Non-Stationary Path-Dependent Queues in Heavy Traffic

Kerry Fendick
Johns Hopkins University Applied Physics Laboratory
Kerry.Fendick@jhuapl.edu

Ward Whitt
Columbia University
ww2040@columbia.edu

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 depend on the history of the arrival process, thus capturing the self-reinforcing property that one might expect for demand at a Covid-19 test site. 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.

Key Words: path-dependent stochastic processes; generalized Polya process; Gaussian Markov process; diffusion approximations; queues; heavy-traffic limit

2010 Mathematics Subject Classification: Primary 60K25

Secondary 60F17, 90B22

1 Introduction

Queueing systems are usually modeled under assumptions that result in an asymptotic loss of memory (ALM) causing the influence of early conditions to dissipate over time. ALM may not apply, though, when the stochastic intensity (or conditional rate) of new arrivals to a queue depends on past history. For 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. A process lacking the ALM property may be described as *path-dependent*. For stationary processes, lack of the ALM property implies lack of ergodicity since even long-term averages can then depend on early conditions.

The Generalized Polya Process (GPP), as defined and characterized in [1], [2], and [3], is a counting process that is path-dependent. It therefore would be a candidate for modeling demand at a COVID19 test site. A heavy-traffic limit theorem (HTLT) was derived in [3] for a sequence of *P/GI/1* queues with stationary GPP arrival processes and independent and identically distributed (i.i.d) service times. The limit for the queue-length process is the reflection of a Gaussian diffusion process (a continuous Gaussian Markov process) with constant drift in a class characterized in [4].

A law of large numbers with convergence to a random limit was also obtained in [3] for the P/GI/1 queue-length process. It shows that the queue-length process (not normalized by time) approaches infinity with positive probability as time increases, regardless of the queue's traffic intensity. The variance of the process grows so large that long-range forecasts would not be very useful, but short-range forecasts based on an understanding of transient behavior may well still be useful. The heavy-traffic limit of the queue-length process in [3] is in a class of reflected processes with a transient distribution explicitly characterized by Theorem 5 of [4]. Corollary 6 of [3] provides an equivalent but more revealing formulation of that result.

An assumption that the arrival process is stationary may not be warranted. For example, social distancing, mask wearing, and vaccinations may influence demand for COVID-19 testing by reducing infection rates over time. The intensity of demand would then depend on changing factors beyond the internal history of the demand process, and the demand would be non-stationary as well as path dependent. Although almost all real-world processes are naturally viewed as non-stationary over long enough time scales, the ALM property is commonly cited to justify modeling them as locally stationary and their increments on widely spaced intervals as independent. Non-stationarity over long time scales may then be ignored over shorter time scales. While a path-dependent process may be locally stationary, its increments on disjoint intervals are strongly dependent even when the intervals are spaced widely apart. Non-stationarity in the past should then not generally be ignored.

A Generalized Polya Process (GPP) can be used to model a non-stationary path-dependent arrival process because stationarity only occurs as a special case . Motivated by the example of a COVID-19 testing site, the current paper studies the behavior of a *P/GI/1* queue with a non-stationary GPP arrival process. We develop an HTLT for the workload (or service backlog) process as time-varying traffic intensities approach one uniformly, and we apply it to derive an approximation for the transient distribution of the workload process with error that approaches zero as traffic intensities approach one. The asymptotic approximation is the reflection of a Gaussian Markov process with time-dependent drift. The result on its transient distribution generalizes results in [3] and [4].

In applications, it may be useful to work with GPPs that depend on a finite set of parameters. We define piecewise stationary GPPs with that property, and show how any non-stationary GPP can be approximated by a piecewise stationary GPP with a sufficiently large number of pieces. The HTLT and the associated asymptotic approximation for the transient distribution then apply as special cases. A piecewise-stationary GPP has stationary increments on each piece, but that characteristic is not sufficient to define a piecewise stationary GPP because it does not describe the dependence between increments on different pieces. For the same reason, a piecewise application of the HTLT from [3] (assuming only that the arrival process is stationary on each piece) would result in an incomplete characterization of a P/D/1 queue with an arrival process that is a piecewise stationary GPP. The HTLT and transient distribution derived here account for the GPP's global dependence structure, as is necessary to understand how early conditions affect subsequent queueing behavior.

HTLTs were previous developed for sequences of queueing systems with non-stationary arrival processes; e.g., [5] and [6]. The limits derived in those examples can exhibit critically loaded regions where traffic intensities approach one but also may exhibit underloaded or overloaded regions where traffic intensities approach other values. Those HTLTs provide insight into the behavior of non-stationary queueing systems, but the limit process in underloaded regions is equal to zero and therefore is not very useful as an approximation. For the HTLT derived here, traffic intensities approach one everywhere, so that a non-degenerate limit is obtained for the entire time domain. Traffic intensities must be uniformly close to one for approximations based on the HTLT to be accurate.

