Queueing Systems (2022) 101:113–135
https://doi.org/10.1007/s11134-021-09728-5

Heavy traffic limits for queues with non-stationary path-dependent arrival processes

Kerry Fendick[1] · **Ward Whitt**[2]

Received: 29 March 2021 / Revised: 5 December 2021 / Accepted: 13 December 2021 /
Published online: 15 March 2022

Abstract
In this paper, we develop a diffusion approximation for the transient distribution of the workload process in a standard single-server queue with a non-stationary Polya arrival process, which is a path-dependent Markov point process. The path-dependent arrival process model is useful because it has the arrival rate depending on the history of the arrival process, thus capturing a self-reinforcing property that one might expect in some applications. The workload approximation is based on heavy-traffic limits for (i) a sequence of Polya processes, in which the limit is a Gaussian–Markov process, and (ii) a sequence of *P/GI/1* queues in which the arrival rate function approaches a constant service rate uniformly over compact intervals.

Keywords Path-dependent stochastic processes · Generalized Polya process · Gaussian Markov process · Diffusion approximations · Queues · Heavy-traffic limit

Mathematics Subject Classification Primary 60K25 · Secondary 60F17 · 90B22

1 Introduction

In this paper, we establish a heavy traffic limit theorem (HTLT) for the standard single-server queue with an exogenous arrival process that is a generalized Polya process

✉ Ward Whitt
ww2040@columbia.edu

Kerry Fendick
Kerry.Fendick@jhuapl.edu

[1] Communications System Branch, Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723, USA

[2] Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699, USA

(GPP), which we refer to as a *P/GI/1* queue. We then apply that limit to develop a diffusion approximation for the transient distribution of the workload process.

As in Konno [13] and Cha [6], a GPP $N \equiv \{N(t) : t \geq 0\}$ is a Markov point process with stochastic intensity (defined in terms of the internal histories $\mathcal{H}_t$) by

$$\lambda^*(t|\mathcal{H}_t) \equiv (\gamma N(t-) + \beta)\kappa(t), \tag{1.1}$$

where $N(0) = 0$, γ and β are positive constants, $\kappa(t)$ is a positive integrable real-valued function, and $\equiv$ denotes equality by definition. The GPP is interesting and important because it is *path-dependent*, i.e., it fails to have the *asymptotic loss of memory* (ALOM) property, i.e., the influence of early conditions fails to dissipate over time.

This paper extends our paper [9], which established results for the special case of a stationary GPP. Theorem 1 of [9] shows that a GPP is a stationary point process (has stationary increments) if

$$\kappa(t) = \frac{1}{\gamma t + 1}, \ t \geq 0. \tag{1.2}$$

As discussed in [9], a stationary GPP satisfies a non-ergodic law of large numbers, which causes the queue length process (not normalized by time) to approach infinity with positive probability as time evolves. (For the stationary GPP, we defined path dependence by lack of ergodicity; for the more general GPP, we define path-dependence by the lack of ALOM.)

The GPP is a self-exciting point process like the Hawkes process. Infinite-server queues with Hawkes arrival processes have been studied by Koops et al. [14] and references there. As discussed in [14], self-exciting point processes are interesting to capture overdispersion found in arrival process data, for example [12]. However, as discussed in Remark 3 of [9], the GPP is quite different from the Hawkes process; for example, unlike the GPP, the Hawkes process is always stationary and ergodic and has the ALOM property. The GPP can capture the long-term impact of initial conditions. Very broadly, we are motivated by a non-stationary worldview brought on by climate change and the pandemic in which early conditions may strongly influence long-term outcomes.

In our opinion, the greatest appeal of the heavy-traffic limit for the *P/GI/1* queue with a stationary GPP arrival process in [9] is its remarkable tractability, as illustrated by the explicit three-dimensional distribution of the limit process in Corollary 6 of [9]. That tractability suggests that the GPP can be useful for modelling in practical applications. Our goal in this paper was to see how much of this appealing tractability we could achieve without requiring the stationarity. We succeed in generalizing both the HTLT and a result obtained in Corollary 6 of [9] describing the transient distribution of the queue's limit process. We also show that a general GPP can be approximated arbitrarily well by a piecewise-stationary GPP, and we use that structure to obtain an explicit transient approximation formula particularly convenient for modeling.

In this paper, we continue to study the much larger class of non-stationary GPPs. Theorem 2 of [9] establishes that a GPP $\widehat{N}$ with parameter triple $(\widehat{\kappa}(t), \gamma, \beta)$ can

be represented as a deterministic time-transformation of a stationary GPP N with parameter triple $(\kappa(t),\ \gamma,\ \beta)$, where κ is of the form in (1.2). That property facilitates applying the composition map and the continuous-mapping theorem to establish the more general heavy-traffic limit, but more is required to obtain useful results.

First, we need to specify a heavy-traffic regime. Our approach is to make the system critically loaded over an initial time interval of interest. To achieve that, we let the instantaneous traffic intensity approach 1 uniformly over such intervals. But that in turn depends on a definition of the arrival rate. For that, Theorem 1 here determines the arrival rate $\lambda(t)$ (which is defined like $\lambda^*(t|\mathcal{H}_t)$ above, but is not conditioned on the history) and shows that a GPP exists with any integrable rate function. As a consequence, Corollary 2 here shows that the parameter triples $(\lambda(t),\ \gamma,\ \beta)$ and $(\kappa(t),\ \gamma,\ \beta)$ are homeomorphic representations of a GPP. Going forward, we use the representation $(\lambda(t),\ \gamma,\ \beta)$, which directly specifies the rate.

Because of the non-stationarity, the diffusion approximation for the transient distribution depends on the entire rate function $\lambda(t)$. In the first paper, we exploit full stationarity, which corresponds to the case where $\lambda(t)$ is constant. In this paper, we allow the rate function to be a general (positive) function. In order to obtain a transient distribution that depends on a finite number of parameters, we apply Theorem 1 in Corollary 3 to characterize a piecewise-stationary GPP, which we refer to as a ψ^k − GPP (when there are k pieces). The rate function for a ψ^k − GPP is constant on each piece, and it follows that a ψ^k − GPP has $k+2$ parameters. Corollary 4 then shows any GPP can be represented as a limit of ψ^k − GPPs. We envision useful applications having relatively small k.

To achieve the desired critical loading here, we consider a sequence of GPPs indexed by n with parameter triples $(\lambda^n(t),\ \gamma^n,\ \beta^n) = (n\zeta^n(t),\ \gamma,\ nb)$. We then specify the scaling by

$$n^{1/2}\big(\zeta^n - bu\big) \to \overline{\eta} \text{ in } D \text{ and } n^{1/2}\big(\mu^n - b\big) \to \overline{\mu} \text{ in} \mathcal{R}, \tag{1.3}$$

where the M_1 topology is used on the function space D, $u(t) \equiv 1$ is the unit function, and the μ^n are scalar service rates. (The M_1 topology is used throughout rather than a more conventional topology because we either require discontinuous limits approached by continuous functions, for example in Corollaries 2 and 3, or we wish to allow them, as we do in (1.3).) The limit function $\overline{\eta}$ and scalar $\overline{\mu}$ define the mean of the limit process for the queue's net input process. In particular, $\overline{\eta}$ reflects its non-stationary nature. As a consequence of (1.3), the queue's instantaneous traffic intensity function $\rho^n(t)$, defined as

$$\rho^n(t) \equiv \lambda^n(t)/\big(n\mu^n\big) = \zeta^n(t)/\mu^n, \tag{1.4}$$

will approach one uniformly over compact intervals (u.o.c.) as $n \to \infty$. With that scaling, we establish a functional central limit theorem (FCLT) for the arrival process and the desired HTLT for the workload process in the *P/GI/1* queue. However, the limit processes both depend on the limit function $\overline{\eta}$ in (1.3). In order to obtain more useful approximations, in Sect. 3.3 we apply the limit theorems to develop an asymptotic approximation that depends instead on the rate function and the other parameters of

the original *P/GI/1* model. To the best of our knowledge, this approach has never been proposed before. In fact, we think that this approximation approach is new even for an M_t arrival process, i.e., for a non-homogeneous Poisson process (NHPP). In that regard, we mention that results for the NHPP arrival process are covered as a special case here by just setting $\gamma = 0$ (which requires minor modifications in some definitions to avoid dividing by 0).

Since the GPP is an arrival process with a time-varying rate, the limits here are related to previous ones for time-varying single-server queues; see [21], especially Sect. 7.2 for a review. The papers [15] and [22] are relevant. Particularly relevant are the HTLTs for the $M_t/M_t/1$ model in [15] obtained from the strong approximation. The scaling in (1.3) here and the approximation scheme in Sect. 3.3 here are different than used previously.

This paper is organized as follows: In Sect. 2, after reviewing GPPs, we establish Theorem 1, providing the new representation of a GPP in terms of the rate function, and show how to construct piecewise-stationary GPPs. In Sect. 3, we establish limit theorems for the arrival and workload processes and asymptotically correct approximations for those processes. Theorem 2 is the FCLT for the arrival process, while Theorem 3 is the HTLT for the workload process. The new asymptotic approximation is then developed in Sect. 3.3. Finally, Theorem 4 in Sect. 3.4 establishes the approximating transient workload distribution. In Sect. 4, we prove two lemmas used in the proof of Theorem 4. Finally, we provide some concluding discussion in Sect. 5.

