Available online at www.sciencedirect.com

ScienceDirect

Stochastic Processes and their Applications xxx (xxxx) xxx

stochastic processes and their applications

www.elsevier.com/locate/spa

A heavy-traffic perspective on departure process variability

Peter W. Glynn[a], Rob J. Wang[b,*]

[a] *Department of Management Science and Engineering, Stanford University, 475 Via Ortega, Stanford, CA, 94305, USA*
[b] *Block, Inc., 1455 Market St Unit 600, San Francisco, CA, 94103, USA*

Received 3 May 2021; received in revised form 17 December 2022; accepted 4 January 2023
Available online xxxx

Abstract

This paper studies the departure process from a single-server queue in heavy-traffic over time scales that are of diffusion time scale, and over time scales that are both shorter and longer than diffusion time scale. In addition, the paper shows how one can compute the variance of such Brownian departure processes using stochastic calculus methods. Furthermore, the paper studies the implications of these results for downstream queues that are fed by such departure processes, and shows that downstream equilibrium congestion depends on upstream departure variability over the downstream queue's characteristic heavy traffic time scale. These results also shed further light on the discontinuity in departure process asymptotic variability that is known as the BRAVO effect.

Keywords: Asymptotic variability; Reflected Brownian motion; Departure processes; BRAVO effect; Heavy-traffic analysis; Tandem networks

1. Introduction

When considering the behavior of a network of queues, the departure process from a station is an object of central interest, since such processes generate the endogenous input traffic to other stations in the network. For example, the Queueing Network Analyzer proposed by Whitt [26] uses two-moment approximations to describe the departure processes from stations in a network to develop numerical approximations to congestion measures within the network. The study of departure processes has a long history, with early contributions by Burke [2], Finch

* Corresponding author.
E-mail addresses: glynn@stanford.edu (P.W. Glynn), robjwang@gmail.com (R.J. Wang).

https://doi.org/10.1016/j.spa.2023.01.002

[9], Chang [3], Reich [22], Daley [6], Daley [7], and Vlach and Disney [25]. In this paper, we will focus on two-moment approximations to the departure process from a single-server station, framed within the context of heavy-traffic limit theory. We will study this departure process both for a single-station network and for a two-station tandem network. The latter network is of particular interest, since it is the simplest network in which one can explore the impact of the first station's departure variability on downstream stations.

Our contribution centers on the behavior in heavy traffic, because such heavy-traffic limit theory arises naturally in the setting of heavily congested networks, and because its associated mathematical theory allows us to develop Brownian approximations for the departure process in which the dependence on the mean and variability coefficients is especially visible. In particular, depending on the time scale at which one is studying the departures, one may see departure limit distributions that depend on the station's arrival and service time distributions only through those of the arrival process (over long time scales), on all the first and second moment parameters of both distributions (over moderate time scales), or only on variability characteristics of the arrival and service time distributions (over short time scales). Our theory is intended to expose the degree to which the choice of time scale affects the way a queue's arrival and service time distribution parameters feed into departure time behavior. In particular, while the departure process of a single station queue in heavy traffic has been studied in "diffusion time scale" by earlier authors (see, for example, Hanbali et al. [12]), our discussion in Sections 3 through 5 also encompasses shorter and longer time scales than those studied previously in the heavy traffic literature.

Another main motivation in this paper is to offer insight into an intriguing phenomenon identified by Nazarathy and Weiss [18] in their study of the variability of the $M/M/1$ queue's departure process. They showed that when the queue is started empty, the cumulative number of departures $N_D(t)$ over $[0, t]$ satisfies

$$\lim_{t\to\infty} \frac{\operatorname{Var} N_D(t)}{E N_D(t)} = \begin{cases} 1, & \lambda \neq \mu \\ 2\left(1 - \frac{2}{\pi}\right), & \lambda = \mu, \end{cases}$$

where λ and μ are the arrival and service rates to the queue, respectively. In view of the fact that the value at the discontinuity is less than 1, this reduction in variability is called the BRAVO effect, for "balancing reduces average variability of outputs". For the general single-server $G/G/1$ queue with independent streams of independent and identically distributed (iid) inter-arrival and service times, Hanbali et al. [12] show that

$$\lim_{t\to\infty} \frac{\operatorname{Var} N_D(t)}{E N_D(t)} = \begin{cases} c_\chi^2, & \lambda < \mu \\ \left(1 - \frac{2}{\pi}\right)(c_\chi^2 + c_V^2), & \lambda = \mu \\ c_V^2, & \lambda > \mu, \end{cases} \tag{1.1}$$

where c_χ^2 and c_V^2 are the squared coefficients of variation of the inter-arrival and service time distributions, respectively. The second limit relation (1.1) makes clear that balancing need not necessarily reduce long-run variability.

Nevertheless, it does point to a discontinuity that is worth understanding better. Of course, the discontinuity appears because of a limit interchange issue, having to do with interchanging the limit in time tending to infinity with the limit in λ tending to μ. This regime is exactly the heavy traffic setting that is studied in this paper. In particular, the various departure process time scales studied in Sections 3 through 5 clarify the settings in which one can expect to see the BRAVO phenomenon. These limit theorems show that BRAVO manifests itself for stable

$G/G/1$ queues only at short time scales and, even then, only if the system is initialized with relatively little work. Furthermore it should be noted that even for $G/G/1$ systems that are in precise balance, the BRAVO variability behavior is seen only when the system is initialized with a small amount of work.

As noted above, Sections 3 through 5 clarify the differing behaviors of the departure process over different time scales, and make clear which system parameters primarily affect the departure behavior over each time scale. However, our paper makes two additional contributions:

1. We show how to use stochastic calculus methods to compute the asymptotic variance parameter for the Brownian limit of the departure process for the perfectly balanced $G/G/1$ queue. This was previously derived by Hanbali et al. [12] via an argument based on using the form of the limit distribution to reduce the calculation to one involving the known distribution of the maximum of a Brownian bridge. Our approach, discussed in Section 6, can be extended to more general Brownian departure process settings. In particular, Theorem 8 provides a new formula for the departure variance parameter when the arrival and service streams are correlated.
2. We show in Section 7 how the theory of Sections 3 through 5 applies to the downstream queue in a two station tandem queueing network. We note that such tandem networks have previously been studied via diffusion limits. In particular, both Harrison [13], Harrison and Shepp [15] take this perspective. A distinguishing feature of this paper is that we adopt a multiscale perspective in which we put both stations into heavy traffic, but allow either the first or second station to be much more congested than the other. To our understanding, this is the first such multiscale analysis for queueing networks. The analysis shows that the downstream queue is affected by departure process variability from the first station that is aligned with the heavy traffic time scale that goes with the second station. This implies, for example, that when the second station is less heavily loaded, its equilibrium depends on the first station's departure process over a time scale that is short as compared to the first station's heavy traffic time scale. This translates into a departure process that behaves as if the server at the first queue is always busy. So, the equilibrium then has no dependence on the parameters of the arrival process to the first station.

We believe this multiscale perspective, illustrated in this paper in the two station tandem network setting, can be useful in other network modeling environments as well.

2. The Brownian approximation for the departure time sequence

We have the choice of studying departures from a queue either via an approximation to the sequence of departure times, or via an approximation to the departure counting process. We choose to study the departure time sequence, because it can be expressed directly in terms of the Skorohod reflection map acting upon the model primitives (i.e. the arrival and service times). In contrast, the departure counting process involves a more complex interaction between the Skorohod reflection map and a random time change corresponding to the cumulative "busyness" process; see, for example, p.146 of Chen and Yao [4]. Consequently, the counting process is more challenging to represent directly in terms of the model primitives. As we shall see, we can recover departure counting process approximations from the departure time sequence.