In the remainder of this paper, Section 2 surveys results on GPPs and provides a new representation of a GPP in terms of its instantaneous mean function. It shows how to construct piecewise stationary GPPs and describes how they approximate general GPPs. Section 3 derives a heavy-traffic limit theorem for a nonstationary *P/GI/1* queue, provides asymptotic justification for approximating its workload process by the reflection of a Gaussian Markov process with a time-dependent drift, and applies the approximation to describe the transient distribution of the workload process conditional on the queue's history. Section 4 proves two lemmas used in the proof of the transient result.

2 A Piecewise-Stationary Generalized Polya Process ($\boldsymbol{\psi^k - GPP}$)

In this section we briefly review generalized Polya processes (GPPs), as developed in [1] - [3]. Then we show how to construct a piecewise-stationary GPP with k pieces, which we refer to as a $\psi^k - GPP$. We then show that, under regularity conditions, a general GPP can be approximated by a $\psi^k - GPP$ for suitably large k.

A GPP with parameter triple $(\kappa(t), \gamma, \beta)$ is defined in [1] as the orderly point process $\{N(t): t \geq 0\}$ with $N(0) = 0$ and stochastic intensity function

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

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 Section 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* [3]) *The GPP* $\widetilde{N}$ *with parameter triple* $(\kappa(t), \gamma, \beta)$, *where*

$$\kappa(t) = 1/(\gamma t + 1) \tag{2.2}$$

is a stationary point process with mean and covariance functions

$$E\big[\widetilde{N}(t)\big] = \beta t \;\; and\; Cov\big[\widetilde{N}(s), \widetilde{N}(t)\big] = \beta s(1 + \gamma t)\; for\; 0 \leq s \leq t. \tag{2.3}$$

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

$$\mathrm{K}(t) = \int_0^t \kappa(s)\, ds, \tag{2.4}$$

then $N(t)$ has a negative binomial distribution with mean $E[N(t)] = \tau p(t)/\big(1 - p(t)\big)$ and $Var[N(t)] = \tau p(t)/\big(1 - p(t)\big)^2$, where $\tau \equiv \gamma/\beta$ and $p(t) = 1 - \exp\big(-\gamma\, \mathrm{K}(t)\big)$. If (2.2) holds, then $\mathrm{K}(t) = \gamma^{-1} \log(\gamma t + 1)$, and $1 - p(t) = \exp\big(-\gamma\, \mathrm{K}(t)\big) = \kappa(t)$. The mean and variance of $\widetilde{N}(t)$ easily follow. The proof of the stationarity of $\widetilde{N}$ from Theorem 1 of [3] uses the property from Theorem 3 and Remark 3 of [1] 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 (2.2) holds, those random variables are uniformly distributed. The covariance function in (2.3) follows easily from the mean and variance functions and the stationarity of $\widetilde{N}$. ■

The GPP $\widetilde{N}$ from Proposition 1 is called a stationary-increment GPP, or a $\psi - GPP$ for short, and is specified by the parameter pair (γ, β). 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 [3], 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 below. For that purpose, let $\overset{d}{=}$ denote equality in distribution for stochastic processes.

Proposition 2. (Theorem 2 of [3]*) Let N be a GPP with parameter tripl*e $(\kappa(t), \gamma, \beta)$ *and $\widetilde{N}$ be the $\psi - GPP$ with parameter pair (γ, β). Then,*

$$\{N(t) \colon t \geq 0\} \overset{d}{=} \{\widetilde{N}\big(M(t)\big) \colon t \geq 0\} \tag{2.5}$$

if and only if

$$M(t) = \gamma^{-1}\left(e^{\gamma K(t)} - 1\right) for\ t \geq 0, \tag{2.6}$$

where $\mathrm{K}(t)$ is *defined in* (2.4).

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 $E[N(t)] = \beta\gamma^{-1}\left(exp\left(\gamma\mathrm{K}(t)\right) - 1\right)$ and $Cov[N(s), N(t)] = \beta\gamma^{-1}exp\left(\gamma\mathrm{K}(t)\right)\left(exp\left(\gamma\mathrm{K}(s)\right) - 1\right)$ for $0 \leq s \leq t$.

Proof. If $\widetilde{N}$ is the $\psi - GPP$ with parameter pair (γ, β), then $E\left[\widetilde{N}(t)\right] = \beta t$ and $Cov\left[\widetilde{N}(s)\widetilde{N}(t)\right] = \beta s(1 + \gamma t)$ 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; e.g., see Remark 2 below.

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 infinitesimal mean function can be expressed as

$$\lambda(t) \equiv \lim_{h\to 0} E\,[N(t+h) - N(t)]/h = \beta\kappa(t)exp\left(\gamma\mathrm{K}(t)\right), \tag{2.7}$$

for $\mathrm{K}(t)$ *in* (2.4), *so that $\lambda(t)$ is integrable. As a consequence, the associated mean function is*

$$\Lambda(t) \equiv \mathrm{E}[N(t)] = \beta M(\mathrm{t}) = \int_0^t \lambda(v)dv = \frac{\beta exp\big(\gamma \mathrm{K}(t)\big) - \beta}{\gamma}, \tag{2.8}$$