2 A piecewise-stationary Generalized Polya Process (ψ^k − GPP)

In this section, we briefly review GPPs, as developed in [6, 9, 13]. Then, we show how to construct a piecewise-stationary GPP with k pieces, which we refer to as a ψ^k − GPP. We then show that, under regularity conditions, a general GPP can be approximated by a ψ^k − GPP for suitably large k.

A GPP with parameter triple $(\kappa(t), \gamma, \beta)$ is defined in [6] as the orderly point process $\{N(t) : t \geq 0\}$ with $N(0) = 0$ and stochastic intensity function

$$\begin{aligned}\lambda^*(t|\mathcal{H}_t) &\equiv \lim_{h\to 0} \frac{P(N(t+h) - N(t) = 1|\mathcal{H}_t)}{h} = \lim_{h\to 0} \frac{E[N(t+h) - N(t)|\mathcal{H}_t]}{h} \\ &= (\gamma N(t-) + \beta)\kappa(t), \end{aligned} \tag{2.1}$$

where $\mathcal{H}_t$ denotes the internal history of N up to time t, $\kappa(t)$ is a positive-integrable real-valued function, while β and γ are positive real numbers. For background on point processes and their intensity functions, see Sect. 3.3 and 7.2 of [7].

A GPP can be a stationary point process (meaning that it possesses stationary increments) although GPPs are not in general stationary processes.

Proposition 1 (*Theorem 1 of* [9]) *The GPP* $\tilde{N}$ *with parameter triple* $(\kappa(t), \gamma, \beta)$, *where*

$$\kappa(t) = 1/(\gamma t + 1)$$

as in (1.2) is a stationary point process with mean and covariance functions

$$E\left[\tilde{N}(t)\right] = \beta t \text{ and } Cov\left[\tilde{N}(s),\, \tilde{N}(t)\right] = \beta s(1+\gamma t) \text{ for } 0 \le s \le t. \tag{2.2}$$

Sketch of proof. By Theorem 1 of [6], if N is a GPP with parameter triple $(\kappa(t),\, \gamma,\, \beta)$ and

$$\mathrm{K}(t) = \int_0^t \kappa(s)ds, \tag{2.3}$$

then $N(t)$ has a negative binomial distribution with mean $E[N(t)] = \tau p(t)/(1-p(t))$ and $Var[N(t)] = \tau p(t)/(1-p(t))^2$, where $\tau \equiv \gamma/\beta$ and $p(t) = 1 - \exp(-\gamma \mathrm{K}(t))$. If (1.2) holds, then $\mathrm{K}(t) = \gamma^{-1}\log(\gamma t + 1)$, and $1 - p(t) = \exp(-\gamma \mathrm{K}(t)) = \kappa(t)$. The mean and variance of $\tilde{N}(t)$ easily follow. The proof of the stationarity of $\tilde{N}$ from Theorem 1 of [9] uses the property from Theorem 3 and Remark 3 of [6] that the times of increase of a GPP on the interval $[s,\, t]$, conditioned on $N(t) - N(s) = n$, have the distribution of the order statistics of n i.i.d random variables. When (1.2) holds, those random variables are uniformly distributed. The covariance function in (2.2) follows easily from the mean and variance functions and the stationarity of $\tilde{N}$. □

The GPP $\tilde{N}$ from Proposition 1 is called a stationary-increment GPP, or a $\psi -$ GPP for short, and is specified by the parameter pair $(\gamma,\, \beta)$. The classical Polya process defined on page 435 of [8] is the $\psi -$ GPP with parameter pair $(\gamma,\, 1)$.

In this paper, we will make strong use of Theorem 2 of [9], which shows that a general GPP can be represented as a deterministic time transformation of a $\psi -$ GPP. We thus restate it here as Proposition 2. For that purpose, let $\stackrel{d}{=}$ denote equality in distribution for stochastic processes.

Proposition 2 *(Theorem 2* of [9]*) Let N be a GPP with parameter triple $(\kappa(t),\, \gamma,\, \beta)$ and $\tilde{N}$ be the $\psi -$* GPP *with parameter pair $(\gamma,\, \beta)$. Then,*

$$\{N(t) : t \ge 0\} \stackrel{d}{=} \left\{\tilde{N}(M(t)) : t \ge 0\right\} \tag{2.4}$$

if and only if

$$M(t) = \gamma^{-1}\left(e^{\gamma K(t)} - 1\right) \text{ for } t \ge 0, \tag{2.5}$$

where $K(t)$ is defined in (2.3).

We can apply Proposition 2 to derive properties of non-stationary GPPs from those of a corresponding $\psi -$ GPP, as illustrated by the following corollary.

Corollary 1 *If N is a GPP with parameter triple $(\kappa(t),\, \gamma,\, \beta)$, then $E[N(t)] = \beta\gamma^{-1}(exp(\gamma \mathrm{K}(t)) - 1)$ and $Cov[N(s),\, N(t)] = \beta\gamma^{-1}exp(\gamma \mathrm{K}(t))(exp(\gamma \mathrm{K}(s)) - 1)$ for $0 \le s \le t$.*

Proof If $\tilde{N}$ is the ψ − GPP with parameter pair (γ, β), then $E[\tilde{N}(t)] = \beta t$ and $Cov[\tilde{N}(s)\tilde{N}(t)] = \beta s(1 + \gamma t)$ by Proposition 1. We then obtain the result by applying Proposition 2. □

It will be helpful to express results for a GPP in terms of its instantaneous mean function $\lambda(t)$ and its mean function $\Lambda(t)$ which we define and characterize next. The instantaneous mean function $\lambda(t)$ is defined as $\lambda(t) \equiv \lim_{h\to 0} E[N(t + h) - N(t)]/h$. The instantaneous mean function $\lambda(t)$ differs from the stochastic intensity function $\lambda^*(t|\mathcal{H}_t)$ in (2.1) by not conditioning on the history. The instantaneous mean function is also known as the arrival rate function. We will sometimes refer to it as the rate or the mean rate; for instance, see Remark 2.

We show that, for given parameter pair (γ, β), the instantaneous mean function $\lambda(t)$ is a one-to-one function of the parameter function $\kappa(t)$.

Theorem 1 (instantaneous mean and mean functions) *If N is a GPP with parameter triple $(\kappa(t), \gamma, \beta)$, then the instantaneous mean function can be expressed as:*

$$\lambda(t) \equiv \lim_{h\to 0} E[N(t + h) - N(t)]/h = \beta\kappa(t)exp(\gamma \mathrm{K}(t)), \tag{2.6}$$

for $\mathrm{K}(t)$ *in* (2.3)*, so that $\lambda(t)$ is integrable. As a consequence, the associated mean function is*

$$\Lambda(t) \equiv \mathrm{E}[N(t)] = \beta M(\mathrm{t}) = \int_0^t \lambda(v)dv = \frac{\beta exp(\gamma \mathrm{K}(t)) - \beta}{\gamma}, \tag{2.7}$$

where $M(\mathrm{t})$ *is defined in (2.5), and*

$$\mathrm{E}[N(s + t) - N(s)] = \Lambda(t + s) - \Lambda(s) = \int_s^{s+t} \lambda(v)dv, \tag{2.8}$$

$$\mathrm{Cov}[N(s + t) - N(s),\, N(s + u) - N(s)] = \left(\int_s^{s+t} \lambda(v)dv\right)\left(1 + \tau \int_s^{s+u} \lambda(v)dv\right), \tag{2.9}$$

$$\kappa(t) = \lambda(t)/(\beta + \gamma\Lambda(t)), \tag{2.10}$$

and

$$\lambda^*(t|\mathcal{H}_t) \equiv \frac{\tau N(t-) + 1}{\tau\Lambda(s) + 1}\lambda(t) \tag{2.11}$$

for $s \geq 0$ and $0 \leq t \leq u$, where $\tau \equiv \gamma/\beta$.

Proof The result in (2.7) follows from Corollary 1. The results in (2.6) and (2.8) follow from (2.7). By Corollary 1 and (2.7), $\Gamma(t, u) \equiv \mathrm{Cov}[N(t), N(u)] = \left(\int_0^t \lambda(v)dv\right)\left(1 + \tau \int_0^u \lambda(v)dv\right)$ for $0 \leq t \leq u$. The result in (2.9) is obtained for $s \geq 0$ through the identity $\mathrm{Cov}[N(s + t) - N(s), N(s + u) - N(s)] = \Gamma(s + t, s + u) - \Gamma(s, s + u) - \Gamma(s, s + t) + \Gamma(s, s)$. By (2.6) and (2.7), the result in (2.10) holds, and (2.11) follows using (2.1). □

We will consider limits of GPPs. For this purpose, we will exploit the function space D of all right-continuous real-valued functions on the semi-infinite interval $[0, \infty)$ with limits from the left, endowed with one of the Skorohod topologies, as in Sects. 3.3 and 11.5 and Chapter 12 of [20]. These topologies reduce to uniform convergence over compact sets (u.o.c.) when the limit function is continuous. In order to allow for continuous functions converging to discontinuous limits, we use the Skorohod M_1 topology. Convergence in D under the M_1 metric is implied by u.o.c. convergence. The use of M_1 to denote a metric should not be confused with the use of M in (2.5) to denote the time-transformation function. Throughout, $\Rightarrow$ will denote weak convergence of a sequence of random elements of a given topological space. Let $\left(D^k, M_1\right)$ be the product space with the product topology. (It is used in the proofs of Theorem 2 and Corollary 5.)