We now describe the departure time sequence $(D(n) : n \geq 0)$, where $D(n)$ is the departure time of the nth customer to arrive to the system. Clearly,

$$D(n) = A_n + W(n) + V_n, \tag{2.1}$$

where A_n is the arrival time of the nth customer, $W(n)$ is its waiting time (exclusive of service), and V_n is its service time. We set $A_0 = 0$ and recall that for a single-server queue with an infinite capacity waiting room that serves customers according to a first-in/first-out (FIFO) queue discipline,

$$W(n+1) = [W(n) + V_n - \chi_{n+1}]^+$$

for $n \geq 0$, where $[x]^+ = \max(x, 0)$ and $\chi_{n+1} = A_{n+1} - A_n$ is the $(n+1)$'st inter-arrival time. It is well-known that

$$W(n) = S(n) + Y(n), \tag{2.2}$$

where

$$\begin{aligned} S(n) &= W(0) + \sum_{k=1}^{n} [V_{k-1} - \chi_k] \\ &= W(0) + \sum_{j=0}^{n-1} V_j - A_n, \end{aligned}$$

and

$$Y(n) = \max_{1 \leq k \leq n} [-S(k)]^+;$$

see p.96 of Asmussen [1]. We note that the mapping sending $(S(n) : n \geq 0)$ into $((W(n), Y(n)) : n \geq 0)$ is precisely the (discrete time) Skorohod reflection map.

We now wish to study the departure sequence in the "heavy traffic" regime. We start by introducing the parameter $\rho \in \mathbb{R}^+$, the so-called *traffic intensity* of the queue, and scale the service times in the ρth system by ρ. We will also later want to study the joint dependence of the departure times as a function of ρ, n, and the initial condition $W(0) = x$. In view of these considerations, let $D_\rho(n, x)$, $W_\rho(n, x)$, $S_\rho(n, x)$, and $Y_\rho(n, x)$ denote the corresponding values of $D(n)$, $W(n)$, $S(n)$, and $Y(n)$ in the ρth system when it is initialized with $W(0) = x$. In particular,

$$\begin{aligned} D_\rho(n, x) &= A_n + W_\rho(n, x) + \rho V_n, \\ W_\rho(n, x) &= S_\rho(n, x) + Y_\rho(n, x), \\ S_\rho(n, x) &= x + \rho \sum_{j=0}^{n-1} V_j - A_n, \end{aligned}$$

and

$$Y_\rho(n, x) = \max_{1 \leq k \leq n} [-S_\rho(k, x)]^+.$$

We note that $D_\rho(n, x)$ can be re-expressed as

$$D_\rho(n, x) = x + \rho \sum_{j=0}^{n} V_j + Y_\rho(n, x).$$

To obtain our Brownian approximation for the departure time sequence, we shall impose a *strong approximation* assumption on the A_n's and V_n's. In particular, we make the following assumption:

(A1) There exist $r \in (0,\ 1/2)$, $\alpha > 0$, and a probability space supporting a two-dimensional mean zero Brownian motion process $(Z_1,\ Z_2) = ((Z_1(t),\ Z_2(t)) : t \geq 0)$ and a sequence $((V'_{k-1},\ A'_k) : k \geq 1)$ such that:

(i) $((V_{k-1},\ A_k)\ :\ k\ \geq\ 1) \overset{\mathscr{D}}{=} ((V'_{k-1},\ A'_k)\ :\ k\ \geq\ 1)$, where $\overset{\mathscr{D}}{=}$ denotes equality in distribution;

(ii) $\sum_{k=0}^{n-1} V'_k = \alpha n + Z_1(n) + o(n^r)$ a.s. as $n \to \infty$, where $o(n^r)$ denotes a sequence of random variables $(E_n\ :\ n \geq 0)$ for which $E_n/n^r \to 0$ a.s. as $n \to \infty$;

(iii) $A'_n = \alpha n + Z_2(n) + o(n^r)$ a.s. as $n \to \infty$.

In view of the fact that the same "centering constant" αn appears on the right-hand side of both (ii) and (iii) above, it follows that when $\rho = 1$, the arrival rate to the queue is perfectly balanced with its service rate. We therefore refer to $\rho = 1$ as the *balanced* case. Thus, when $\rho < 1$, this should correspond to a *stable* queue, whereas when $\rho > 1$, this corresponds to an *unstable* queue.

Loosely speaking, the strong approximation assumption A1 holds in great generality when the $(V_{j-1},\ \chi_j)$'s evolve in a stationary environment. For example, when $(V_j\ :\ j \geq 0)$ is an iid sequence independent of the iid sequence $(\chi_j\ :\ j \geq 1)$, A1 holds whenever $EV_0^{1/r} + E\chi_1^{1/r} < \infty$; see [16]. In this context, $\alpha = EV_0 = E\chi_1$, $\operatorname{Var} Z_1(1) = \operatorname{Var} V_0$, $\operatorname{Var} Z_2(1) = \operatorname{Var} \chi_1$, and $\operatorname{Cov}(Z_1(1),\ Z_2(1)) = 0$. We shall henceforth refer to this setting as the *fully independent* assumption on the model primitives. We can also allow the $(V_{j-1},\ \chi_j)$'s to be iid, but permit dependence between V_{j-1} and χ_j, provided $r < 1/3$ and $EV_0^{1/r} + E\chi_1^{1/r} < \infty$; see [8]. Here, $\alpha = EV_0 = E\chi_1$, $\operatorname{Var} Z_1(1) = \operatorname{Var} V_0$, $\operatorname{Var} Z_2(1) = \operatorname{Var} \chi_1$, and $\operatorname{Cov}(Z_1(1),\ Z_2(1)) = \operatorname{Cov}(V_0,\ \chi_1)$. assumption A1 also often holds when the χ_j's and V_j's are dependent in j. For example, if $(V_j\ :\ j \geq 0)$ and $(\chi_j\ :\ j \geq 1)$ are independent sequences that can each be described as bounded functionals of geometrically ergodic Markov chains, one can apply the results of Csáki and Csörgő [5] or Merlevède and Rio [17].

We henceforth take the view that our original probability space supporting the A_n's and V_n's also supports Z_1 and Z_2 satisfying A1 (ii) and (iii), so we need not differentiate between $((V'_{k-1},\ A'_k) : k \geq 1)$ and $((V_{k-1},\ A_k) : k \geq 1)$ in the sequel. Let

$$X_\rho(t,\ x) = x - \alpha(1-\rho)t + \rho Z_1(t) - Z_2(t) + L_\rho(t,\ x),$$

where

$$L_\rho(t,\ x) = \max_{0 \leq s \leq t} [-x + \alpha(1-\rho)s - \rho Z_1(s) + Z_2(s)]^+.$$

We note that $(X_\rho(t,\ x)\ :\ t \geq 0)$ is a one-dimensional *reflected Brownian motion* (RBM) satisfying the stochastic differential equation (SDE)

$$dX(t) = -\alpha(1-\rho)dt + \rho dZ_1(t) - dZ_2(t) + dL(t),$$

where $(L(t) : t \geq 0)$ is the local time (at the origin) that is non-decreasing and satisfies

$$\mathbb{I}(X(t) > 0)dL(t) = 0.$$

Note that $(X_\rho(t,\ x) : t \geq 0)$ is the Brownian analog to $(W_\rho(\lfloor t \rfloor,\ x) : t \geq 0)$. This suggests that

$$\tilde{D}_\rho(t,\ x) = x + \rho\alpha t + \rho Z_1(t) + L_\rho(t,\ x)$$

is the natural Brownian approximation to $D_\rho(\lfloor t \rfloor,\ x)$. The result below makes precise the sense in which $(\tilde{D}_\rho(t,\ x) : t \geq 0)$ approximates $(D_\rho(\lfloor t \rfloor,\ x) : t \geq 0)$.

