. 8. 2nd ed. North-Holland Publishing Co., Amsterdam.

Cox, D.R. 1955. Some statistical models connected with series of events. *Journal of the Royal Statistical Society* **17**(2) 129–164.

de Bruijn, N.G. 1981. *Asymptotic methods in analysis*. Dover Publications.

Ding, S., G. Koole. 2014. Optimal call center forforecast and staffing under arrival rate uncertainty.

Gans, N., H. Shen, Y. Zhou, N. Korolev, A. McCord, H. Ristock. 2012. Parametric stochastic programming models for call-center workforce scheduling.

Glasserman, P. 2003. *Monte Carlo methods in financial engineering*. Springer.

Grandell, J. 1997. *Mixed Poisson Processes*. Monographs on Statistics and Applied Probability, Chapman & Hall.

Green, L.V. 2004. *Operations research and health care: A handbook of methods and applications*, chap. 1. Kluwer.

Green, L.V., S. Savin, M. Murray. 2007. Providing timely access to care: What is the right patient panel size? *The Joint Commission Journal on Quality and Patient Safety* **33** 211–218.

Gurvich, I., J. Luedtke, T. Tezcan. 2010. Staffing call-centers with uncertain demand forecasts: a chance-constrained optimization approach. *Manage. Sci.* **56**(7) 1093–1115.

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29**(3) 567–588.

Izady, N. 2015. Appointment capacity planning in specialty clinics: A queueing approach.

Janssen, A.J.E.M, J.S.H. van Leeuwaarden. 2005. Analytic computation schemes for the discrete-time bulk service queue. *Queueing Syst.* **50** 141–163.

Janssen, A.J.E.M, J.S.H. van Leeuwaarden. 2006. On Lerchs transcendent and the Gaussian random walk. *Ann.of Appl. Probab.* **17**(2) 421–439.

Janssen, A.J.E.M., J.S.H. van Leeuwaarden, B.W.J. Mathijsen. 2015. Novel heavy-traffic regimes for large-scale service systems. *SIAM Journal of Applied Mathematics* **75**(2) 787–812.

Jelenkovic, P., A. Mandelbaum, P. Momcilovic. 2004. Heavy traffic limits for queues with many deterministic servers. *Queueing Syst.* **47** 53–69.

Johnson, N.L., A.W. Kemp, S. Kotz. 1993. *Univariate discrete distributions*. 2nd ed. John Wiley & Sons.

Jongbloed, G., G. Koole. 2001. Managing uncertainty in call centres using Poisson mixtures. *Applied Stoch. Mod. Bus.* **17**(4) 307–318.

Kim, S-H., P. Vel, W. Whitt, W.C. Cha. 2015a. Poisson and non-Poisson properties in appointment-generated arrival processes: The case of an endocrinology clinic. *Operations Research Letters* **43** 247–253.

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

Kim, S-H., W. Whitt, W.C. Cha. 2015b. A data-driven model of an appointment-generated arrival process at an outpatient clinic.

Koçaga, Y.L., M. Armony, A.R. Ward. 2014. Staffing call centers with uncertain arrival rate and co-sourcing.

Maman, S. 2009. Uncertainty in the demand for service: The case of call centers and emergency departments. Master's thesis, Technion - Israel Institute of Technology, Haifa, Israel.

Mehrotra, V., O. Ozlük, R. Saltzmann. 2010. Intelligent procedures for intra-day updating of call center agent schedules. *Production and operations management* **19**(3) 353–367.

Murray, M., T. Bodenheimer, D. Rittenhouse, K. Grumbach. 2003. Improving timely access to primary care: Case studies of the advanced access model. *Journal of the American Medical Association* **289**(8) 1042–1046.

Murray, M., C. Tantau. 1999. Redefining open access to primary care. *Managed Care Quarterly* **7**(3) 45–55.

Nagaev, S.V. 1979. Large deviations of sums of independent random variables. *Ann. Probab.* **7**(5) 745–789.

Robbins, T.R., D.J. Medeiros, T.P. Harrison. 2010. Does the Erlang C model fit in real call centers? *Proceedings of the 2010 Winter Simulation Conference*.