Corollary 2 The function $\kappa(t)$ *is an element of* (D, M_1) if and only if $\lambda(t)$ *is an element of* (D, M_1), *and* $(\kappa(t), \gamma, \beta)$ *and* $(\lambda(t), \gamma, \beta)$ *constitute homeomorphic representations of a GPP.*

Proof The one-to-one relationship is established by Theorem 1. The continuity map from $\kappa(t)$ to $\lambda(t)$ and its inverse follow from their explicit representations in (2.6) and (2.10), because converge of functions in (D, M_1) implies convergence of their integrals; see [17] for background. □

Theorem 1 implies one-to-one relationships between $(\kappa(t), \gamma, \beta)$, $(\lambda(t), \gamma, \beta)$, $(\mathrm{K}(t), \gamma, \beta)$, $(\Theta(t), \gamma, \beta)$, and $(\mathrm{M}(t), \gamma, \beta)$. Convergence of $(\kappa(t), \gamma, \beta)$ or $(\lambda(t), \gamma, \beta)$ implies convergence of any of the others, but the converse is not true because convergence of functions does not imply convergence of their derivatives. Therefore, Corollary 2 describes the only homeomorphic representation of a GPP from those among those one-to-one relationships.

Remark 1 (*instantaneous mean representation of a GPP*). We will use the $(\lambda(t), \gamma, \beta)$ representation of a GPP for results that follow. In that representation, the first element is the GPP's instantaneous mean, and the second and third elements are always positive, just as for the $(\kappa(t), \gamma, \beta)$ representation. As an example, GPPs with parameter triples $(\lambda(t), \gamma, \beta)$ and $(\lambda(t), \gamma, n\beta)$ have the same instantaneous mean.

We now apply Theorem 1 to characterize a $\psi^k -$ GPP, a piecewise-stationary GPP with k pieces.

Corollary 3 (characterization of a $\psi^k -$ GPP). If N *is a GPP with parameter triple* $(\lambda(t), \gamma, \beta)$, *where*

$$\lambda(t) = \lambda_i u(t) \text{ for } t_{i-1} \le t < t_i \text{ and } 1 \le i \le k \le \infty \tag{2.12}$$

for real $\lambda_i > 0$, $u(t) \equiv 1$, and $t_0 \equiv 0$, *then*

$$\Lambda(t) \equiv \mathrm{E}[N(t)] = \beta M(\mathrm{t}) = \sum_{j=1}^{i-1} \lambda_j \left(t_j - t_{j-1}\right) + \lambda_i (t - t_{i-1}) \tag{2.13}$$

for $t_{i-1} \le t < t_i$ and $1 \le i \le k \le \infty$, *so that the instantaneous mean* $\lambda(t)$ *is piecewise constant and the mean* $\Lambda(t)$ *is continuous and piecewise linear,*

$$\mathrm{E}[N(s+t) - N(s)] = \lambda_i t \tag{2.14}$$

and

$$Cov[N(s+t) - N(s),\, N(s+u) - N(s)] = \lambda_i t(1 + \tau\lambda_i \mathrm{u}) \tag{2.15}$$

for $t_{i-1} \le s < t_i$ and $0 \le t \le u < t_i - s$, *where* $\tau = \gamma/\beta$. *Furthermore,* N *is stationary on* $t_{i-1} \le t < t_i$ *for each* $i \ge 1$.

Proof The expressions in (2.13)-(2.15) are special cases of the results in Theorem 1. By Theorem 1 and (2.13), a $\psi^k -$ GPP can be represented as a piecewise linear time transformation of a $\psi -$GPP, in which time is scaled by a constant on each piece. The stationarity of the $\psi -$GPP on each piece is then preserved by the time transformation, so that a $\psi^k -$ GPP is piecewise stationary. □

Remark 2 (*GPPs with constant or piecewise-constant rates*). As a consequence of Corollary 3, a GPP has a piecewise-constant instantaneous mean function $\lambda(t)$ if and only if it is a $\psi^k -$ GPP. In particular, a GPP has a constant rate c if and only if it is a $\psi -$GPP with parameter triple $(\lambda(t) = cu(t),\, \gamma,\, \beta)$. The $\psi -$GPP $\tilde{N}$ with parameter pair $(\gamma,\, \beta)$ defined in terms of $\kappa(t)$ in (1.2) arises as the special case when $c = \beta$.

We will consider a sequence of GPPs indexed by n with parameter triples $(\lambda^n(t),\, \gamma^n,\, \beta^n)$, where $\lambda^n(t)$ will denote the instantaneous mean function of the n^{th} GPP in the sequence. We will then define the mean function $\Lambda^n(t)$ and time-transformation function $M^n(t)$ to be

$$\Lambda^n(t) \equiv \int_0^t \lambda^n(s)ds = \beta^n M^n(t), \tag{2.16}$$

consistently with the definitions in Theorem 1.

Proposition 3 (continuity for GPPs) *If* $\widehat{N}^n$ *is a GPP with parameter triple* $(\lambda^n(t),\, \gamma,\, \beta)$ *for* $n \ge 1$, *where* λ^n *is in* D, *and* $\lambda^n \to \lambda > 0$ *in* $(D,\, M_1)$ *as* $n \to \infty$, *then* $\widehat{N}^n \Rightarrow N$ *in* $(D,\, M_1)$, *where* N *is a GPP with parameter triple* $(\lambda(t),\, \gamma,\, \beta)$.

Proof Under the assumptions, $M^n = \beta^{-1}\Lambda^n \to \beta^{-1}\Lambda = M$ in $(D,\, M_1)$, where $\Lambda^n(t)$ and $M^n(t)$ are defined in (2.16) and $\Lambda(t)$ and $M(t)$ are defined in (2.7). Applying Proposition 2 twice,

$$\widehat{N}^n \overset{d}{=} \tilde{N} \circ M^n \Rightarrow \tilde{N} \circ M \overset{d}{=} N \text{ in } (D,\, M_1),$$

where the weak convergence step follows from continuity of the composition map by applying Theorem 13.2.3 of [20], which uses the fact that M is continuous and strictly increasing. □

Corollary 4 *If N is a GPP with parameter triple $(\lambda(t), \gamma, \beta)$ where λ is in D, then there exists a sequence $\widehat{N}^n$ of $\psi^k -$ GPPs such that $\widehat{N}^n \Rightarrow N$ in (D, M_1) as $n \to \infty$.*

Proof The limit follows from Proposition 3 and Theorem 12.2.2 of [20], which states that any function in D can be represented as the u.o.c. convergence of a sequence of piecewise-constant functions. At this point, the M_1 topology is used only to ensure that the space D is endowed with the usual Kolmogorov $\sigma-$ field; see Sect. 11.5.3 of [20] for further discussion. We can obtain u.o.c. convergence because we can choose the discontinuity points of the converging function to match those of the limit function. □

3 The P/GI/1 workload in heavy traffic

Our purpose now is to obtain a HTLT for a sequence of *P/GI/1* queues as the associated sequence of instantaneous time-dependent traffic intensities approaches one u.o.c. and to apply that limit to develop tractable approximations. In Sect. 3.1, we derive a FCLT for the arrival processes. In Sect. 3.2, we define the net input and workload processes for a *P/GI/1* queue and derive an HTLT describing them. In Sect. 3.3, we apply the HTLT to develop asymptotic approximations for the net input and workload processes as functions of their parameters. Finally, in Sect. 3.4 we provide a tractable approximation for the transient distribution of the workload process.

3.1 The functional central limit theorem for the arrival process

We first state a FCLT for a sequence of $\psi -$ GPPs approaching a zero-mean Gaussian Markov process $\overline{N}$ with stationary increments, referred to as an $\psi -$ GMP in [9, 10].

Because the limit process $\overline{N}$ is a $\psi -$ GMP, it is a zero-mean Gaussian process, so that its distribution (as a process, i.e., its finite-dimensional distributions) is determined by its covariance function. As shown in [14], if A is a $\psi -$ GMP with parameter pair $(\alpha^* > 0, \beta^* \leq 0)$, then

$$Cov[A(s), A(t)] = s\left(\alpha^* - \beta^* t\right) \text{ for } 0 \leq s \leq t. \tag{3.1}$$

A $\psi -$ GMP is continuous with probability one. We will apply the following FCLT for $\psi -$ GPPs from [9] together with Proposition 2 to obtain a FCLT for non-stationary GPPs.