Theorem 1. *Fix $\rho_0 \in (0, \infty)$. Under A1, $D_\rho(n, x) = \tilde{D}_\rho(n, x) + o(n^r)$ a.s. as $n \to \infty$, where the term $o(n^r)$ is uniform in $x \in \mathbb{R}_+$ and $\rho \in [0, \rho_0]$.*

Proof. In view of A1, it is easy to see that

$$D_\rho(n, x) = x + \rho\alpha n + \rho Z_1(n) + \max_{1\le k\le n}[-x + \alpha k(1-\rho) - \rho Z_1(k) + Z_2(k)]^+ + o(n^r) \text{ a.s.} \quad (2.3)$$

as $n \to \infty$. Note that

$$\left|\max_{0\le s\le n}[-x + \alpha s(1-\rho) - \rho Z_1(s) + Z_2(s)]^+ - \max_{0\le k\le n}[-x + \alpha k(1-\rho) - \rho Z_1(k) + Z_2(k)]^+\right|$$
$$\le \alpha|\rho - 1| + \max_{0\le k\le n}\left(\rho \max_{0\le s\le 1}|Z_1(k+s) - Z_1(k)|, \max_{0\le s\le 1}|Z_2(k+s) - Z_2(s)|\right).$$

Since $E\exp\left(\theta \max_{0\le s\le 1}|Z_1(s)|\right) < \infty$ for some $\theta > 0$ (which follows easily from the explicit distributions that are known for

$$\max_{0\le s\le 1} Z_1(s) \text{ and } \max_{0\le s\le 1} -Z_1(s);$$

see, for example, [14] p. 13–14), it follows from Markov's inequality and the Borel–Cantelli lemma that

$$\max_{0\le k\le n}\max_{0\le s\le 1}|Z_1(k+s) - Z_1(k)| = O(\log n) \text{ a.s.}$$

as $n \to \infty$. We similarly conclude that

$$\max_{0\le k\le n}\max_{0\le s\le 1}|Z_2(k+s) - Z_2(k)| = O(\log n) \text{ a.s.}$$

as $n \to \infty$, from which Theorem 1 follows from (2.3). □

The RBM-quantity $\tilde{D}_\rho(n, x)$ has a magnitude of order n, with stochastic fluctuations of order $n^{1/2}$, each of which is of larger order than $o(n^r)$. As a consequence, Theorem 1 ensures that the Brownian approximation to $D_\rho(n, x)$ is suitable for n large, uniformly in x and $\rho \in [0, \rho_0]$.

3. Setting 1: Diffusion time scale

We now develop some consequences of Theorem 1, in the setting that the time scale n corresponds to what is conventionally called the *diffusion time scale*. Suppose $\rho \ne 1$, so that the RBM $X_\rho(\cdot, x)$ has non-zero drift. In the diffusion time scale, the effect of the stochastic volatility in $\rho Z_1 - Z_2$ is roughly of the same order of the magnitude as that induced by the non-zero drift. In particular, we are considering a time scale n in which the volatility, which is of order $n^{1/2}$, is of the same order as the cumulative drift contribution $\alpha n(1-\rho)$, so that n is of the order of $1/(1-\rho)^2$. In this temporal scale, both the stochastic variability and the cumulative effect of the drift are of the order $1/|1-\rho|$.

In addition, the initial condition x plays a significant role. If $x \gg 1/|1-\rho|$, it is unlikely that the RBM $X_\rho(\cdot, x)$ will hit the origin over $[0, n]$ (when n is of the order of $1/(1-\rho)^2$), so the increasing process $L_\rho(\cdot, x)$ will then be identically zero over $[0, n]$. As a consequence, we expect $D_\rho(n, x)$ to then approximately equal $x + \rho\alpha n + \rho Z_1(n)$ in this setting. In other words, the departure process from the queue behaves as if there is an infinite supply of customers present in the queue, so that the Brownian approximation Z_2 to the arrival process is irrelevant in this context. On the other hand, if $x \ll 1/|1-\rho|$, it is likely that the stochastic fluctuations in $\rho Z_1 - Z_2$ will cause the RBM $X_\rho(\cdot, x)$ to hit the origin early in $[0, n]$, so that $D_\rho(n, x)$ will then effectively behave in the same way as $D_\rho(n, 0)$. For x of the order of $1/|1-\rho|$, the departure quantity $D_\rho(n, x)$ will be significantly influenced by the initial condition x.

The following result summarizes this behavior.

Theorem 2. *Assume A1. If $x|1-\rho| \to y \in \mathbb{R}_+$ as $\rho \to 1$, then*

$$|1-\rho|\left(D_\rho\left(\left\lfloor\frac{t}{(1-\rho)^2}\right\rfloor, x\right) - x - \frac{\alpha t}{(1-\rho)^2}\right) \Rightarrow -\alpha t + Z_1(t) + \max_{0\le s\le t}[-y + \alpha s - Z_1(s) + Z_2(s)]^+ \tag{3.1}$$

as $\rho \nearrow 1$ (where $\Rightarrow$ denotes weak convergence) and

$$|1-\rho|\left(D_\rho\left(\left\lfloor\frac{t}{(1-\rho)^2}\right\rfloor, x\right) - x - \frac{\alpha t}{(1-\rho)^2}\right) \Rightarrow \alpha t + Z_1(t) + \max_{0\le s\le t}[-y - \alpha s - Z_1(s) + Z_2(s)]^+ \tag{3.2}$$

as $\rho \searrow 1$. If $x|1-\rho| \to \infty$ as $\rho \to 1$, then

$$|1-\rho|\left(D_\rho\left(\left\lfloor\frac{t}{(1-\rho)^2}\right\rfloor, x\right) - x - \frac{\alpha t}{(1-\rho)^2}\right) \Rightarrow -\alpha t + Z_1(t) \tag{3.3}$$

as $\rho \nearrow 1$ and

$$|1-\rho|\left(D_\rho\left(\left\lfloor\frac{t}{(1-\rho)^2}\right\rfloor, x\right) - x - \frac{\alpha t}{(1-\rho)^2}\right) \Rightarrow \alpha t + Z_1(t) \tag{3.4}$$

as $\rho \searrow 1$.

Remark 1. We note that if $W_\rho(0, x) = x$, then the departure time for the customer arriving at $A_0 = 0$ is $x + V_0$, so all the subsequent departure times are offset by the same x. As a consequence, our limit theorems "center" the departure times by x.

Remark 2. The departure time limit in (3.3) is smaller than that in (3.4), because our limit regime is obtained by scaling the service times by the factor ρ. Consequently, the service times when $\rho \nearrow 1$ are slightly smaller than those arriving when $\rho \searrow 1$.