where $M(\mathrm{t})$ is defined in (2.6), *and*

$$\mathrm{E}[N(s+t) - N(s)] = \Lambda(t+s) - \Lambda(s) = \int_s^{s+t} \lambda(v)dv, \tag{2.9}$$

$$\mathrm{Cov}[N(s+t) - N(s), N(s+u) - N(s)] \ = \left(\int_s^{s+t} \lambda(v)dv\right)\left(1 + \tau \int_s^{s+u} \lambda(v)dv\right), \tag{2.10}$$

and

$$\kappa(t) = \lambda(t)/\big(\beta + \gamma\Lambda(t)\big) \tag{2.11}$$

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

Proof: The result in (2.8) follows from Corollary 1. The results in (2.7) and (2.9) follow from (2.8). By Corollary 1 and (2.8), $\Gamma(t,u) \equiv \mathrm{Cov}[N(t), N(u)] = \left(\int_0^t \lambda(v)dv\right)\left(1 + \tau \int_0^u \lambda(v)dv\right)$ for $0 \leq t \leq u$. The result in (2.10) 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 (2.7) and (2.8), the result in (2.11) holds.■

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 Sections 3.3 and 11.5 and Chapter 12 of [9]. 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 (2.6) to denote the time-transformation function. Throughout, ⇒ will denote weak convergence of a sequence of random elements of a given topological space.

Corollary 2. Suppose that $\kappa(t)$ is an element of the function space (D, M_1). Then, so is $\lambda(t)$, 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 (2.7) and (2.11), because converge of functions in (D, M_1) implies convergence of their integrals; see [10] 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 3 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 tripl*e $(\lambda(t), \gamma, \beta)$*,* where

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

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

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

$for\ t_{i-1} \le t < t_i\ and\ 1 \le \mathrm{i} \le \mathrm{k} \le \infty$, 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 $\psi - GPP\ \widetilde{N}$ with parameter pair (γ, β) defined in terms of $\kappa(t)$ in (2.2) 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

$$\Lambda^n(t) \equiv \int_0^t \lambda^n(s)ds = \beta^{-1}M^n(t), \tag{2.16}$$

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* $n \geq 1$, 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, $M^n = \beta\Lambda^n \to \beta\Lambda = M$ 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 (2.8). Applying Proposition 2 twice*,*

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

where the weak convergence step follows from continuity of the composition map by applying Theorem 13.2.3 of [9], 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 [9], 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 σ −field; see Section 11.5.3 of [9] 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 heavy-traffic limit theorem (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 Section 3.1, we provide motivation for the HTLT assumptions. In Section 3.2, we derive a FCLT for the arrival processes. In Section 3.3, we define the net input and workload processes for a *P/GI/1* queue and derive an HTLT describing them. In Section 3.4 we apply the HTLT to develop asymptotic approximations for the net input and workload processes as functions of their parameters. Finally, in Section 3.5 we provide a tractable approximation for the transient distribution of the workload process.

3.1 Motivation for the Assumptions

We will derive an HTLT for a sequence of *P/GI/1* queues as the arrival and service rates become large and the instantaneous traffic intensities approach one. The arrival process for the n^{th} queue in the sequence is a GPP with parameter triple $(\lambda^n(t), \gamma^n, \beta^n) = (n\,\zeta^n(t), \gamma, nb)$, where λ^n is the instantaneous mean, ζ^n is a deterministic nonnegative element of D, and b is a positive real constant. The service rate for the n^{th} queue is $n\mu^n$, where μ^n is a positive real constant. The remaining properties required of the parameters are contained in their assumed limits in (3.5) and (3.18) below.

To provide motivation for the dependence of the parameters on n, we remark that the assumed limits imply that

$$\zeta^n(t) \to bu(t)\; u.o.c \;\; and \;\; \mu^n \to b \; as \; n \to \infty, \tag{3.1}$$

where $u(t)$ is the unit function. As a consequence, the queue's instantaneous traffic intensity function $\rho^n(t)$, defined as

$$\rho^n(t) \equiv \lambda^n(t)/(n\mu^n) = \zeta^n(t)/\mu^n, \tag{3.2}$$

will approach one u.o.c as $n \to \infty$.

3.2 The Functional Central Limit Theorem for the Arrival Process

We first state a Functional Central Limit Theorem (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 [3, 4].

Because the limit process $\bar{N}$ 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 [4], if A is a $\psi - GMP$ with parameter pair $(\alpha^* > 0, \beta^* \leq 0)$, then

$$Cov[A(s), A(t)] = s(\alpha^* - \beta^* t) \; for \; 0 \leq s \leq t. \tag{3.3}$$

A $\psi - GMP$ is continuous with probability one. We will apply the following FCLT for $\psi - GPPs$ from [3] together with Proposition 2 to obtain a FCLT for non-stationary GPPs.

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

Proof: By Proposition 3 of [3], $\tilde{N}^n$ 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 [3], 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 with properties discussed in Section 3.1.