Proposition 4 (FCLT for $\psi -$ GPPs from [3]) *If $\tilde{N}_n(t) \equiv n^{-1/2}(\tilde{N}^n(t) - nbt)$ for $n \geq 1$, where $\tilde{N}^n$ is a $\psi -$GPP with parameter pair (γ, nb), then $\tilde{N}_n \Rightarrow \overline{N}$ in (D, M_1) as $n \to \infty$, where $\overline{N}$ is the $\psi -$ GMP with parameter pair $(\alpha^*, \beta^*) = (b, -b\gamma)$.*

Proof By Proposition 3 of [9], $\tilde{N}^n$ has the same distribution as the superposition of n i.i.d $\psi -$ GPPs each with parameter pair (γ, b). The result is then implied by Theorem 4 of [9], since u.o.c. convergence there to a continuous limit is equivalent to M_1 convergence. □

We now establish convergence of a sequence of non-stationary GPPs.

Theorem 2 (convergence to a ψ − GMP with time-dependent drift) *If*

$$N_n(t) \equiv n^{-1/2}\left(N^n(t) - nbt\right) \text{ for } n \geq 1, \tag{3.2}$$

where N^n *is a GPP with parameter triple* $(\lambda^n(t), \gamma^n, \beta^n) = (n\zeta^n(t), \gamma, nb)$ *for* $\zeta^n > 0$ *a deterministic element of* D *and* $b > 0$, *and*

$$n^{1/2}\left(\zeta^n - bu\right) \to \overline{\eta} \text{ in } (D, M_1) \text{ as } n \to \infty \tag{3.3}$$

(where $\overline{\eta}$ *need not be continuous), then*

$$N_n \Rightarrow \overline{N} + \overline{\nu} \text{ in } (D, M_1), \tag{3.4}$$

where $\overline{N}$ *is the* ψ − GMP *with parameter pair* $(\alpha^*, \beta^*) = (b, -b\gamma)$ *and* $\overline{\nu}(t) = \int_0^t \overline{\eta}(s)ds$.

Proof Using (2.16),

$$M^n(\mathrm{t}) = \frac{1}{\beta}\int_0^t \lambda^n(v)dv = \frac{1}{b}\int_0^t \zeta^n(v)dv \equiv \frac{1}{b}\mathrm{Z}^n(t). \tag{3.5}$$

Then, (3.3) implies that

$$n^{1/2}\left(\mathrm{Z}^n - be\right) \to \overline{\nu} \text{ in } (D, M_1) \text{ as } n \to \infty, \tag{3.6}$$

so that

$$M_n \equiv n^{1/2}\left(M^n - e\right) \to b^{-1}\overline{\nu} \text{ in } (D, M_1) \text{ as } n \to \infty. \tag{3.7}$$

By Proposition 2 and the definitions from Proposition 4,

$$N_n(t) \equiv n^{-1/2}\left(N^n(t) - nbt\right) \overset{d}{=} n^{-1/2}\left(\tilde{N}^n\left(M^n(t)\right) - nbt\right) = \tilde{N}_n\left(M^n(t)\right) + bM_n(t)$$

Then, $(M^n, M_n) \to \left(e, b^{-1}\overline{\nu}\right)$ in $\left(D^2, M_1\right)$ by (3.7). Applying Theorem 11.4.5 of [20], $\left(\tilde{N}_n, M^n, M_n\right) \to \left(\overline{N}, e, b^{-1}\overline{\nu}\right)$ in $\left(D^3, M_1\right)$. The limit preservation in Theorem 13.3.1 of [20] then yields

$$N_n \overset{d}{=} \left(\tilde{N}_n \circ M^n + bM_n\right) \Rightarrow \overline{N} + \overline{\nu} \text{ in } (D, M_1) \text{ as } n \to \infty. \qquad \square$$

Remark 3 For an example of a sequence satisfying (3.3), let $\zeta^n = bu + n^{-1/2}\overline{\eta}$ for any $\overline{\eta}$ in D.

3.2 Heavy traffic limit theorem for the queue

We apply Theorem 2 to develop the HTLT for a sequence of *P/GI/1* models, where the arrival process for each model n is the GPP N^n defined in Proposition 5. Let $\{V_k : k \geq 1\}$ be the sequence of service requirements of successive arrivals, which we assume for each of the models. There are two key assumptions. The first is that the service requirements are independent of the arrival processes. (That conditions could be replaced by joint convergence.) The second key assumption is that the associated sequence of partial sums satisfies a FCLT. In particular, let

$$S_n(t) = n^{-1/2}\left(\sum_{k=1}^{\lfloor nt \rfloor} V_k - nt\right), \; t \geq 0 \tag{3.8}$$

Our key assumption is that

$$S_n \Rightarrow c_s B \text{ in } (D, M_1) \text{ as } n \to \infty, \tag{3.9}$$

where B is standard (0 drift, unit variance) Brownian motion. This is the classical Donsker's theorem in Sect. 4.3 of [20].

A sufficient condition for (3.9) is for the sequence $\{V_k : k \geq 1\}$ to be i.i.d. with $E[V_k] = 1$ and $Var[V_k] = c_s^2$. That puts us in the setting of the P/GI/1 queue, but the i.i.d. assumption can be relaxed, as illustrated by Sect. 4.4 of [20].

Then, let

$$T^n(t) \equiv \sum_{k=1}^{N^n(t)} V_k \tag{3.10}$$

be the total input process over the interval $[0, t]$ for model n. It represents the total service requirements of all arrivals in N^n over the interval $[0, t]$. In this context, the net input process is

$$X^n(t) \equiv T^n(t) - n\mu^n t, \; t \geq 0, \tag{3.11}$$

where μ^n is the constant deterministic rate that service is performed when there is work waiting to be served. The corresponding workload process is then defined as the reflection of the net input process, i.e.,

$$W^n \equiv \phi\big(X^n, W^n(0)\big), \tag{3.12}$$

where $\phi : D \times R \to D$ is the reflection map, defined by

$$\phi(x)(t, w(0)) = w(0) + x(t) - \inf_{0 \leq s \leq t} \{\min\{w(0) + x(s), 0\}\} \text{ for } t \geq 0. \tag{3.13}$$

The reflection map describes the workload (or service backlog) for a single-server queue with an infinite buffer. For additional properties of the reflection map, see Sect. 2 of Chapter 2 on pages 19–21 of [11].

We can obtain an FCLT for $T^n(t)$ because it is a random sum, as discussed in Sect. 7.4 of [20], or more generally in Sect. 13.3 of [20] (as needed here, because we will apply Proposition 4, which has the ψ − GMP limit $\overline{N}$ instead of a Brownian motion limit). We can then use the FCLT for $T^n(t)$ to obtain limits $X^n(t)$ and $W^n(t)$. In particular, let

$$N_n(t) \equiv n^{-1/2}\big(N^n(t) - nbt\big),\ T_n(t) \equiv n^{-1/2}\big(T^n(t) - nbt\big), \tag{3.14}$$

$$X_n(t) \equiv n^{-1/2}X^n(t),\ \text{and}\ W_n(t) \equiv n^{-1/2}W^n(t). \tag{3.15}$$

Theorem 3 (HTLT for a P/GI/1 queue with non-stationary arrival process) *If* (N_n, T_n, X_n, W_n) *is defined by (3.14)–(3.15), where the definitions in (3.10)-(3.12) apply,* N^n *is a GPP with parameter triple* $(\lambda^n(t), \gamma^n, \beta^n) = (n\zeta^n(t), \gamma, nb)$ *for* $\zeta^n > 0$ *in D, and*

$$\begin{aligned} &n^{1/2}\big(\zeta^n - bu\big) \to \overline{\eta}\ \text{in}\ (D, M_1),\ n^{1/2}\big(\mu^n - b\big) \to \overline{\mu}\ \text{in}\ \mathcal{R}, \\ &\quad \text{and}\ W_n(0) \to \overline{W}(0)\ \text{in}\ \mathcal{R}^+\ \text{as}\ n \to \infty, \end{aligned} \tag{3.16}$$

then $(N_n, T_n, X_n, W_n) \Rightarrow \big(\overline{N} + \overline{\nu}, \overline{T} + \overline{\nu}, \overline{X}, \overline{W}\big)$ *in* $\big(D^4, M_1\big)$, *where* $\overline{N}$ *is the* ψ − GMP *with parameter pair* $(\alpha^*, \beta^*) = (b, -b\gamma)$, $\overline{T}$ *is the* ψ − GMP *with parameter pair* $(\alpha^*, \beta^*) = \big(b + bc_s^2, -b\gamma\big)$, $\overline{\nu}(t) = \int_0^t \overline{\eta}(s)ds$, $\overline{X} \equiv \overline{\nu} - \overline{\mu}e + \overline{T}$, *and* $\overline{W} \equiv \phi\big(\overline{X}, \overline{W}(0)\big)$.