Proof. We note that for $n = \lfloor t/(1-\rho)^2\rfloor$, the term $o(n^r)$ in Theorem 1 satisfies

$$|1-\rho|o(n^r) = |1-\rho|o((1-\rho)^{-2r}) \to 0 \text{ a.s.}$$

as $\rho \to 1$, so that

$$|1-\rho||D_\rho(\lfloor t/(1-\rho)^2\rfloor, x) - \tilde{D}_\rho(\lfloor t/(1-\rho)^2\rfloor, x)| \to 0 \text{ a.s.} \tag{3.5}$$

uniformly in $x \in \mathbb{R}_+$ and $\rho \in [0, \rho_0]$. The scaling properties of Brownian motion guarantee that

$$|1-\rho|\left(\tilde{D}_\rho\left(\frac{t}{(1-\rho)^2}, x\right) - x - \frac{\alpha t}{(1-\rho)^2}\right) \overset{\mathcal{D}}{=} \begin{cases} -\alpha t + \rho Z_1(t) + \max_{0\le s\le t}[-x|1-\rho| + \alpha s - \rho Z_1(s) + Z_2(s)]^+, & \rho < 1 \\ \alpha t + \rho Z_1(t) + \max_{0\le s\le t}[-x|1-\rho| - \alpha s - \rho Z_1(s) + Z_2(s)]^+, & \rho > 1. \end{cases} \tag{3.6}$$

Sending ρ to 1 and applying (3.5) then yields (3.1) and (3.2). On the other hand, if $x|1-\rho| \to \infty$,

$$\max_{0\le s\le t}[-x|1-\rho| - \alpha s - \rho Z_1(s) + Z_2(s)]^+ \to 0$$

as $\rho \to 1$, so that (3.5) and (3.6) together imply (3.3) and (3.4). □

When ρ is exactly 1 (so that the queue is precisely balanced), the drift in $\rho Z_1 - Z_2$ is exactly zero. In this case, the analog to Theorem 2 is our next result.

Theorem 3. *Assume A1. If $xn^{-1/2} \to y \in \mathbb{R}_+$ as $n \to \infty$, then*

$$n^{-\frac{1}{2}}(D_1(n, x) - x - \alpha n) \Rightarrow Z_1(1) + L_1(1, y)$$

as $n \to \infty$, while if $xn^{-1/2} \to \infty$ as $n \to \infty$, then

$$n^{-\frac{1}{2}}(D_1(n, x) - x - \alpha n) \Rightarrow Z_1(1)$$

as $n \to \infty$.

We note that in the balanced case,

$$n^{-1/2}(D_1(n, x) - \tilde{D}_1(n, x)) = n^{-1/2}o(n^r) \to 0 \text{ a.s.}$$

uniformly in $x \in \mathbb{R}_+$. The proof of Theorem 3 then follows immediately from the scaling properties of Brownian motion.

When $\rho = 1$ and the queue is started empty (so $y = 0$), the Brownian approximation to $D_\rho(n, 0)$ has the scaling property

$$\begin{aligned} Z_1(t) + L_1(t, 0) &= \max_{0\le s\le t}(Z_2(s) + Z_1(t) - Z_1(s)) \\ &\overset{\mathscr{D}}{=} \sqrt{t}\max_{0\le s\le 1}(Z_2(s) + Z_1(1) - Z_1(s)). \end{aligned} \tag{3.7}$$

This is precisely the rv that arises in the perfectly balanced setting analyzed by Hanbali et al. [12], and within the BRAVO literature.

We conclude this section with a brief discussion of how to recover a Brownian departure counting process approximation from a departure time approximation, as promised in Section 2. We start by noting that the cumulative number of departures of customers indexed by $n \ge 0$ up to time $W_\rho(0, x)$ is zero. So, the departure counting process is interesting only for times $s \ge W_\rho(0, x) = x$. So, for $s \ge x$, consider

$$\varDelta_\rho(s, x) = \max\{n \ge 0 : D_\rho(n, x) \le s\}.$$

In what follows, we will consider $\varDelta_\rho(x + t/(1-\rho)^2, x)$, so that time is again measured in the diffusion time scale.

Theorem 4. *Assume A1. If $x|1-\rho| \to y \in \mathbb{R}_+$ as $\rho \to 1$, then*

$$\begin{aligned} &|1-\rho|\left(\varDelta_\rho\left(x + \frac{t}{(1-\rho)^2}, x\right) - \frac{t}{\alpha(1-\rho)^2}\right) \\ &\Rightarrow \frac{t}{\alpha} - \alpha^{-\frac{3}{2}}Z_1(t) - \max_{0\le s\le t}\left[-\frac{y}{\alpha} + \frac{s}{\alpha} - \alpha^{-\frac{3}{2}}Z_1(s) + \alpha^{-\frac{3}{2}}Z_2(s)\right]^+ \end{aligned} \tag{3.8}$$

as $\rho \nearrow 1$, *and*

$$|1-\rho|\left(\Delta_\rho\left(x+\frac{t}{(1-\rho)^2},\, x\right)-\frac{t}{\alpha(1-\rho)^2}\right)$$
$$\Rightarrow -\frac{t}{\alpha}-\alpha^{-\frac{3}{2}}Z_1(t)-\max_{0\le s\le t}\left[-\frac{y}{\alpha}-\frac{s}{\alpha}-\alpha^{-\frac{3}{2}}Z_1(s)+\alpha^{-\frac{3}{2}}Z_2(s)\right]^+ \tag{3.9}$$

as $\rho \searrow 1$. *If* $x|1-\rho| \to \infty$ *as* $\rho \to 1$, *then*

$$|1-\rho|\left(\Delta_\rho\left(x+\frac{t}{(1-\rho)^2},\, x\right)-\frac{t}{\alpha(1-\rho)^2}\right)\Rightarrow \frac{t}{\alpha}-\alpha^{-\frac{3}{2}}Z_1(t) \tag{3.10}$$

as $\rho \nearrow 1$ *and*

$$|1-\rho|\left(\Delta_\rho\left(x+\frac{t}{(1-\rho)^2},\, x\right)-\frac{t}{\alpha(1-\rho)^2}\right)\Rightarrow -\frac{t}{\alpha}-\alpha^{-\frac{3}{2}}Z_1(t) \tag{3.11}$$

as $\rho \searrow 1$.

Proof. For the purposes of this argument, we abbreviate $\Delta_\rho(x+t/(1-\rho)^2,\, x)$ as Δ_ρ. Then,

$$D_\rho(\Delta_\rho,\, x)\le x+\frac{t}{(1-\rho)^2}\le D_\rho(\Delta_\rho+1,\, x)$$

for $t \ge 0$. It follows from A1 that

$$D_\rho(\Delta_\rho,\, x)=x+\frac{t}{(1-\rho)^2}+o(\Delta_\rho^r)\ \text{ a.s.} \tag{3.12}$$

as $\rho \to 1$. As a consequence of Theorem 1 and the fact that $|Z_i(t)| = O(t^{2/3})$ a.s. as $t \to \infty$, we find that

$$x+\frac{t}{(1-\rho)^2}=x+\rho\alpha\Delta_\rho+O(\Delta_\rho^{\frac{2}{3}})\ \text{ a.s.}$$

as $\rho \to 1$, so that

$$(1-\rho)^2\Delta_\rho \to \frac{t}{\alpha}\ \text{ a.s.}$$

as $\rho \to 1$.