Siegmund, D. 1978. Corrected diffusion approximations in certain random walk problems. Tech. rep., Stanford university.

Sigman, K., W. Whitt. 2011a. Heavy-traffic limits for nearly deterministic queues. *Journal of Applied Probability* **48**(3) 657–678.

Sigman, K., W. Whitt. 2011b. Heavy-traffic limits for nearly deterministic queues: stationary distributions. *Queueing Systems* **69** 145–173.

Steckley, S.G., S.G. Henderson, V. Mehrotra. 2009. Forecast errors in service systems. *Probability in the Engineering and Informational Sciences* **23** 305–332.

Whitt, W. 1999. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters* **24** 205–212.

Whitt, W. 2006. Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management* **15**(1) 88–102.

Zacharias, C., M. Armony. 2014. Joint panel sizing and appointment scheduling in outpatient care.

Zan, J. 2012. Staffing service centers under arrival-rate uncertainty. Ph.D. thesis, University of Texas.

Appendix A: Numerical procedures

An alternative characterization of the stationary distribution is based on the analysis in Boudreau et al. (1962) and considers a factorization in terms of (complex) roots:

$$Q_n(w) = \frac{(s_n - \mathbb{E}A_n)(w-1)}{w^{s_n} - A_n(w)} \prod_{k=1}^{s_n-1} \frac{w - z_k^n}{1 - z_k^n}, \tag{A.1}$$

where $z_1^n, z_2^n ..., z_{s_n-1}^n$ are the $s_n - 1$ zeros of $z^{s_n} - A_n(z)$, in $|z| < 1$, yielding

$$\mathbb{E}Q_n = \frac{\sigma_n^2}{2(s_n - \mu_n)} - \frac{s_n - 1 + \mu_n}{2} + \sum_{k-1}^{s_n-1} \frac{1}{1 - z_k^n}, \tag{A.2}$$

$$\mathbb{P}(Q_n = 0) = \frac{s_n - \mu_A}{A_n(0)} \prod_{k=1}^{s-1} \frac{z_k^n}{z_k^n - 1}, \tag{A.3}$$

which for our choice of $A_n(z)$ becomes

$$\mathbb{E}Q_n = \frac{a_n b_n (b_n + 1)}{2\gamma\sqrt{a_n} b_n} - \frac{2a_n b_n + \gamma\sqrt{a_n b_n (b_n+1)} - 1}{2} + \sum_{k=1}^{s_n-1} \frac{1}{1 - z_k^n}, \tag{A.4}$$

$$\mathbb{P}(Q_n = 0) = \gamma\sqrt{a_n b_n (b_n+1)}(1 + b_n)^{a_n} \prod_{k=1}^{s_n-1} \frac{z_k^n}{z_k^n - 1}. \tag{A.5}$$

where $z_1, ..., z_{s_n-1}$ denote the zeros of $z^{s_n} - G_n^A(z)$ in $|z| < 1$. 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 Janssen and van Leeuwaarden (2005).

Lemma 4. *Define the iteration scheme*

$$z_k^{n,l+1} = w_k^n [A_n(z_k^{n,l})]^{1/s_n}, \tag{A.6}$$

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

Proof The successive substitution scheme given in (A.6) is the fixed point iteration scheme described in Janssen and van Leeuwaarden (2005), applied to the pgf of our arrival process. The authors show that, under the assumption of $A_n(z)$ being zero-free within $|z| \leq 1$, the zeros can be approximated arbitrarily closely, given that the function $[A_n(z)]^{1/s_n}$ is a contraction for $|z| \leq 1$, i.e.

$$\Big| \frac{d}{dz} [A_n(z)]^{1/s_n} \Big| < 1. \tag{A.7}$$

In our case,

$$\Big| \frac{d}{dz} [A_n(z)]^{1/s_n} \Big| = \Big| \frac{d}{dz} \left(1 + (1-z)b_n\right)^{-a_n/s_n} \Big| = \frac{a_n b_n}{s_n} \Big| 1 + (1-z)b_n \Big|^{-a_n/s_n - 1}, \tag{A.8}$$