Proof Let $S^n(t) = \sum_{k=1}^{\lfloor nt \rfloor} V_k$, so that $S_n(t) = n^{-1/2}(S^n(t) - nt)$ by (3.8). Corollary 13.3.2 of [20] plus Theorem 2 then imply that

$$\begin{aligned} &T_n = n^{-1/2}\big(S^n \circ \big(n^{-1}N^n\big) - nbe\big) = S_n \circ \big(n^{-1}N^n\big) + N_n = S_n \circ \big(n^{-1/2}N_n + be\big) + N_n \\ &\quad \Rightarrow c_s B \circ (be) + \overline{N} + \overline{\nu} = \sqrt{b}c_s B + \overline{N} + \overline{\nu} \equiv \overline{T} + \overline{\nu}\ \text{in}\ (D, M_1). \end{aligned} \tag{3.17}$$

Using (3.16) and (3.17),

$$\begin{aligned} X_n &\equiv n^{-1/2}X^n = n^{-1/2}\big(T^n - n\mu^n e\big) = n^{-1/2}\big(T^n - nbe\big) - n^{-1/2}\big(n\mu^n e - nbe\big) \\ &= T_n - n^{1/2}\big(\mu^n - b\big)e \Rightarrow \overline{T} + \overline{\nu} - \overline{\mu}e\ \text{in}\ (D, M_1). \end{aligned} \tag{3.18}$$

The conclusion about joint convergence then follows by the continuous mapping theorem. □

Remark 4 (Double sequences) It might be more natural to assume that there is a double sequence of service requirements, i.e., that there is a sequence $\big\{V_k^n : k \geq 1\big\}$ of service requirements of successive arrivals in model n for each $n \geq 1$. We would then need a generalization of Donsker's theorem in Sect. 4.3 of [20] to double sequences

or triangular arrays, because we have a sequence k for each n. An early statement of the direct extension of Donsker's theorem to double sequences or triangular arrays appears on p. 220 of [18]. The extension is also discussed in Sect. 2.4. of the Internet Supplement to [20]. It requires an additional regularity condition. It would be natural to require that V_k^n have uniformly bounded third moments. Under appropriate assumptions, the same conclusions from Theorem 2 would be obtained when there is a double sequence of service requirements.

3.3 Asymptotic approximation for the prelimit sequence

In order to develop approximations that depend on the parameter triples of the converging processes, we want to replace the unspecified function $\overline{\nu}$ in the limit from Theorem 3 by a function depending directly on the parameter triple. We provide asymptotic justification for that step now. In Sect. 3.4, we apply the resulting asymptotic approximation to obtain explicit distributional results under additional assumptions about the instantaneous mean function of the arrival process.

A new sequence will now be defined and its asymptotic equivalence to the prelimit sequence from Theorem 2 proven. For that purpose, $d_{M_1}(x,\, y)$ will denote the M_1 metric for x and y in D or D^2.

Corollary 5 (asymptotically equivalent sequence) *Using the definitions and assumptions from Theorem 3, let*

$$\acute{X}^n \equiv nZ^n - n\mu^n e + n^{1/2}\overline{T} \text{ and } \acute{W}^n \equiv \phi\Big(\acute{X}^n,\, \acute{W}^n(0)\Big) \text{ for } n \geq 1, \tag{3.19}$$

where $Z^n(t) \equiv \int_0^t \zeta^n(v)dv$, $\acute{W}^n(0) \stackrel{d}{=} W^n(0)$, *and* $\overline{T}$ *is the* $\psi -$ GMP *with parameter pair* $(\alpha^*,\, \beta^*) = \big(b + bc_s^2,\, -b\gamma\big)$. *Then*

$$d_{M_1}\Big(\Big(n^{-1/2}X^n,\, n^{-1/2}W^n\Big),\, \Big(n^{-1/2}\acute{X}^n,\, n^{-1/2}\acute{W}^n\Big)\Big) \Rightarrow 0 \text{ in } \mathcal{R} \text{ as } n \to \infty. \tag{3.20}$$

Proof By Theorem 3,

$$(T_n,\, X_n,\, W_{\rm n}) \Rightarrow \big(\overline{T} + \overline{\nu},\, \overline{X},\, \overline{W}\big) \text{ in } \Big(D^3,\, M_1\Big), \tag{3.21}$$

where $\overline{X} = \overline{\nu} - \overline{\mu}e + \overline{T}$. Let $\acute{X}_n \equiv n^{\frac{-1}{2}}\acute{X}^n$ and $\acute{W}_n \equiv n^{\frac{-1}{2}}\acute{W}^n$. Applying (3.5), (3.7), and (3.16),

$$\begin{aligned} \acute{X}_n &= n^{-1/2}\Big(nbM^n - n\mu^n e + n^{1/2}\overline{T}\Big) \\ &= bn^{1/2}\big(M^n - e\big) - n^{1/2}\big(\mu^n - b\big)e + \overline{T} \\ &\Rightarrow \overline{\nu} - \overline{\mu}e + \overline{T} = \overline{X} \text{ in } (D,\, M_1). \end{aligned} \tag{3.22}$$

Using the assumption that $\acute{W}^n(0) \stackrel{d}{=} W^n(0)$, the continuous mapping theorem then implies that

$$\left(\overline{T},\, \acute{X}_n,\, \acute{W}_n\right) \Rightarrow \left(\overline{T},\, \overline{X},\, \overline{W}\right) \text{in } \left(D^3,\, M_1\right). \tag{3.23}$$

By (3.21), $T_n - \overline{\nu} \Rightarrow \overline{T}$ in $(D,\, M_1)$. We apply the Skorohod representation theorem from Theorem 3.2.2 of [20] to obtain $d_{M_1}\left(T_n - \overline{\nu},\, \overline{T}\right) \Rightarrow 0$ in $\mathcal{R}$. The convergence together theorem from Theorem 11.4.7 of [20] then implies that.

$$\left(T_n - \overline{\nu},\, \overline{T}\right) \Rightarrow \left(\overline{T},\, \overline{T}\right) \text{in } \left(D^2,\, M_1\right). \tag{3.24}$$

Since X_n and W_{n} are functions of T_n, and $\acute{X}_n$ and $\acute{W}_n$ are functions of $\overline{T}$, we obtain

$$\left(n^{-1/2}X^n,\, n^{-1/2}\acute{X}^n,\, n^{-1/2}W^n,\, n^{-1/2}\acute{W}^n\right) \Rightarrow \left(\overline{X},\, \overline{X},\, \overline{W},\, \overline{W}\right) \text{in } \left(D^4,\, M_1\right) \tag{3.25}$$

using (3.21), (3.23), (3.24), and the continuous mapping theorem. The conclusion in (3.20) is then a consequence of (3.25) and the converse of the convergence together theorem in Theorem 11.4.8 of [20]. □

For the prelimit sequence satisfying the conditions of Corollary 5, (3.22) implies that

$$\left(X^n \equiv N^n - n\mu^n e,\, W^n\right) \stackrel{d}{\approx} \left(\acute{X}^n \equiv nZ^n - n\mu^n e + n^{1/2}\overline{T},\, \acute{W}^n\right) \tag{3.26}$$

with error that is $o\left(n^{1/2}\right)$ as $n \to \infty$ on bounded intervals, i.e., the error is asymptotically negligible for large n after dividing by $n^{1/2}$. On the right-hand side, $n^{1/2}\overline{T}$ is the zero-mean Gaussian process with distribution determined by $Cov\left[n^{1/2}\overline{T}(s),\, n^{1/2}\overline{T}(t)\right] = nbs\left(1 + c_s^2 + \gamma t\right)$ for $0 \le s \le t$. By (3.3), it is therefore the ψ − GMP with parameter pair $(\alpha^*,\, \beta^*) = \left(nb\left(1 + c_s^2\right),\, -nb\gamma\right)$. On the left-hand side, N^n is a GPP with parameter triple $(\lambda^n(t),\, \gamma^n,\, \beta^n) = (n\zeta^n(t),\, \gamma,\, nb)$. We can therefore eliminate explicit reference to n from (3.26) for any particular n by substituting $\lambda(t) \equiv n\zeta^n(t)$, $\Lambda(t) \equiv nZ^n(t)$, $\beta \equiv nb$, $\mu \equiv n\mu_n$, and $\overline{\overline{T}}(t) \equiv n^{1/2}\overline{T}(t)$. With those substitutions, (3.26) becomes

$$(X \equiv N - \mu e,\, W \equiv \phi(X,\, W(0))) \stackrel{d}{\approx} \left(\acute{X} \equiv \Lambda - \mu e + \overline{\overline{T}},\, \acute{W} \equiv \phi\left(\acute{X},\, \acute{W}(0)\right)\right), \tag{3.27}$$

where N is then the GPP with parameter triple $(\lambda(t),\, \gamma,\, \beta)$, $\overline{\overline{T}}$ is the ψ − GMP with parameter pair $(\alpha^*,\, \beta^*) = \left(\beta + \beta c_s^2,\, -\beta\gamma\right)$, and μ is the service rate. The parameters $\beta,\, \mu,\, \lambda(t)$ are large when the index n is large before the substitutions.