Theorem 1 and (3.12) then imply that if $\rho \nearrow 1$,

$$\begin{aligned}x+\frac{t}{(1-\rho)^2} &= x+\rho\alpha\Delta_\rho+\rho Z_1\left(\frac{t}{\alpha(1-\rho)^2}\right)\\ &\quad+\max_{0\le s\le t}\left[-x+\frac{s}{1-\rho}-\rho Z_1\left(\frac{s}{\alpha(1-\rho)^2}\right)+Z_2\left(\frac{s}{\alpha(1-\rho)^2}\right)\right]^+\\ &\quad+o((1-\rho)^{-2r})\ \text{ a.s.}\end{aligned}$$

and hence

$$\begin{aligned}(1-\rho)\left(\Delta_\rho-\frac{t}{\alpha\rho(1-\rho)^2}\right) &= -\frac{\rho(1-\rho)}{\alpha}Z_1\left(\frac{t}{\alpha(1-\rho)^2}\right)-\frac{1}{\alpha}\\ &\cdot \max_{0\le s\le t}\left[-x(1-\rho)+s-\rho(1-\rho)Z_1\left(\frac{s}{\alpha(1-\rho)^2}\right)\right.\\ &\left.+\ (1-\rho)Z_2\left(\frac{s}{\alpha(1-\rho)^2}\right)\right]^+ +o((1-\rho)^{-2r}).\end{aligned}$$

The scaling properties of $Z_1(\cdot)$ and $Z_2(\cdot)$ then immediately yield (3.8), upon recognizing that

$$\frac{t}{\rho(1-\rho)} = \frac{t}{1-\rho} + t + o(1)$$

as $\rho \nearrow 1$. The proofs of (3.9) through (3.11) follow a similar argument. □

4. Setting 2: Short time scales

In this section, we study the departure time sequence in time scales shorter than diffusion time scale. In particular, when $\rho \neq 1$, we consider time scales n in which n is large, ρ is close to 1, but $n \ll (1-\rho)^{-2}$. In this setting, the magnitude of the cumulative drift $\alpha n(1-\rho)$ is then small relative to the stochastic variability $n^{1/2}$ that is present in the pre-limit and Brownian approximation. As a consequence, the limit processes that arise in this setting all essentially involve RBM with zero drift, and the result closely resembles Theorem 3.

Theorem 5. *Assume A1 and suppose that $n \to \infty$ with $n(1-\rho)^2 \to 0$. If $xn^{-1/2} \to y \in \mathbb{R}_+$ as $n \to \infty$, then*

$$n^{-\frac{1}{2}}(D_\rho(n, x) - x - \alpha n) \Rightarrow Z_1(1) + L_1(1, y)$$

as $n \to \infty$, while if $xn^{-1/2} \to \infty$ as $n \to \infty$, then

$$n^{-\frac{1}{2}}(D_\rho(n, x) - x - \alpha n) \Rightarrow Z_1(1)$$

as $n \to \infty$.

The proof of this result closely resembles that of Section 3, so is omitted. We note that when $y = 0$, this is precisely the limit rv associated with the BRAVO literature.

5. Setting 3: Long time scales

We now consider the remaining case, in which the time n is long compared to diffusion time scale, so that $n \gg (1-\rho)^{-2}$ when $\rho \neq 1$. In this setting, the magnitude of the cumulative drift $\alpha n(1-\rho)$ is large compared to the stochastic variability $\sqrt{n}$ that is present at time n. For a stable queue with $\rho < 1$ and initial work x, the drift will empty the queue at a time roughly of order $x/((1-\rho)\alpha)$. Hence, if $x/((1-\rho)\alpha)$ is small enough relative to n, the system will then empty repeatedly prior to time n, so that the queue will effectively be in equilibrium at time n. In this case, we expect the departure characteristics to be those associated with the arrival process, as is characteristic of a stable queue. On the other hand, if $x/((1-\rho)\alpha)$ is large relative to n, then the likelihood that the queue will empty prior to n is small, so that the departure characteristics then should be those associated with a queue that effectively starts with infinite work, so that the service time process dominates the departure behavior.

For an unstable queue with $\rho > 1$, the effect of the cumulative drift $\alpha n(\rho - 1)$ will be much larger than the stochastic variability $n^{1/2}$, so that any emptying time of the system would have to occur early in $[0, n]$. As a consequence, the queue's departure process again behaves as if the system effectively started with infinite work, so that the service time process again dominates the departure behavior.

Our next theorem makes rigorous this discussion.

Theorem 6. *Assume A1 with $n(1-\rho)^2 \to \infty$ as $n \to \infty$.*

(a) If $\rho \nearrow 1$ as $n \to \infty$ and $\overline{\lim}_{n\to\infty} x/(n(1-\rho)) < \alpha$, then

$$n^{-\frac{1}{2}}(D_\rho(n, x) - x - \alpha n) \Rightarrow Z_2(1) \tag{5.1}$$

as $n \to \infty$.

(b) If $\rho \nearrow 1$ as $n \to \infty$ and $\underline{\lim}_{n\to\infty} x/(n(1-\rho)) > \alpha$, then

$$n^{-\frac{1}{2}}(D_\rho(n, x) - x - \rho\alpha n) \Rightarrow Z_1(1) \tag{5.2}$$

as $n \to \infty$.

(c) If $\rho \searrow 1$ as $n \to \infty$, then

$$n^{-\frac{1}{2}}(D_\rho(n, x) - x - \rho\alpha n) \Rightarrow Z_1(1) \tag{5.3}$$

as $n \to \infty$.

Remark 3. We observe that over long time scales, the departure time sequence inherits either the behavior of the arrival sequence (when the system starts with a moderate amount of work) or that of the service time sequence (when the system starts with a substantial amount of work). In no long time scale setting does the BRAVO limit (3.7) arise.

Proof. We start by noting that

$$D_\rho(n, x) - \alpha n = Z_2(n) + X'_{\rho-1}(n, x) + o(n^{\frac{1}{2}}) \text{ a.s.} \tag{5.4}$$

as $n \to \infty$, where $X'_\mu(\cdot, z)$ is an RBM with initial position z that satisfies

$$X'_\mu(t, z) = z + \mu\alpha t + Z_1(t) - Z_2(t) + \max_{0\le s\le t}[-z - \mu\alpha s - Z_1(s) + Z_2(s)]^+. \tag{5.5}$$

Then, (5.1) follows if $X'_{\rho-1}(n, x)/\sqrt{n} \Rightarrow 0$ as $n \to \infty$.

Due to the scaling properties of Brownian motion, it is evident that

$$X'_{\rho-1}(n, z) \overset{\mathscr{D}}{=} \frac{1}{1-\rho} X'_{-1}(n(1-\rho)^2, z(1-\rho)), \tag{5.6}$$

where $X'_{-1}(\cdot, z)$ has drift $-\alpha$. Since X'_{-1} is a positive recurrent diffusion, there exists a finite-valued random variable $X'_{-1}(\infty)$ such that for each $z \ge 0$,

$$X'_{-1}(t, z) \Rightarrow X'_{-1}(\infty) \tag{5.7}$$

as $t \to \infty$. Hence, if $x = O((1-\rho)^{-1})$, (5.6) immediately implies that $\frac{1}{\sqrt{n}} X'_{\rho-1}(n, x) \Rightarrow 0$ as $n \to \infty$. On the other hand, if $x(1-\rho) \to \infty$ with $\overline{\lim}\, x/(n(1-\rho)) < \alpha$, we note that $X'_{-1}(\cdot, x(1-\rho))$ then hits the origin prior to time $n(1-\rho)^2(1-\delta)$ for some $\delta > 0$ with probability converging to 1 as $n \to \infty$, by virtue of the explicit distribution for the first hitting times of Brownian motion with drift; see, for example, [14] p.14. It follows from the strong Markov property applied at the hitting time, and (5.7), that $X'_{\rho-1}(n, x)/\sqrt{n} \Rightarrow 0$ as $n \to \infty$.