Proposition 5 (convergence to a $\psi - GMP$ with time-dependent drift) If

$$N_n(t) \equiv n^{-1/2}(N^n(t) - nbt) \; for \; n \geq 1, \tag{3.4}$$

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

$$n^{1/2}(\zeta^n - bu) \to \bar{\eta} \;\; in \; (D, M_1) \; as \; n \to \infty \tag{3.5}$$

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

$$N_n \Rightarrow \bar{N} + \bar{v} \; in \; (D, M_1), \tag{3.6}$$

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

Proof. Using (2.16),

$$M^n(\mathrm{t}) = \frac{1}{\beta}\int_0^t \lambda^n(v)dv = \frac{1}{b}\int_0^t \zeta^n(v)dv \equiv \frac{1}{b}\mathrm{Z}^n(t). \tag{3.7}$$

Then, (3.5) implies that

$$n^{1/2}(\mathrm{Z}^n - be) \to \bar{v} \;\; in \; (D, M_1) \; as \; n \to \infty, \tag{3.8}$$

so that

$$M_n \equiv n^{1/2}(M^n - e) \to b^{-1}\bar{v} \;\; in \; (D, M_1) \; as \; n \to \infty. \tag{3.9}$$

By Proposition 2 and the definitions from Proposition 4,

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

Then $(M^n, M_n) \to (e, b^{-1}\bar{v})$ in (D^2, M_1) by (3.9). Applying Theorem 11.4.5 of Whitt [9], $(\tilde{N}_n, M^n, M_n) \to (\bar{N}, e, b^{-1}\bar{v})$ in (D^3, M_1). The limit preservation in Theorem 13.3.1 of Whitt [9] then yields

$$N_n \overset{d}{=} (\tilde{N}_n \circ M^n + bM_n) \Rightarrow \bar{N} + \bar{v} \; in \; (D, M_1) \; as \; n \to \infty. \blacksquare$$

Remark 3. In (3.1), $\zeta^n \to bu$ u.o.c because (3.5) implies convergence to a continuous limit. ■

Remark 3. For an example of a sequence satisfying (3.5), let $\zeta^n = bu + n^{-1/2}\bar{\eta}$ for any $\bar{\eta}$ in *D.*■

3.3 Heavy Traffic Limit Theorem for the Queue

We apply Proposition 5 to develop the HTLT for a sequence *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

$$S_n(t) = n^{-1/2}\left(\sum_{k=1}^{\lfloor nt \rfloor} V_k - nt\right), t \geq 0. \tag{3.10}$$

Our key assumption is that

$$S_n \Rightarrow c_s B \; in \; (D, M_1) \; as \; n \to \infty, \tag{3.11}$$

where B is standard (0 drift, unit variance) Brownian motion. This is the classical Donsker's theorem in Section 4.3 of [9].

A sufficient condition for (3.11) 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 Section 4.4 of [9].

Then, let

$$T^n(t) \equiv \sum_{k=1}^{N^n(t)} V_k \tag{3.12}$$

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

$$X^n(t) \equiv T^n(t) - n\mu^n t, \quad t \geq 0, \tag{3.13}$$

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.,

$$W^n \equiv \phi\big(X^n, W^n(0)\big), \tag{3.14}$$

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

$$\phi(x)\big(t, w(0)\big) = w(0) + x(t) - \inf_{0 \leq s \leq t} \big\{min\{w(0) + x(s), 0\}\big\} \; for \; t \geq 0. \tag{3.15}$$

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 Section 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 discuss in Section 7.4 of [9], or more generally in Section 13.3 of [9] (as needed here, because we will apply Proposition 4, which has the $\psi - GMP$ limit $\bar{N}$ 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

$$N_n(t) \equiv n^{-1/2}(N^n(t) - nbt), \qquad T_n(t) \equiv n^{-1/2}(T^n(t) - nbt), \tag{3.16}$$

$$X_n(t) \equiv n^{-1/2} X^n(t), \quad and \quad W_n(t) \equiv n^{-1/2} W^n(t). \tag{3.17}$$

Theorem 2 (HTLT for a P/GI/1 queue with non-stationary arrival process). If $(N_n, T_n, X_n, W_\mathrm{n})$ *is defined by (3.16)-(3.17), where the definitions in (3.12)-(3.14) 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*

$$n^{1/2}(\zeta^n - bu) \to \bar{\eta} \;\; in\ (D, M_1), and\ \ n^{1/2}(\mu^n - b) \to \bar{\mu} \;\; in\ \mathcal{R}\ as\ n \to \infty, \tag{3.18}$$

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

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 (3.10). Corollary 13.3.2 of [9] plus Proposition 5 then imply that

$$T_n = n^{-1/2}(S^n \circ (n^{-1}N^n) - nbe) = S_n \circ (n^{-1}N^n) + N_n = S_n \circ \big(n^{-1/2}N_n + be\big) + N_n$$

$$\Rightarrow c_s B \circ (be) + \bar{N} + \bar{v} \;= \sqrt{b}\, c_s B + \bar{N} + \bar{v} \equiv \bar{T} + \bar{v}\ in\ (D, M_1). \tag{3.19}$$