Recall that $\Lambda(t) = \int_0^t \lambda(s)ds$ is the mean function of N and observe that the right-hand side of (3.22) is then determined by the parameter triple $(\lambda(t),\, \gamma,\, \beta)$, the squared

coefficient of variation c_s^2, and the service rate μ. As with approximations obtained from classical HTLTs, the approximation in (3.27) is not necessarily accurate for particular choices of those parameters, but Theorem 1 provides the qualitative criteria that μ and $\lambda(t)$ both should be close to β for the approximations to be accurate.

3.4 The transient distribution

According to the results in Sect. 3.3, we can approximate the workload process for a *P/GI/1* queue by the reflection of a $\psi -$ GMP with time-dependent drift. A $\psi -$ GMP is a generalization of Brownian motion, and the transient distribution of reflected Brownian motion (RBM) with constant drift is well known; see Chapter 1 of [11], Chapter 8 of [16, 1], and [2]. The transient distribution of a reflected $\psi -$ GMP with constant drift was derived in [10] and applied in [9] for $\psi -$ GPPs. We generalize that result for the case when the drift is time-dependent to describe the transient distribution of the reflection on an interval conditional on history up to the start of the interval. That holds when the drift is any time-dependent function in D prior to the interval but is constant on the interval. The time-dependent drift prior to the interval enters into the transient distribution on the interval because the increments of a $\psi -$ GMP are dependent.

The proof uses two lemmas from Sect. 4. Lemma 1 restates a result from (30) in [9] on the transient distribution of a reflected $\psi -$ GMP with constant drift. A new proof based on the proof for RBM in [11] is provided. Lemma 2 is a new result describing the transient distribution of a $\psi -$GMP with time-dependent drift conditional on its history. That result is analogous to the restart property for GPPs described in Proposition 1 of [9] and originally derived in [6]. The lemmas are applied using the memoryless property of the reflection map from Proposition 10 on page 21 of [11].

The approximation in (3.27) reduces as a special case to

$$(X(s),\ W(s),\ W(s+t)) \stackrel{d}{\approx} \left(\acute{X}(s),\ \acute{W}(s),\ \acute{W}(s+t)\right) \text{ for } s,\ t \geq 0, \tag{3.28}$$

where X is the net input process and W is the workload process for a P/D/1 queue with service rate μ, squared coefficient of variation c_s^2, and arrival process with parameter triple $(\lambda(t),\ \gamma,\ \beta)$, and where

$$\acute{X}(t) \equiv \int_0^t \lambda(s)ds - \mu t + \overline{\overline{T}}(t) \text{ and } \acute{W}(t) \equiv \phi\left(\acute{X}\right)\left(t,\ \acute{W}(0)\right) \text{ for } t \geq 0, \tag{3.29}$$

while $\overline{\overline{T}}$ is the $\psi -$ GMP with parameter pair $(\alpha^*,\ \beta^*) = \left(\beta + \beta c_s^2,\ -\beta\gamma\right)$.

Then,

$$P(W(s+h) \leq w_{s+h} | X(s),\ W(s)) \approx P\left(\acute{W}(s+h) \leq w_{s+h} \middle| \acute{X}(s),\ \acute{W}(s)\right) \text{ for } s,\ t \geq 0. \tag{3.30}$$

We provide an explicit expression for the cumulative distribution function (cdf) on the right-hand side.

Theorem 4 *If $\acute{X}$ and $\acute{W}$ are defined as in (3.29), where $\lambda(v) = \lambda(s)$ for $s \le v \le s+t$ and $P\left(\acute{W}(0) = w_0\right) = 1$, then*

$$P\left(\acute{W}(s+t) \le w_{s+t} \middle| \acute{X}(s) = x_s,\ \acute{W}(s) = w_s\right) = F(t,\ w_{s+t}), \tag{3.31}$$

where

$$F(t,\ w) = \Phi\left(\frac{w - w_s - \omega_s t}{\sqrt{t\left(\alpha^* - \beta_s^* t\right)}}\right) - e^{\frac{-2w(\beta_s^* w - \alpha^* \omega_s)}{\alpha^{*2}}} \Phi\left(\frac{\left(2\beta_s^* w - \alpha^* \omega_s\right)t - \alpha^*(w + w_s)}{\alpha^* \sqrt{t\left(\alpha^* - \beta_s^* t\right)}}\right), \tag{3.32}$$

while $\Phi(t)$ is the standard normal cdf, and

$$\alpha^* = \beta\left(1 + c_s^2\right),\ \beta_s^* \equiv \frac{-\beta\gamma\left(1 + c_s^2\right)}{1 + c_s^2 + \gamma s},\ \text{and}\ \omega_s \equiv \lambda(s) - \mu + \frac{\gamma\left(x_s - \left(\int_0^s \lambda(v)dv - \mu s\right)\right)}{1 + c_s^2 + \gamma s}. \tag{3.33}$$

Proof Let $\acute{W}_s(h) \equiv \acute{W}(s+h)$, $d_s\acute{X}(h) \equiv \acute{X}(s+h) - \acute{X}(s)$, and $\acute{X}_s(h) \equiv \left(d_s\acute{X}(h) \middle| \acute{X}(s) = x_s\right)$ for $0 \le h \le t$. By the memoryless property of the reflection map from Proposition 10 on page 21 of [11], $\acute{W}_s \equiv \phi\left(d_s\acute{X},\ \acute{W}(s)\right)$ with probability 1. Then, with probability 1,

$$\begin{aligned}\left(\acute{W}_s \middle| \acute{X}(s) = x_s,\ \acute{W}(s) = w_s\right) &= \left(\phi\left(d_s\acute{X},\ \acute{W}(s)\right) \middle| \acute{X}(s) = x_s,\ \acute{W}(s) = w_s\right) \\ &= \left(\phi\left(d_s\acute{X},\ w_s\right) \middle| \acute{X}(s) = x_s\right) = \phi\left(w_s,\ \acute{X}_s\right)\end{aligned} \tag{3.34}$$

where the next-to-last equality holds by the Markov property of $\acute{X}$, the definition of $\acute{W}$, and the assumption that $P\left(\acute{W}(0) = w_0\right) = 1$ (which implies that $\acute{X}$ is independent of $\acute{W}(0)$). By Lemma 2 in Sect. 4, $\acute{X}_s$ in the final expression of (3.34) is a ψ − GMP on $[0,\ t]$ with parameter pair $\left(\alpha^*,\ \beta_s^*\right)$ and drift ω_s. The result in (3.31)–(3.33) then follows from Lemma 1 in Sect. 4 with the substitutions there of $\alpha^* = \beta$, $\beta^* = \beta_s^*$ and $\omega = \omega_s$ in (3.33). □

Theorem 4 may be applied when the instantaneous mean function has been estimated for the past, a forecast is needed, and the best available estimate of the mean function over the forecast horizon is the estimate that has been obtained for its value at the current time. To estimate the instantaneous mean function, a parametric form would generally need to be assumed. Corollary 4 suggests that not much generality will

be lost by assuming that the arrival process is a $\psi^k -$ GPP, for which the instantaneous mean function is piecewise constant.

We describe how Theorem 4 applies when the arrival process is a $\psi^k -$ GPP.

Corollary 6 *If* $\acute{X}$ and $\acute{W}$ *are defined as in (3.29), where* $P\left(\acute{W}(0) = w_0\right) = 1$ and $\lambda(t) = \lambda_i u(t)$ *for* $t_{i-1} \le t < t_i$ *and* $1 \le i \le k \le \infty$, *and if* $t_{i-1} \le s < s + t < t_i$ *for some* i, *then* (3.31)-(3.33) *hold, where*

$$\omega_s = \lambda_i - \mu + \frac{\gamma\left(x_s - \left(\lambda_i(s - t_{i-1}) + \left(\sum_{j=1}^{i-1} \lambda_j\left(t_j - t_{j-1}\right)\right) - \mu s\right)\right)}{1 + c_s^2 + \gamma s}. \quad (3.37)$$

4 Lemmas for Theorem 4

We state and prove two lemmas used in the proof of Theorem 4. The lemmas are discussed in Sect. 3.4.

A zero-mean real-valued Gaussian process $\{A(t) : 0 \le t < \mathrm{T}\}$ is defined in [10] to be a $\psi -$ GMP with parameter pair (α^*, β^*) if $A(0) = 0$ and $Cov[A(s), A(t)] = s(\alpha^* - \beta^* t)$ $for\, 0 \le s \le t < \mathrm{T}$, where $\alpha^* > 0$ and $\infty < \beta^* < \infty$. If $\beta^* > 0$, then it is necessary that $\mathrm{T} \le \alpha^*/\beta^*$; otherwise, $\mathrm{T} \le \infty$. When A is defined in that way, the process

$$X^*(t) \equiv \omega t + A(t) \text{ for } 0 \le t < T \quad (4.1)$$

is called a $\psi -$ GMP on [0, T) with parameter pair (α^*, β^*) and drift ω. If $\beta^* = 0$, then X^* is Brownian motion with $Var\left[X^*(t)\right] = \alpha^* t$ and drift ω.