For (5.2), we note that the emptying time of $X'_{-1}(\cdot, x(1-\rho))$ will, with probability converging to 1, occur after $n(1-\rho)^2$ when

$$\underline{\lim}_{n\to\infty} \frac{x}{n(1-\rho)} > \alpha.$$

Hence, with probability converging to 1,

$$D_\rho(n, x) = \alpha n + Z_2(n) + x + (\rho - 1)\alpha n + Z_1(n) - Z_2(n) + o(n^{\frac{1}{2}})$$

$$= x + \rho\alpha n + Z_1(n) + o(n^{\frac{1}{2}})$$

as $n \to \infty$, yielding (5.2).

For (5.3), (5.4) remains valid, with (5.6) replaced by

$$X'_{\rho-1}(n, z) \overset{\mathscr{D}}{=} \frac{1}{\rho - 1} X'_1(n(\rho-1)^2, z(\rho-1)).$$

Since $X'_1(\cdot, z)$ is now an RBM with positive drift α, it is evident that $X'_1(\cdot, z)$ either never visits the origin, or last visits at some finite time. Hence,

$$X_1(t, z) = z + \alpha t + Z_1(t) - Z_2(t) + O(1) \text{ a.s.}$$

as $t \to \infty$. Consequently, we find that

$$D_\rho(n, x) = x + \rho\alpha n + Z_1(n) + o_p(n^{\frac{1}{2}})$$

as $n \to \infty$, yielding (5.3). □

6. Computing moments for the Brownian departure process

We now show how stochastic calculus can be used to compute various moments of the Brownian departure process. All of the limits derived in Sections 3 through 5 can be expressed in terms of the process

$$D(t) = m_A t + Z_2(t) + X(t) - X(0),$$

where $(X(t) : t \geq 0)$ is an RBM for which $X(0) = x$ and satisfies

$$dX(t) = mdt + dZ(t) + dL(t),$$

where L satisfies $\mathbb{I}(X(t) > 0)dL(t) = 0$ and $(Z(t) : t \geq 0)$ is a mean zero Brownian motion that can be expressed as

$$Z(t) = Z_1(t) - Z_2(t),$$

with $\operatorname{Var} Z_1(1) = \sigma_S^2$, $\operatorname{Var} Z_2(1) = \sigma_A^2$, and $\operatorname{Cov}(Z_1(1), Z_2(1)) = \tau\sigma_A\sigma_S$ with the coefficient of correlation τ lying in $[-1, 1]$. Let $\sigma^2 \overset{\Delta}{=} \operatorname{Var} Z(1) = \sigma_S^2 - 2\tau\sigma_S\sigma_A + \sigma_A^2$. Note that m_A and σ_A^2 are the mean and variance of the Brownian motion corresponding to arrivals, while $m_S \overset{\Delta}{=} m_A + m$ and σ_S^2 are the mean and variance parameters of the service time sequence Brownian motion.

Remark 4. In order to streamline notation, we utilize the symbol $D(\cdot)$ to represent the Brownian departure process. This quantity is analyzed entirely within Section 6, and should not cause confusion with our earlier usage of the same symbol for pre-limit departure times.

We start by deriving a partial differential equation (PDE) that computes the moment generating function (mgf) of $D(t)$. For $\theta \in \mathbb{R}$, let $\varphi(t, x) = E_x \exp(\theta D(t))$ be the mgf of $D(t)$, where $E_x(\cdot)$ is the expectation operator defined by $E_x(\cdot) = E(\cdot \mid X(0) = x)$. We note that if $\mathscr{F}_t = \sigma((Z_1(s), Z_2(s))| : 0 \leq s \leq t)$, then for $s \leq t$,

$$E_x\left[\exp(\theta D(t)) \mid \mathscr{F}_s\right] = \exp(\theta D(s))\varphi(t-s, X(s)) \overset{\Delta}{=} M(s)$$

and $M(\cdot)$ should be a martingale adapted to $(\mathscr{F}_s : 0 \leq s \leq t)$. Then, if φ is smooth,

$$dM(s) \quad = \quad d(\exp(\theta D(s))\varphi(t-s, X(s))) + \exp(\theta D(s))d\varphi(t-s, X(s))$$

$$+ \, d(\exp(\theta D(s)))d\varphi(t-s,\ X(s)),$$

where

$$\begin{aligned} d(\exp(\theta D(s))) &= \theta \exp(\theta D(s))[m_A ds + dZ_2(s) + (m ds + dZ(s) + dL(s))] \\ &\quad + \frac{\theta^2}{2}\exp(\theta D(s))\sigma_S^2 ds \\ &= \exp(\theta D(s))\left[\left(\theta m_S + \frac{\theta^2}{2}\sigma_S^2\right) ds + \theta dZ_1(s) + \theta dL(s)\right], \end{aligned}$$

$$\begin{aligned} d\varphi(t-s,\ X(s)) \quad = \quad & -\frac{\partial}{\partial s}\varphi(t-s,\ X(s))ds + \frac{\partial}{\partial x}\varphi(t-s,\ X(s))[m ds + dZ(s) + dL(s)] \\ & + \frac{1}{2}\frac{\partial^2}{\partial x^2}\varphi(t-s,\ X(s))\sigma^2 ds, \end{aligned}$$

and

$$\begin{aligned} d(\exp(\theta D(s)))d\varphi(t-s,\ X(s)) &= \exp(\theta D(s))\theta\frac{\partial}{\partial x}\varphi(t-s,\ X(s))dZ_1(s)dZ(s) \\ &= \exp(\theta D(s))\theta\frac{\partial}{\partial x}\varphi(t-s,\ X(s))(\sigma_S^2 - \sigma_S\sigma_A\tau)ds. \end{aligned}$$

It follows that

$$\begin{aligned} dM(s) \quad = \quad & \exp(\theta D(s))\Bigg[\left(\theta m_S + \frac{\theta^2\sigma_S^2}{2}\right)\varphi(t-s,\ X(s)) - \frac{\partial}{\partial s}\varphi(t-s,\ X(s)) \\ & + \frac{\partial}{\partial x}\varphi(t-s,\ X(s))[m + \theta(\sigma_S^2 - \sigma_S\sigma_A\tau)] + \frac{\sigma^2}{2}\frac{\partial^2}{\partial x^2}\varphi(t-s,\ X(s))\Bigg] ds \\ & + \exp(\theta D(s))\left[\theta\varphi(t-s,\ X(s)) + \frac{\partial}{\partial x}\varphi(t-s,\ X(s))\right] dL(s) \\ & + \exp(\theta D(s))\left[\theta\varphi(t-s,\ X(s))dZ_1(s) + \frac{\partial}{\partial x}\varphi(t-s,\ X(s))dZ(s)\right]. \end{aligned}$$

Hence, if φ satisfies the PDE

$$\frac{\partial}{\partial s}\varphi(s,\ x) = [m + \theta(\sigma_S^2 - \sigma_S\sigma_A\tau)]\frac{\partial}{\partial x}\varphi(s,\ x) + \frac{\sigma^2}{2}\frac{\partial^2}{\partial x^2}\varphi(s,\ x) + \left(\theta m_S + \frac{\theta^2}{2}\sigma_S^2\right)\varphi(s,\ x) \tag{6.1}$$

for $(s,\ x) \in \mathbb{R}_+ \times \mathbb{R}_+$, subject to the boundary condition

$$\theta\varphi(s,\ 0) + \frac{\partial}{\partial x}\varphi(s,\ 0) = 0 \tag{6.2}$$

for $s \in \mathbb{R}_+$, $M(\cdot)$ will be a stochastic integral (and consequently a martingale under suitable integrability conditions). Finally, it is evident that φ must satisfy the initial condition $\varphi(0,\ x) = 1$ for $x \in \mathbb{R}_+$, in view of the fact that $\varphi(t,\ x) = E_x \exp(\theta D(t))$.