Using (3.18) and (3.19)

$$X_n \equiv n^{-1/2}X^n = n^{-1/2}(T^n - n\mu^n e) = n^{-1/2}(T^n - nbe) - n^{-1/2}(n\mu^n e - nbe)$$

$$= \; T_n - n^{1/2}(\mu^n - b)e \;\Rightarrow \bar{T} + \bar{v} - \bar{\mu}e\ in\ (D, M_1). \tag{3.20}$$

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 $\{V_k^n : k \geq 1\}$ 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 Section 4.3 of [9] 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 [12]. The extension is also discussed in Section 2.4. of the Internet Supplement to [9]. 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. ■

3.4 Asymptotic Approximation for the Prelimit Sequence

In classical HTLTs for queues with stationary arrival and service process, each elements of the prelimit sequence depends on a constant traffic intensity. Classical HTLTs are commonly applied to obtain approximations as a function of traffic intensity; see Section 5.5.2 of [9] for an example. The approximation for a particular traffic intensity is not necessarily accurate, but the HTLT assures that it will be accurate for any traffic intensity sufficiently close to one. The limit theorem therefore provides the qualitative criteria that the traffic intensity should be near one for the approximation to be accurate. The dependence structure for a stationary queue's arrival and service process can be complicated, but the HTLT shows that only their asymptotic variances are relevant in sufficiently heavy traffic.

For the HTLT in Theorem 2, the elements of the prelimit sequence each depend on the parameter triple for the arrival process as well as on the squared coefficient of variation of service requirements and the service rate. Our goal is to apply the HTLT to obtain an approximation that is a function of those parameters. As for classical HTLTs, the approximation will not necessarily be accurate for particular choices of those parameters, but we can interpret the HTLT to provide qualitative criterial for the choices. The dependence structure for the arrival process is complicated, but the limit theorem shows that only the asymptotic covariance structure, as expressed by the parameters of the $\psi - GMP$ appearing in the limit, is relevant in sufficiently heavy traffic.

In order to develop approximations that depend on the parameter triples of the converging processes, we want to replace the unspecified function $\bar{v}$ in the limit from Theorem 2 by a function depending directly on the parameter triple. We provide asymptotic justification for that step now. In Section 3.5, 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 2, let

$$\acute{X}^n \equiv n\mathrm{Z}^n - n\mu^n e + n^{1/2}\bar{T} \;\; and \;\; \acute{W}^n \equiv \phi\left(\acute{X}^n, \acute{W}^n(0)\right) \; for \; n \geq 1, \tag{3.21}$$

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

$$d_{M_1}\left(\left(n^{-1/2}X^n, n^{-1/2}W^n\right), \left(n^{-1/2}\acute{X}^n, n^{-1/2}\,\acute{W}^n\right)\right) \Rightarrow 0 \; in \;\; \mathcal{R} \text{ as } n \to \infty. \tag{3.22}$$

Proof. By Theorem 2,

$$(T_n, X_n, W_\mathrm{n}) \Rightarrow (\bar{T} + \bar{v}, \bar{X}, \bar{W}) \text{ in } (D^3, M_1), \tag{3.23}$$

where $\bar{X} = \bar{v} - \bar{\mu}e + \bar{T}$. Let $\acute{X}_n \equiv n^{-1/2}\acute{X}^n$ and $\acute{W}_n \equiv n^{-1/2}\acute{W}^n$. Applying (3.7), (3.9), and (3.18),

$$\begin{aligned}\acute{X}_n &= n^{-1/2}\left(nbM^n - n\mu^n e + n^{1/2}\bar{T}\right) \\ &= bn^{1/2}\left(M^n - e\right) - n^{1/2}(\mu^n - b)e + \bar{T}\end{aligned}$$

$$\Rightarrow \bar{v} - \bar{\mu}e + \bar{T} = \bar{X} \text{ in } (D, M_1). \tag{3.24}$$

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

$$\left(\bar{T}, \acute{X}_n, \acute{W}_n\right) \Rightarrow (\bar{T}, \bar{X}, \bar{W}) \; in \; (D^3, M_1). \tag{3.25}$$

By (3.23), $T_n - \bar{v} \Rightarrow \bar{T} \; in \; (D, M_1)$. We apply the Skorohod representation theorem from Theorem 3.2.2 of [9] to obtain $d_{M_1}(T_n - \bar{v}, \bar{T}) \Rightarrow 0$ in $\mathcal{R}$. The convergence together theorem from Theorem 11.4.7 of [9] then implies that

$$(T_n - \bar{v}, \bar{T}) \Rightarrow (\bar{T}, \bar{T}) \; in \; (D^2, M_1). \tag{3.26}$$

Since X_n and W_n are functions of T_n, and $\acute{X}_n$ and $\acute{W}_n$ are functions of $\bar{T}$, we obtain

$$\left(n^{-1/2}X^n, n^{-1/2}\acute{X}^n, n^{-1/2}W^n, n^{-1/2}\, \acute{W}^n\right) \Rightarrow (\bar{X}, \bar{X}, \bar{W}, \bar{W}) \; in \; (D^4, M_1) \tag{3.27}$$

using (3.23), (3.25), (3.26), and the continuous mapping theorem. The conclusion in (3.22) is then a consequence of (3.27) and the converse of the convergence-together theorem in Theorem 11.4.8 of [9]. ■