The first lemma is a special case of Theorem 5 from [10]. We provide a different proof below derived from first principles and closely following the proof in [11] for the RBM case. The proof of Theorem 3 will apply the lemma when $\beta^* < 0$. The result below, which holds regardless of the sign of the parameter β^*, is therefore more general than we will require for Theorem 3. Recall that ϕ is the reflection map defined in (3.13).

Lemma 1 *If* X^* *is defined as in (4.1) and* $W^* \equiv \phi(w_0, X^*)$, *then*

$$\begin{aligned} F^*(h, w) &\equiv P\left(W^*(h) \le w\right) \\ &= \Phi\left(\frac{w - w_0 - \omega h}{\sqrt{h(\alpha^* - \beta^* h)}}\right) \\ &\quad - e^{\frac{-2w(\beta^* w - \alpha^* \omega)}{\alpha^{*2}}} \Phi\left(\frac{(2\beta^* w - \alpha^* \omega)h - \alpha^*(w + w_0)}{\alpha^* \sqrt{h(\alpha^* - \beta^* h)}}\right), \end{aligned} \quad (4.2)$$

where $w \ge 0, 0 \le h < T$, *and* $\Phi(z) \equiv (2\pi)^{-1/2} \int_{-\infty}^{z} exp\left(-y^2/2\right) dy$ *is the standard normal cdf.*

Proof Case 1: $\beta^* > 0$.

Let

$$B(t) = \frac{1+t\beta^*}{\alpha^*} A\left(\frac{t\alpha^*}{1+t\beta^*}\right) \text{ for } t \geq 0. \tag{4.3}$$

Then, B is standard Brownian motion because it is a zero-mean Gaussian process with $Cov[B(s), B(t)] = s$ for $s \leq t$; see page 184 of Adler [3] for a discussion of that definition. Furthermore

$$Y(t) \equiv w_0 + X^*\left(\frac{t\alpha^*}{1+t\beta^*}\right) = w_0 + \omega\frac{t\alpha^*}{1+t\beta^*} + \frac{\alpha^*}{1+t\beta^*}B(t) \tag{4.4}$$

using (4.1) and (4.3).

Because $1 + s\beta^* > 0$ when $s \geq 0$, it follows from (4.4) that

$$\begin{aligned}
&\inf_{0\leq s\leq t} Y(s) \leq y \Leftrightarrow \inf_{0\leq s\leq t}\left(\frac{\omega\alpha^* s + \alpha^* B(s) + (w_0 - y)(1+s\beta^*)}{1+s\beta^*}\right) \leq 0 \\
&\Leftrightarrow \inf_{0\leq s\leq t}\left(\omega\alpha^* s + \alpha^* B(s) + (w_0 - y)\left(1+s\beta^*\right)\right) \leq 0 \\
&\Leftrightarrow \inf_{0\leq s\leq t}\left(\eta s + \alpha^* B(s)\right) \leq y - w_0,
\end{aligned} \tag{4.5}$$

where $\eta \equiv \omega\alpha^* + (w_0 - \mathrm{y})\beta^*$.

By (4.4)–(4.5),

$$\begin{aligned}
G(x, y) &\equiv P\left(Y(t) \leq x, \inf_{0\leq s\leq t} Y(s) \leq y\right) \\
&= \int_{-\infty}^{y-w_0} \int_{b}^{(x-y)\ (1+t\beta^*)+y-w_0} P\left(\eta t + \alpha^* B(t) \in da, \inf_{0\leq s\leq t}\left(\eta s + \alpha^* B(s)\right) \in db\right).
\end{aligned} \tag{4.6}$$

Applying the change of measure theorem on page 10 of [11] followed by the Reflection Principle on pages 7–9 of [11], we obtain

$$\begin{aligned}
&P\left(\eta t + \alpha^* B(t) \in da, \inf_{0\leq s\leq t}\left(\eta s + \alpha^* B(s)\right) \in db\right) \\
&= \exp\left(\frac{\eta a}{\alpha^{*2}} - \frac{\eta^2 t}{2\alpha^{*2}}\right) P\left(\alpha^* B(t) \in da, \inf_{0\leq s\leq t}\left(\alpha^* B(s)\right) \in db\right) \\
&= \exp\left(\frac{\eta a}{\alpha^{*2}} - \frac{\eta^2 t}{2\alpha^{*2}}\right) \frac{\sqrt{2}(a-2b)\exp\left(\frac{(a-2b)^2}{2\alpha^{*2}t}\right) da db}{\sqrt{\pi}\alpha^{*3}t^{3/2}}.
\end{aligned} \tag{4.7}$$

Using (4.6) and (4.7),

$$g(x,\,y) \equiv \frac{d}{dy}\frac{d}{dx}G(x,\,y) = \frac{\sqrt{2}(w_0+x-2y)(1+t\beta^*)^2 e^{-\left(\frac{t\omega^2}{2}+\frac{(1+t\beta^*)(w_0-x)\omega}{\alpha^*}+\frac{(1+t\beta^*)\left((1+\beta^*)\left(x^2+w_0^2\right)+4y(y-x-w_0)+2w_0x(1-t\beta^*)\right)}{2t\alpha^{*2}}\right)}}{\sqrt{\pi}t^{3/2}\alpha^{*3}}. \tag{4.8}$$

Using the definitions from (3.15), (4.2), (4.4), (4.6), and (4.8),

$$\begin{aligned}\frac{d}{dw}F^*\left(\frac{t\alpha^*}{1+t\beta^*},\,w\right) &= \frac{d}{dw}\left(P\left(Y(t)-\inf_{0\le s\le t}Y(s)\le w,\ \inf_{0\le s\le t}Y(s)\le 0\right)\right.\\ &\quad\left.+P\left(Y(t)\le w,\ \inf_{0\le s\le t}Y(s)>0\right)\right)\\ &= \frac{d}{dw}\left(\int_{-\infty}^{0}\int_{y}^{w+y}g(x,\,y)dxdy+\int_{0}^{w_0}\int_{y}^{w}g(x,\,y)dxdy\right)\\ &= \int_{-\infty}^{0}g(w+y,\,y)dy+\int_{0}^{w_0}g(w,\,y)dy. \end{aligned}\tag{4.9}$$

Substituting (4.8) into (4.9), the integrals on the right-hand side of the final equality in (4.9) can be solved by completing the squares in the exponent; see page 13 of Harrison [11] for an example where completing the squares is applied in the RBM case. We conclude that

$$\begin{aligned} f^*(h,\,w) &\equiv \frac{d}{dw}F^*(h,\,w)\\ &= \frac{1}{\sqrt{h(\alpha^*-\beta^*h)}}\Phi'\left(\frac{w-w_0-\omega h}{\sqrt{h(\alpha^*-\beta^*h)}}\right)\\ &\quad+e^{\frac{-2w(\beta^*w-\alpha^*\omega)}{\alpha^{*2}}}\left[\begin{array}{l}\dfrac{4\beta^*w-2\alpha^*\omega}{\alpha^{*2}}\left(\Phi\left(\dfrac{(2\beta^*w-\alpha^*\omega)h-\alpha^*(w+w_0)}{\alpha^*\sqrt{h(\alpha^*-\beta^*h)}}\right)\right)\\ +\dfrac{(\alpha^*-2\beta^*h)}{\alpha^*\sqrt{h(\alpha^*-\beta^*h)}}\Phi'\left(\dfrac{(2\beta^*w-\alpha^*\omega)h-\alpha^*(w+w_0)}{\alpha^*\sqrt{h(\alpha^*-\beta^*h)}}\right)\end{array}\right] \end{aligned}\tag{4.10}$$

where $\Phi'(z) \equiv (d/dz)\Phi(z)$ is the standard normal pdf. Differentiating the cdf in (4.2), we confirm that it agrees with the probability density function in (4.10).

Case 2: $\beta^* < 0$.

Replace the condition in (4.3) that $t \ge 0$ with the condition that $0 \le t < T/(\alpha^* - T\beta^*)$. Then, the argument of $A(\cdot)$ in (4.3) is still constrained to the interval $[0,\ T)$, and the term $1 + t\beta^*$ in (4.3) is still always positive. With that modification, the remainder of the proof for Case 1 holds with no additional changes. □

The second lemma describes the distribution of a ψ − GMP with time-dependent drift conditional on its history.