This, of course, is a special case of the constant coefficient linear PDE

$$\frac{\partial}{\partial t}w(t,\ x) = cw(t,\ x) + b\frac{\partial}{\partial x}w(t,\ x) + a\frac{\partial^2}{\partial x^2}w(t,\ x), \tag{6.3}$$

subject to

$$w(0,\ x) = 1, \tag{6.4}$$

$$\theta w(t,\, 0) + \frac{\partial}{\partial x} w(t,\, 0) = 0, \tag{6.5}$$

for $(t,\, x) \in \mathbb{R}_+ \times \mathbb{R}_+$. The PDE (6.3), with initial condition (6.4) and boundary condition (6.5), can be explicitly solved. The solution takes the form

$$\begin{aligned} w(t,\, x) \quad &= \quad \int_0^\infty \frac{1}{2\sqrt{\pi a t}} \exp\left[\frac{b(\xi - x)}{2a} + \left(c - \frac{b^2}{4a}\right) t\right] \\ &\quad \times \left\{ \exp\left[-\frac{(x-\xi)^2}{4at}\right] + \exp\left[-\frac{(x+\xi)^2}{4at}\right] \right. \\ &\quad \left. - \; 2s \int_0^\infty \exp\left[-\frac{(x+\xi+\eta)^2}{4at} - s\eta\right] d\eta \right\} d\xi, \end{aligned}$$

where $s = -\theta + \frac{b}{2a}$; see [20] Section 1.1.5. This simplifies to

$$\begin{aligned} w(t,\, x) \quad &= \quad \frac{a\theta}{b - a\theta} e^{ct - \frac{bx}{a}} \Phi\left(\frac{bt - x}{\sqrt{2at}}\right) + e^{ct} \Phi\left(\frac{bt + x}{\sqrt{2at}}\right) \\ &\quad + \frac{b - 2a\theta}{b - a\theta} e^{-x\theta + (c - b\theta + a\theta^2)t} \Phi\left(-\frac{(b - 2a\theta)t + x}{\sqrt{2at}}\right) \end{aligned} \tag{6.6}$$

whenever $b \neq a\theta$ and

$$w(t,\, x) = e^{ct - x\theta} \Phi\left(\frac{bt - x}{\sqrt{2at}}\right) + e^{ct} \Phi\left(\frac{bt + x}{\sqrt{2at}}\right)$$

whenever $b = a\theta$, where

$$\Phi(x) \overset{\Delta}{=} P(N(0,\, 1) \leq x),$$

with $N(0,\, 1)$ representing a normal random variable (rv) with mean 0 and unit variance. We are now ready to state our theorem that summarizes this discussion.

Theorem 7. *For $(t,\, x) \in \mathbb{R}_+ \times \mathbb{R}_+$, $E_x \exp(\theta D(t))$ equals $w(t,\, x)$ as given by (6.6), where $a = \sigma^2/2$, $b = m + \theta(\sigma_S^2 - \sigma_S \sigma_A \tau)$, and $c = \theta m_S + \theta^2 \sigma_S^2/2$.*

Proof. Put $\tilde{M}(s) = \exp(\theta D(s)) w(t - s,\, X(s))$. In view of the arguments above, the main remaining issue involves establishing that $(\tilde{M}(s) \,:\, 0 \leq s \leq t)$ possesses the integrability necessary to ensure that its stochastic integral representation is a martingale. But $w(s,\, \cdot)$ and $\frac{\partial}{\partial x} w(s,\, \cdot)$ have uniformly (in s) bounded exponential growth at infinity, whereas the distribution of $X(s)$ has Gaussian tails that are uniform over $s \in [0,\, t]$ (see, for example, Section 2 of Glynn and Wang [10]), so that

$$\sup_{0 \leq s,\, u \leq t} E_x \left[w(s,\, X(u))^2 + \left(\frac{\partial}{\partial x} w(s,\, X(u)) \right)^2 \right] < \infty,$$

thereby proving that $(\tilde{M}(s) : 0 \leq s \leq t)$ is a square-integrable martingale. Hence,

$$E_x \exp(\theta D(s) w(t - s,\, X(s)))$$

is independent of $s \in [0,\, t]$, so that

$$E_x \exp(\theta D(t)) = w(t,\, x). \quad \square$$

We can now (in principle) compute any moment of $D(t)$ through successive differentiation of its mgf. As a consequence, we have the ability to compute "finite time" moments of the Brownian departure time approximations for arbitrary initial conditions x and arbitrary correlation structures for Z_1 and Z_2. The existing literature provides such a first and second moment computation only in the special case in which $x = 0$, $m = 0$, and $\tau = 0$; see [12].

Given the complexity of these derivatives (in θ) for w, we now provide an alternative approach that allows us to compute the first two moments of $D(t)$. The first moment is easy to derive, since

$$E_x D(t) = m_A t + E_x X(t), \tag{6.7}$$

where $E_x X(t)$ can be computed explicitly, since $P_x(X(t) \in \cdot)$ is known in closed form (where $P_x(\cdot) = P(\cdot \mid X(0) = x)$); see, for example, Section 2 of Glynn and Wang [10], where both the spectral representation and representation in terms of $\Phi(\cdot)$ can be found.

For the variance, we note that

$$\mathrm{Var}_x\, D(t) = \mathrm{Var}\, Z_2(t) + \mathrm{Var}_x\, X(t) + 2E_x Z_2(t) X(t), \tag{6.8}$$

where $\mathrm{Var}_x\, W \overset{\Delta}{=} E_x W^2 - (E_x W)^2$ for any square-integrable rv W. Again, $\mathrm{Var}_x\, X(t)$ can be obtained from the known closed form for $P_x(X(t) \in \cdot)$. Since $\mathrm{Var}\, Z_2(t) = \sigma_A^2 t$, this leaves the computation of the covariance term $E_x Z_2(t) X(t)$.

We start by recalling that $u(t, x) = E_x X(t)$ satisfies the Kolmogorov backwards partial differential equation

$$\frac{\partial}{\partial t} u(t, x) = m \frac{\partial}{\partial x} u(t, x) + \frac{\sigma^2}{2} \frac{\partial^2}{\partial x^2} u(t, x),$$

subject to $u(0, x) = x$ and $\frac{\partial}{\partial x} u(t, 0) = 0$ for $t > 0$ and $x \in \mathbb{R}_+$. It follows that

$$\begin{aligned} du(t-s, X(s)) \quad &= \quad \Big[-\frac{\partial}{\partial s} u(t-s, X(s)) + m \frac{\partial}{\partial x} u(t-s, X(s)) \\ &\quad + \frac{\sigma^2}{2} \frac{\partial^2}{\partial x^2} u(t-s, X(s)) \Big] ds \\ &\quad + \frac{\partial}{\partial x} u(t-s, X(s)) dL(s) \\ &\quad + \frac{\partial}{\partial x} u(t-s, X(s)) dZ(s) \\ &= \quad \frac{\partial}{\partial x} u(t-s, X(s)) dZ(s). \end{aligned}$$

Consequently,

$$u(0, X(t)) - u(t, X(0)) = \int_0^t \frac{\partial}{\partial x} u(t-s, X(s)) dZ(s),$$

so that conditional on $X(0) = x$,

$$X(t) = u(t, x) + \int_0^t \frac{\partial}{\partial x} u(t-s, X(s)) dZ(s).$$

Then,

$$E_x Z_2(t) X(t) = E_x Z_2(t) \int_0^t \frac{\partial}{\partial x} u(t-s, X(s)) dZ(s).$$