According to (3.22),

$$(X^n \equiv N^n - n\mu^n e, W^n) \approx \left(\acute{X}^n \equiv n\text{Z}^n - n\mu^n e + n^{1/2}\bar{T}, \acute{W}^n\right) \tag{3.28}$$

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, $n^{1/2}\bar{T}$ is the zero-mean Gaussian process with distribution determined by $Cov\left[n^{1/2}\bar{T}(s), n^{1/2}\bar{T}(t)\right] = nbs(1 + c_s^2 + \gamma t)$ for $0 \leq s \leq t$. By (3.3), it is therefore the $\psi - GMP$ with parameter pair $(\alpha^*, \beta^*) = (nb(1 + c_s^2), -nb\gamma)$. 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 (3.28) for any particular n by substituting $\lambda(t) \equiv n\zeta^n(t),\ \Lambda(t) \equiv n\mathrm{Z}^n(t), \beta \equiv nb$, $\mu \equiv n\mu_n$, and $\bar{\bar{T}}(t) \equiv n^{1/2}\bar{T}(t)$. With those substitutions, (3.28) becomes

$$\left(X \equiv N - \mu e, W \equiv \phi(X, W(0))\right) \overset{d}{\approx} \left(\acute{X} \equiv \Lambda - \mu e + \bar{\bar{T}}, \acute{W} \equiv \phi\left(\acute{X}, \acute{W}(0)\right)\right), \tag{3.29}$$

where N is then the GPP with parameter triple $(\lambda(t), \gamma, \beta)$, $\bar{\bar{T}}$ is the $\psi - GMP$ with parameter pair $(\alpha^*, \beta^*) = (\beta + \beta c_s^2, -\beta\gamma)$, and μ is the service rate. The parameters β, μ, $\lambda(t)$ 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 approximation obtained from classical HTLTs, the approximation in (3.29) 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. In applying (3.29), we can reduce the number of parameters by one by assuming that $\beta = \mu$. In that case, the criterion is that $\lambda(t) \approx \mu$.

3.5 The Transient Distribution

According to the results in Section 3.4, 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 [13], [14], and [15]. The transient distribution of a reflected $\psi - GMP$ with constant drift was derived in [4] and applied in [3] 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 Section 4. Lemma 1 restates a result from (30) in [3] 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 [3] and originally derived in [1]. The lemmas are applied using the memoryless property of the reflection map from Proposition 10 on page 21 of [11].

According to (3.29), if $P\big(\acute{W}(0) = w_0\big) = 1$, then

$$\big(X(s), W(s), W(s+t)\big) \approx \big(\acute{X}(s), \acute{W}(s), \acute{W}(s+t)\big)\; for\; s,t \geq 0, \tag{3.30}$$

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

$$\acute{X}(t) \equiv \int_0^t \lambda(s)ds - \mu t + \bar{\bar{T}}(t) \qquad and \qquad \acute{W}(t) \equiv \phi(\acute{X})\big(t, \acute{W}(0)\big)\;\; for\; t \geq 0, \tag{3.31}$$

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

Then

$$P(W(s+h) \leq w_{s+h} |\, X(s), W(s)) \approx P(\acute{W}(s+h) \leq w_{s+h} |\, \acute{X}(s), \acute{W}(s))\; for\; s,t \geq 0. \tag{3.32}$$

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

Theorem 3. If $\acute{X}$ and $\acute{W}$ are defined as in (3.31), where $\lambda(v) = \lambda(s)$ for $s \leq v \leq s + t$ and $P\big(\acute{W}(0) = w_0\big) = 1$, then

$$P\big(\acute{W}(s+t) \leq w_{s+t}\big|\, \acute{X}(s) = x_s, \acute{W}(s) = w_s\big) = F(t, w_{s+t}), \tag{3.33}$$

where

$$F(t,w) = \Phi\left(\frac{w - w_s - \omega_s t}{\sqrt{t(\alpha^* - \beta_s^* t)}}\right) \\ - e^{\frac{-2w(\beta_s^* w - \alpha^* \omega_s)}{\alpha^{*2}}}\, \Phi\left(\frac{(2\beta_s^* w - \alpha^* \omega_s)t - \alpha^*(w + w_s)}{\alpha^* \sqrt{t(\alpha^* - \beta_s^* t)}}\right), \tag{3.34}$$

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

$$\alpha^* = \beta(1 + c_s^2),\ \ \beta_s^* \equiv \frac{-\beta\gamma(1 + c_s^2)}{1 + c_s^2 + \gamma s}, and\ \omega_s \equiv \lambda(s) - \mu + \frac{\gamma\left(x_s - \left(\int_0^s \lambda(v)dv - \mu s\right)\right)}{1 + c_s^2 + \gamma s}. \tag{3.35}$$