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

$$|1 + (1-z)b_n| \geq |1 + b_n| - |z| b_n = 1 + (1 - |z|) b_n \geq 1, \tag{A.9}$$

when $|z| \leq 1$. Hence the expression in (A.8) is less than 1 for all $|z| \leq 1$. Evidently, $A_n(z)$ is also zero-free in $|z| \leq 1$. Thus (Janssen and van Leeuwaarden 2005, Lemma 3.8) shows that $z_k^{n,l}$ as in (A.6) 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 (A.6) equals $\frac{d}{dz}]A_n(z)]^{1/s_n}$ 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 $w_k^n = e^{2\pi i k/s_n}$ and $\alpha_n = a_n/s_n$. Then the zeros of $z^{s_n} - A_n(z)$ are given by*

$$z_k^n = \sum_{l=1}^{\infty} \frac{1}{l!} \frac{\Gamma[l\alpha_n + l - 1)}{\Gamma(l\alpha_n)} \frac{b_n + 1}{b_n} \Big(\frac{b_n}{(b_n+1)^{\alpha_n+1}} \Big)^l (w_k^n)^l, \tag{A.10}$$

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

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

$$z\,[A_n(z)]^{-1/s_n} = z\,(1 + (1-z)b_n)^{a_n/s_n} = w, \tag{A.11}$$

where $w = w_k = e^{2\pi i k/s_n}$. The Bürmann-Lagrange formula for $z = z(w)$, as can be found in (de Bruijn 1981, Sec. 2.2) for $z = z(w)$ is given by

$$\begin{aligned} z(w) &= \sum_{l=1}^{\infty} \frac{1}{l!} \left(\frac{d}{dz} \right)^{l-1} \left[\left(\frac{z}{z(1 + (1-z)b_n)^{a_n/s_n}} \right)^l \right]_{z=0} w^l \\ &= \sum_{l=1}^{\infty} \frac{1}{l!} \left(\frac{d}{dz} \right)^{l-1} \left[(1 + (1-z)b_n)^{-l\,a_n/s_n} \right]_{z=0} w^l. \end{aligned} \tag{A.12}$$

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

$$\left(\frac{d}{dz} \right)^{l-1} \left[(1 + (1-z)b_n)^{-l\alpha_n} \right]_{z=0} = \frac{\Gamma(l\alpha_n + l - 1)}{\Gamma(l\alpha_n)} \frac{1 + b_n}{b_n} \left(\frac{b_n}{(1+b_n)^{\alpha_n+1}} \right)^l. \tag{A.13}$$

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

$$z(w) = d_n \sum_{l=1}^{\infty} \frac{\Gamma(l\alpha_n + l - 1)}{\Gamma(l+1)\Gamma(l\alpha_n)} c_n^l\, w^l. \tag{A.14}$$

By Stirling's formula

$$\frac{\Gamma(l\alpha_n + l - 1)}{\Gamma(l+1)\Gamma(l\alpha_n)} = \frac{D}{l\sqrt{l}} \left(\frac{(\alpha_n + 1)^{\alpha_n+1}}{\alpha_n^{\alpha_n}} \right)^l, \tag{A.15}$$

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

$$\frac{(\alpha_n+1)^{\alpha_n+1}}{\alpha_n^{\alpha_n}}\,c_n = \frac{(\alpha_n+1)^{\alpha_n+1}}{\alpha_n^{\alpha_n}}\cdot\frac{b_n}{(1+b_n)^{\alpha_n+1}} = \left(\frac{b_n+\rho_n}{b_n+1}\right)^{\rho_n/b_n+1}\left(\frac{1}{\rho_n}\right)^{\rho_n/b_n}. \tag{A.16}$$

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

$$\frac{1}{r_{\mathrm{BL}}} := \left(\frac{b_n+\rho_n}{b_n+1}\right)^{\rho_n/b_n+1}\left(\frac{1}{\rho_n}\right)^{\rho_n/b_n}. \tag{A.17}$$

The derivative with respect to ρ_n of the quantity

$$\left(1+\frac{\rho_n}{b_n}\right)\ln\left(\frac{b_n+\rho_n}{b_n+1}\right)+\frac{\rho_n}{b_n}\ln\left(\frac{1}{\rho_n}\right) \tag{A.18}$$

is given by

$$\frac{1}{b_n}\ln\Big(\frac{b_n+\rho_n}{b_n\rho_n+\rho_n}\Big) > 0 \tag{A.19}$$

for $0<\rho_n<1$ and $b_n>0$. Furthermore, the quantity in (A.18) 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 (A.17) clearly shows the decremental effect of both having a large b_n and or having ρ_n close to 1. The quantities $\Gamma(l\alpha+l-1)/(\Gamma(l+1)\Gamma(l\alpha))$ 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.