Lemma 2 *Let* $\acute{X}(t) \equiv \Lambda(t) - \mu t + \acute{\overline{T}}(t)$ *for* $t \geq 0$ *where* μ *is real,* $\Lambda(t) = \int_0^t \lambda(v)dv$ *for* $\lambda(v)$ *real and integrable, and* $\acute{\overline{T}}$ *is the* ψ − GMP *with parameter pair* $(\alpha^*, \beta^*) = \left(\beta + \beta c_s^2, -\beta\gamma\right)$. *If* $\lambda(v) = \lambda(s)$ *for* $0 \leq s \leq v < s + T$, *then*

$$\acute{X}_s(t) \equiv \left(\acute{X}(t+s) - \acute{X}(s) \middle| \acute{X}(s) = x_s\right) \tag{4.11}$$

is a ψ − GMP *on* $[0,\ T)$ *with parameter pair* $\left(\alpha^*,\ \beta_s^*\right)$ *and constant drift* ω_s, *where*

$$\beta_s^* \equiv \frac{-\beta\gamma\left(1 + c_s^2\right)}{1 + c_s^2 + \gamma s} \text{ and } \omega_s \equiv \lambda(s) - \mu + \frac{\gamma(x_s - (\Lambda(s) - \mu s))}{1 + c_s^2 + \gamma s}. \tag{4.12}$$

Proof Under the assumptions,

$$E\left[\acute{X}(s+t)\right] = \Lambda(s+t) - \mu(s+t) = \Lambda(s) - \mu s + (\lambda(s) - \mu)t \tag{4.13}$$

for $s \geq 0$ and $0 \leq t <$ T, and

$$\Gamma(s,\ t) \equiv \text{Cov}\left[\acute{X}(s),\ \acute{X}(t)\right] = s\left(\alpha^* - \beta^* t\right) = \beta s\left(1 + c_s^2 + \gamma t\right) \text{ for } 0 \leq s \leq t < s + T. \tag{4.14}$$

Since $\acute{X}$ is a Gaussian process, so is $\acute{X}_s$. We substitute (4.13) and (4.14) into well-known formulas for the conditional mean and covariance of the multivariate normal distribution, for example from Sect. 6.2.2 of [19], to obtain

$$\begin{aligned} E\left[\acute{X}_s(t)\right] &= E\left[\acute{X}(t+s) - \acute{X}(s) \middle| \acute{X}(s) = x_s\right] = E\left[\acute{X}(t+s) \middle| \acute{X}(s) = x_s\right] - x_s \\ &= E\left[\acute{X}(t+s)\right] + \Gamma(s,\ t+s)\Gamma(s,\ s)^{-1}\left(x_s - E\left[\acute{X}(s)\right]\right) - x_s \\ &= \Lambda(s) - \mu s + (\lambda(s) - \mu)t + \frac{s(\alpha^* - \beta^*(s+t))(x_s - (\Lambda(s) - \mu s))}{s(\alpha^* - \beta^* s)} - x_s \\ &= \omega_s t \end{aligned}$$

and

$$\begin{aligned} Cov\left[\acute{X}_s(t),\ \acute{X}_s(u)\right] &= Cov\left[\acute{X}(t+s) - \acute{X}(s),\ \acute{X}(u+s) - \acute{X}(s) \middle| \acute{X}(s) = x_s\right] \\ &= Cov\left[\acute{X}(t+s),\ \acute{X}(u+s) \middle| \acute{X}(s) = x_s\right] \\ &= \Gamma(t+s,\ u+s) - \Gamma(s,\ u+s)\Gamma(s,\ s)^{-1}\Gamma(s,\ t+s) = t\left(\alpha^* - \beta_s^* u\right) \end{aligned}$$

for $0 \leq t \leq u < \mathrm{T}$. The result that $\acute{X}_s$ is a $\psi -$ GMP on [0, T) with parameter pair $\left(\alpha^*, \beta_s^*\right)$ and drift ω_s then follows from the definition of a $\psi -$ GMP with constant drift. □

5 Concluding discussion

5.1 Extensions of GPPs

GPPs were described in [6] as a tractable generalization of NHPPs allowing for dependence between increments. According to (2.11), a GPP is defined by only a single scaler parameter beyond the instantaneous rate function that defines a NHPP. As discussed in Sect. 3.4, a GPP possesses a restart property that is useful for modeling: its future increments given its history are another GPP with modified parameters. The asymptotic approximation obtained in this paper for a GPP is even more tractable than a GPP itself, as Theorem 4 illustrates, and possesses its own restart property, as Lemma 2 shows.

Generalizations of GPPs have been considered. Section 4.1 of [13] considered a generalization where the parameters β and γ in (2.1) are themselves functions of time, but showed that the method used in that paper to obtain an exact analytic solution for the marginal distributions of a GPP's state fails for the generalization. A generalization of a GPP considered in [4] is where the linear function of $N(t-)$ that multiplies $\kappa(t)$ in (1.1) is replaced by a general positive function. That paper derives several properties of the resulting process including the marginal distribution of its state and a version of the restart property.

In order to apply results on stationary GPPs from [9], the proofs of the limit theorems here rely on Proposition 2, which shows that a process is a GPP if and only if it is a time-transformation of a stationary GPP. (The time transformation must be the integral of a density function.) The proofs here therefore do not extend directly for either of the generalizations discussed above. It remains to study how limit theorems might be derived for such generalizations.

5.2 Motivation in application

As indicated in [9], there has long been interest in systems with path-dependent behavior. It is important to consider the possibility that ALOM may not be satisfied. For example, with a queue representing the backlog of tests at a COVID-19 testing site, a small cluster of infections randomly occurring in the area of the site early in the epidemic may spread and influence subsequent infection rates. The intensity of demand for testing would then depend on prior demand, increases in demand would be self-reinforcing, and the influence of early conditions would persist. The spread of COVID-19 was assumed to have such characteristics in [5] and modeled there as a GPP; see (8) of [5] for the particular GPP intensity function used. It remains to seriously study models of queues arising in such applications.

The asymptotic approximation obtained here for the transient distribution of a *P/GI/1* queue is applicable in a critically loaded regime where the GPP's instantaneous rate $\lambda(t)$ does not vary too much from the service rate μ. (The function $\kappa(t)$ in (2.1) therefore cannot differ too much from the form in (1.2) for a stationary GPP.) Although such conditions may not hold at all times in practice, the approximation developed here may be applied over any interval where such conditions do hold, as justified by the GPP's restart property. Critically loaded intervals are particularly interesting because that is where significant queue lengths are likely and their evolution differs significantly from that of the queue's unreflected net input process.

Declarations

Conflict of interest The authors declared that there is no conflict of interest.

References

1. Abate, J., Whitt, W.: Transient behavior of regulated Brownian motion, I: starting at the origin. Adv. Appl. Probab. **19**(3), 560–598 (1987)
2. Abate, J., Whitt, W.: Transient behavior of regulated Brownian motion, II: non-zero initial conditions. Adv. Appl. Probab. **19**(3), 599–631 (1987)
3. Adler, R.J.: The Geometry of Random Fields. Wiley, Chichester (1981)
4. Badia, F.G., Mercier, S., Sanguesa, C.: Extensions of the generalized Polya process. Method Comput Appl Prob **21**, 1057–1085 (2019)
5. Barraza, N.R, Pena, G., Moreno, V.: A non-homogeneous Markov early epidemic growth dynamics model. *Chaos, Solitions & Fractals,* vol. 139: 110297, (2020).
6. Cha, T.H.: Characterization of the generalized Polya process and its applications. Adv. Appl. Probab. **46**(4), 1148–1171 (2014)
7. Daley, D.J., Vere-Jones, D. An Introduction to the Theory of Point Processes: Volume I; Elementary Theorems and Methods, 2nd edn, Springer, New York (2003)
8. Feller, W., An Introduction to Probability Theory and its Applications, Vol. 1, 3rd edn, John Wiley, New York (1968)
9. Fendick, K.W., Whitt, W.: Queues with path-dependent arrival processes. J. Appl. Probab. **58**, 484–504 (2021)
10. Fendick, K.: Brownian motion minus the independent increments: representation and queuing application. *Probability in the Engineering and Informational Sciences,* To appear.
11. Harrison, J.M.: Brownian Motion and Stochastic Flow Systems. Wiley, New York (1985)
12. Kim, S.-H., Whitt, W.: Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes. Manuf. Serv. Oper. Manag. **16**(3), 464–480 (2014)
13. Konno, T.H.: On the exact solution of a generalized Polya process. Adv. Math. Phys., vol. Article ID 504267 (2010)
14. Koops, D., Saxena, M., Boxma, O.J., Mandjes, M.: Infinite-serve queues with Hawkes input. J. Appl. Probab. **55**, 920–943 (2018)
15. Mandelbaum, A., Massey, W.A.: Strong approximations for time-dependent queues. Math. Oper. Res. **20**(1), 33–64 (1995)
16. Newell, G.F.: Applications of Queueing Theory, 2nd edn. Chapman and Hall, New York (1982)
17. Pang, G., Whitt, W.: Continuity of a queueing integral representation in the M_1 topology. Annals of Applied Probability **1**(1), 214–237 (2010)
18. Parthasarathy, K.R.: Probability Measures on Metric Spaces. Academic Press, New York (1967)
19. Puntanen, S., Styan, G.P.: Schur complements in statistics and probability. In: Schur Complement and Its Applications, Numerical Methods and Algorithms, vol. 4. Springer, pp. 163–226 (2005)

20. Whitt, W.: Stochastic-Process Limits. Springer, New York (2002)
21. Whitt, W.: Time varying queues. Que. Mod. Serv. Manag. **1**(2), 79–164 (2018)
22. Whitt, W., You, W.: Time-varying robust queueing. Oper. Res. **67**(6), 1766–1782 (2019)