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 $\acute{X}_s(h) \equiv \big(d_s\acute{X}(h)\big|\acute{X}(s) = x_s\big)$ for $0 \leq h \leq t$. By the memoryless property of the reflection map from Proposition 10 on page 21 of [11], $\acute{W}_s \equiv \phi\big(w_0, d_s\acute{X}\big)$ with probability 1. Then, with probability 1,

$$\big(\acute{W}_s|\, \acute{X}(s) = x_s, \acute{W}(s) = w_s\big) = (\phi(w_s, d_s\acute{X})|\, \acute{X}(s) = x_s, \acute{W}(s) = w_s) \\ = (\phi(w_s, d_s\acute{X})|\, \acute{X}(s) = x_s) = \phi\big(w_s, \acute{X}_s\big) \tag{3.36}$$

where the next-to-last equality holds by the Markov property of $\acute{X}$, the definition of $\acute{W}$, and the assumption that $P\big(\acute{W}(0) = w_0\big) = 1$ (which implies that $\acute{X}$ is independent of $\acute{W}(0)$). By Lemma 2 in Section 4, $\acute{X}_s$ in the final expression of (3.36) is a $\psi - GMP$ on $[0, t]$ with parameter pair (α^*, β_s^*) and drift ω_s. The result in (3.33)-(3.34) then follows from Lemma 1 in Section 6 with the substitutions there of $\alpha^* = \beta,\ \beta^* = \beta_s^*$ and $\omega = \omega_s$ in (3.35). ∎

Theorem 3 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 3 applies when the arrival process is a $\psi^k - GPP$.

Corollary 6. If $\acute{X}$ and $\acute{W}$ are defined as in (3.31), where $P(\acute{W}(0) = w_0) = 1$ and $\lambda(t) = \lambda_i u(t)$ for $t_{i-1} \le t < t_i$ and $1 \le \mathrm{i} \le \mathrm{k} \le \infty$, and if $t_{i-1} \le s < s + t < t_i$ *for some i, then* (3.33)-(3.35) *hold, where*

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

4 Lemmas for Theorem 3

We state and prove two lemmas used in the proof of Theorem 3. The lemmas are discussed in Section 3.5.

A zero-mean real-valued Gaussian process $\{A(t): 0 \le t < \mathrm{T}\}$ is defined in [4] 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

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

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

The first lemma is a special case of Theorem 5 from [4]. 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 (3.15).

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

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

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

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

Let

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

Then, B is standard Brownian motion because it is a zero-mean Gaussian process with $Cov[B(s), B(t)] = s$ for $s \le t$; see page 184 of Adler [16] 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 $s \ge 0$, it follows from (4.4) that

$$
\begin{aligned}
\underset{0\le s\le t}{inf}\, Y(s) \le y \Leftrightarrow & \underset{0\le s\le t}{inf}\left(\frac{\omega\alpha^{*}\mathrm{s} + \alpha^{*}B(s) + (w_0 - y)(1 + s\beta^{*})}{1 + s\beta^{*}}\right) \le 0 \\
\Leftrightarrow & \underset{0\le s\le t}{inf}\big(\omega\alpha^{*}\mathrm{s} + \alpha^{*}B(s) + (w_0 - y)(1 + s\beta^{*})\big) \le 0 \\
\Leftrightarrow & \underset{0\le s\le t}{inf}\,\big(\eta\mathrm{s} + \alpha^{*}B(s)\big) \le y - w_0 \qquad (4.5)
\end{aligned}
$$

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

By (4.4)-(4.5),

$$
\begin{aligned}
G(x, y) &\equiv P\left(Y(t) \le x,\ \underset{0\le s\le t}{inf}\, Y(s) \le y\right) \\
&= \int_{-\infty}^{y-w_0}\int_{b}^{(x-y)(1+t\beta^{*})+y-w_0} P\left(\eta\mathrm{t} + \alpha^{*}B(t)\ \in \mathrm{d}a,\ \underset{0\le s\le t}{inf}\big(\eta\mathrm{s} + \alpha^{*}B(s)\big) \in \mathrm{d}b\ \right) \qquad (4.6)
\end{aligned}
$$

Applying the Change of Measure Theorem on page 10 of [11] followed by the Reflection Principle on pages 7-9 of [11], we obtain

$$
\begin{aligned}
&P\left(\eta\mathrm{t} + \alpha^{*}B(t)\ \in \mathrm{d}a,\ \underset{0\le s\le t}{inf}\big(\eta\mathrm{s} + \alpha^{*}B(s)\big) \in \mathrm{d}b\ \right) \\
&\qquad = exp\left(\frac{\eta a}{{\alpha^{*}}^{2}} - \frac{\eta^{2}t}{2{\alpha^{*}}^{2}}\right) P\left(\alpha^{*}B(t)\ \in \mathrm{d}a,\ \underset{0\le s\le t}{inf}\big(\alpha^{*}B(s)\big) \in \mathrm{d}b\ \right) \\
&\qquad = ex\,p\left(\frac{\eta a}{{\alpha^{*}}^{2}} - \frac{\eta^{2}t}{2{\alpha^{*}}^{2}}\right)\frac{\sqrt{2}(a - 2b)exp\left(\frac{(a - 2b)^{2}}{2{\alpha^{*}}^{2}t}\right)da\ db}{\sqrt{\pi}{\alpha^{*}}^{3}t^{3/2}}. \qquad (4.7)
\end{aligned}
$$