□

Appendix B: Statistical procedures

To calibrate our model to real data, we now discuss some statistical procedures to show the presence of overdispersion and to estimate the parameters of the mixed Gamma-Poisson distribution. Let $x_1,...,x_n$ denote the observed number of arrivals in consecutive time slots. These observations can be interpreted as realizations of the random variables $A_1,...,A_N$, and

$$\bar{a}_N = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad \bar{s}_N^2 = \frac{1}{N-1}\sum_{i=1}(x_i-\bar{x}_i)^2, \tag{B.1}$$

the sample mean and variance with equivalent random variables $\hat{A}_n$ and S_N^2, respectively. Several tests with null hypothesis that $x_1,...,x_N$ originate from a (constant rate) Poisson distribution were discussed by Brown and Zhao (2002). We mention two of them. The first is frequently referred to as the *dispersion test*, and is based on the test statistic

$$D_N = \frac{(N-1)S_N^2}{\bar{A}_N}, \tag{B.2}$$

which is approximately chi-squared distributed with $N-1$ degrees of freedom. When using a significance level α, the critical value is equal to the $(1-\alpha)$-th quantile of chi-squared distribution

$\chi^2_{N-1,1-\alpha}$. The second test, which is also used in Jongbloed and Koole (2001), involves the test statistic

$$T_N = \sqrt{N/2}\left(\frac{S_N^2}{\bar{A}_N} - 1\right), \tag{B.3}$$

which is known as the Neyman-Scott test statistic. Under the null hypothesis T_N tends to a standard normal random variable for large N. Hence both test statistics rely on the ratio of the sample variance and sample mean, which should be 1 if $A_1, ..., A_N$ are indeed i.i.d. Poisson distributed. Excessive values of D_N and T_N therefore raise the suspicion of overdispersed arrivals.

Once either (or both) of the Poisson tests rejects the hypothesis of arrivals originating from a unicomponent Poisson process, we fit the data to the Gamma-Poisson mixture. Note that if we assume A_i to be distributed as a Poisson random variable with random rate Λ_i, which is in turn Gamma distributed with parameters a and $1/b$, then A_i is in fact a negative binomial random variable with parameters $r = a$ and $p = b/(b+1)$. Finding estimators $\hat{a}$ and $\hat{b}$ therefore is equivalent to fitting a negative binomial distribution to the data to obtain $\hat{r}$ and $\hat{p}$, followed by retrieving $\hat{a} = \hat{r}$ and $\hat{b} = \hat{p}/(1-\hat{p})$. We proceed by applying the maximum likelihood estimation method described in Jongbloed and Koole (2001) to find $\hat{r}$ and $\hat{p}$. This method prescribes to set $\hat{r}$ to be the value of r for which the *profile loglikelihood function* defined by

$$L(r) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{a_i}\ln(r+j+1) + r\ln r - (r+\bar{a}_N)\ln(r+\bar{a}_N), \tag{B.4}$$

is attained. Subsequently, $\hat{p} = \hat{r}/(\hat{r}+\bar{a}_N)$, so that $\hat{a} = \hat{r}$ and $\hat{b} = \hat{r}/\bar{a}_N$.

Finally, given the estimators $\hat{a}$ and $\hat{b}$, we need statistical evidence that the obtained Poisson mixture indeed fits the data reasonably well. Here we again cite on Jongbloed and Koole (2001), who give a method to retrieve the p-value for the goodness-of-fit based on bootstrap and Monte-Carlo simulation. In our procedures we work with 1.000.000 replications of the Monte-Carlo simulation to obtain the approximated p-value. We refer to the appendix of Jongbloed and Koole (2001) for further details on this method.