Using (4.6) and (4.7),

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

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

$$\frac{d}{dw}F^*\left(\frac{t\alpha^*}{1+t\beta^*},w\right)=\frac{d}{dw}\left(P\left(Y(t)-\inf_{0\le s\le t}Y(s)\le w,\ \inf_{0\le s\le t}Y(s)\le 0\right)+\ P\left(Y(t)\le w,\ \inf_{0\le s\le t}Y(s)>0\right)\right)$$

$$=\frac{d}{dw}\left(\int_{-\infty}^{0}\int_{y}^{w+y}g(x,y)dx\,dy+\int_{0}^{w_0}\int_{y}^{w}g(x,y)dx\,dy\right)=\int_{-\infty}^{0}g(w+y,y)dy+\int_{0}^{w_0}g(w,y)dy\,. \tag{4.9}$$

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

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

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 $\psi-GMP$ with time-dependent drift conditional on its history.

Lemma 2. Let $\acute{X}(t) \equiv \Lambda(t) - \mu t + \bar{\bar{T}}(t)$ *for* $t \geq 0$ *where* μ *is real,* $\Lambda(t) = \int_0^t \lambda(v)dv$ *for* $\lambda(v)$ *real and integrable, and* $\bar{\bar{T}}$ *is the* $\psi - GMP$ with parameter pair $(\alpha^*, \beta^*) = (\beta + \beta c_s^2, -\beta\gamma)$. *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 $\psi - GMP$ *on* $[0, T)$ with parameter pair (α^*, β_s^*) and constant drift ω_s, where

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

Proof. Under the assumptions,

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

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

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

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, e.g., from Section 6.2.2 of [17], to obtain

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

and

$$\begin{aligned} Cov\left[\acute{X}_s(t), \acute{X}_s(u)\right] &= Cov[\acute{X}(t+s) - \acute{X}(s), \acute{X}(u+s) - \acute{X}(s) | \, \acute{X}(s) = x_s] \\ &= Cov[\acute{X}(t+s), \acute{X}(u+s) | \, \acute{X}(s) = x_s] \end{aligned}$$

$$= \Gamma(t+s, u+s) - \Gamma(s, u+s)\Gamma(s,s)^{-1}\Gamma(s, t+s) = t(\alpha^* - \beta_s^* u)$$

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

REFERENCES

[1] T. H. Cha, "Characterization of the generalized Polya process and its applications," *Advances in Applied Probability,* vol. 46, no. 4, pp. 1148-1171, 2014.

[2] T. H. Konno, "On the exact solution of a generalized Polya process," *Advances in Mathematical Physics,* vol. Article ID 504267, 2010.

[3] K. W. Fendick and W. Whitt, "Queues with path-dependent arrival processes," *Journal of Applied Probability,* To appear.

[4] K. Fendick, "Brownian motion minus the independent increments: representation and queuing application," *Probability in the Engineering and Informational Sciences,* To appear.

[5] W. Whitt, "Heavy-traffic limits for a single-server queue leading up to a critical point," *Operations Research Letters,* vol. 44, pp. 796-800, 2016.

[6] H. Honnappa, R. Jain and A. Ward, "A queueing model with independent arrials and its fluid and diffusion limits," *Queueing Systems,* vol. 80, pp. 71-103., 2015.

[7] D. J. Daley and D. Vere-Jones, An Introduction to the Theory of Point Processes: Volume I; Elementary Theorems and Methods, Second Edition, New York: Springer, 2003.

[8] W. Feller, An Introduction to Probability Theory and its Applications, Vol. 1, 3rd edn, New York: John Wiley, 1968.

[9] W. Whitt, "Stochastic-Process Limits," Springer, New York, 2002.

[10] G. Pang and W. Whitt, "Continuity of a queuing integral representation in the M1 topology," *Annals of Applied Probability,* vol. 1, no. 1, pp. 214-237, 2010.

[11] J. M. Harrison, Brownian Motion and Stochastic Flow Systems, New York: Wiley, 1985.

[12] K. R. Parthasarathy, Probability Measures on Metric Spaces, New York: Academic Press, 1967.

[13] G. F. Newell, Applications of Queuing Theory, 2nd Edition, New York: Chapman and Hall, 1982.

[14] J. Abate and W. Whitt, "Transient behavior of Regulated Brownian Motion, I: starting at the origin," *Advances in Applied Probability,* vol. 19, no. 3, pp. 560-598, 1987.

[15] J. Abate and W. Whitt, "Transient behavior of Regulated Brownian Motion, II: non-zero initial conditions," *Advances in Applied Probability,* vol. 19, no. 3, pp. 599-631, 1987.

[16] R. J. Adler, The geometry of random fields, Chichester: Wiley, 1981.

[17] S. Puntanen and G. P. Styan, "Schur complements in statistics and probability," in *Schur Complement and Its Applications, Numerical Methods and Algorithms, Volume 4*, U.S., Springer, 2005, pp. 163-226.