The integration by parts formula for semimartingales (see, for example, p.60 of Protter [21]) then implies that

$$\begin{aligned} Z_2(t)\int_0^t \frac{\partial}{\partial x}u(t-s,\ X(s))dZ(s) \quad &= \quad \int_0^t \int_0^s \frac{\partial}{\partial x}u(t-r,\ X(r))dZ_2(r)dZ(s) \\ &\quad + \int_0^t Z_2(s)\frac{\partial}{\partial x}u(t-s,\ X(s))dZ(s) \\ &\quad + \int_0^t \frac{\partial}{\partial x}u(t-s,\ X(s))d[Z_2,\ Z](s), \end{aligned} \tag{6.9}$$

where $[Z_2,\ Z](\cdot)$ is the quadratic covariation of Z_2 and Z. In particular,

$$[Z_2,\ Z](s) = (\sigma_A\sigma_S\tau - \sigma_A^2)s$$

for $s \geq 0$. The integrability arguments used in the proof of Theorem 7 again apply here, so that the local martingale stochastic integrals in (6.9) are true martingales, yielding the identity

$$E_x Z_2(t)X(t) = (\sigma_A\sigma_S\tau - \sigma_A^2)\int_0^t E_x\frac{\partial}{\partial x}u(t-s,\ X(s))ds. \tag{6.10}$$

Since the right-hand side of (6.10) can be evaluated in terms of the known closed form for $P_x(X(t) \in \cdot)$, we have arrived at an alternative formula for $\mathrm{Var}_x\ D(t)$ (in view of (6.8)).

We now show that (6.10) can be easily computed when $x = 0$ (i.e. the queue starts empty) and $m = 0$ (i.e. the queue is perfectly balanced, corresponding to $\rho = 1$). It is known that when $m = 0$, $X(\cdot) \overset{\mathscr{D}}{=} |\sigma B(\cdot)|$, where $B(\cdot)$ is a standard Brownian motion; see p.27 of Rogers and Williams [23]. As a consequence,

$$\begin{aligned} u(t,\ x) &= E|x + \sigma B(t)| \\ &= \int_{-x}^{\infty}(x+y)\phi\left(\frac{y}{\sigma\sqrt{t}}\right)\frac{dy}{\sigma\sqrt{t}} - \int_{-\infty}^{-x}(x+y)\phi\left(\frac{y}{\sigma\sqrt{t}}\right)\frac{dy}{\sigma\sqrt{t}}, \end{aligned}$$

where $\phi(\cdot)$ is the density of a $N(0,\ 1)$ rv. It follows that

$$\begin{aligned} \frac{\partial}{\partial x}u(t,\ x) &= 1 - 2P(\sigma B(t) \leq -x) \\ &= P(|\sigma B(t)| \leq x) \end{aligned}$$

for $t > 0$ and $x \geq 0$. Then,

$$E_0\frac{\partial}{\partial x}u(t-s,\ X(s)) = P(|\sigma B(t-s)| \leq |\sigma B'(s)|),$$

where $(B'(t) : t \geq 0)$ is a standard Brownian motion independent of $(B(t) : t \geq 0)$ (since $X(s) \overset{\mathscr{D}}{=} |\sigma B'(s)|$). Note that

$$\begin{aligned} \int_0^t P(|B(t-s)| \leq |B'(s)|)ds &= \int_0^{\frac{t}{2}} P(|B(t-s)| \leq |B'(s)|)ds \\ &\quad + \int_{\frac{t}{2}}^{t} P(|B(t-s) \leq |B'(s)|)ds \\ &= \int_0^{\frac{t}{2}} P(|B(t-s)| \leq |B'(s)|)ds \\ &\quad + \int_0^{\frac{t}{2}} P(|B(r) \leq |B'(t-r)|)dr \end{aligned}$$

$$= \int_0^{\frac{t}{2}} [P(|B(t-s)| \leq |B'(s)|)$$
$$+ P(|B'(s) \leq |B(t-s)|)]ds$$
$$= \frac{t}{2}.$$

It follows from (6.10) that

$$E_0 Z_2(t) X(t) = (\sigma_A \sigma_S \tau - \sigma_A^2)\frac{t}{2}.$$

Also, in this setting in which $m = 0$,

$$E_0 X^2(t) = E_0 \sigma^2 B(t)^2 = \sigma^2 t,$$

and

$$E_0 X(t) = E|\sigma B(t)| = \sqrt{\frac{2}{\pi}} \sigma t^{\frac{1}{2}}.$$

Consequently, (6.8) yields the formula

$$\mathrm{Var}_0\, D(t) = \left[(\sigma_A^2 - 2\tau\sigma_A\sigma_S + \sigma_S^2) \left(1 - \frac{2}{\pi}\right) + \sigma_A\sigma_S\tau \right] t. \tag{6.11}$$

We conclude our above discussion with the following theorem.

Theorem 8. *When $m = 0$, $\mathrm{Var}_0\, D(t)$ is given by (6.11).*

This result generalizes the Brownian calculation established by Hanbali et al. [12] for the special case $\tau = 0$ to a general dependence structure. We further note that our argument relies on very different ideas than the Brownian bridge derivation developed there. Unlike [12], we do not show that the Brownian variance is the heavy-traffic limit obtained from the pre-limit variances. However, our argument does compute the distributional variability of $D(t)$ which is equally important.

We note that the exact finite-time variance of the departure process for the Brownian limit in stationarity has been computed by Whitt and You [27] when $\tau = 0$ and the RBM drift is negative (so that a stationary version exists). They also show that the pre-limit $G/G/1$ departure variance converges to the Brownian limit.

7. Implications for downstream stations

We now study the implication of the limit theory developed in Sections 3 through 5 on queueing at downstream stations that are fed by the departure process thus far analyzed in this paper. In particular, we consider a tandem network of two queues, in which customers from station 1 are immediately fed into station 2 upon their departure from station 1. We assume that both stations are single server systems with infinite capacity waiting rooms that process their respective customers according to a FIFO queue discipline.

In this setting, we will need to add in a service time sequence at station 2 to complement the inter-arrival and service time sequences at station 1. In particular, we will assume throughout this section that we have three independent iid sequences $(\chi_k : k \geq 1)$, $(V_k : k \geq 0)$, and $(\tilde{V}_k : k \geq 0)$, where the $\tilde{V}_k$'s correspond to the service times at station 2. Since we will be considering these stations in heavy traffic, we require that these sequences are all "in balance" with one another, so that

$$E\chi_1 = EV_0 = E\tilde{V}_0 = \alpha.$$

We now follow the approach of Section 2 in scaling the service times at stations 1 and 2 by ρ_1 and ρ_2, respectively. At station 1, we can again define the random variables $W_{\rho_1}(n, x)$, $D_{\rho_1}(n, x)$, etc. (in which the V_i's are scaled by ρ_1). Our main interest here will be in the queueing effects at the downstream station 2, as measured by the waiting time sequence $(\tilde{W}_{\rho_2}(n, x, y) : n \geq 0)$, where our notation emphasizes the fact that the nth waiting time at station 2 depends on both the initial work $x\, (= W_{\rho_1}(0, x))$ at station 1 and also the initial work $y = (\tilde{W}_{\rho_2}(0, x, y))$ at station 2. We can now study the network in heavy traffic, so that ρ_1 and ρ_2 both converge to 1, with limits taken so that

$$\frac{1-\rho_1}{1-\rho_2} \to c \tag{7.1}$$

as $\rho_1 \to 1$. We allow c to take on the value 0 (so that station 1 is more congested than station 2) or the value ∞ (so that station 2 is more congested than station 1).

We will argue in this section that the behavior of the second station depends on the first station's departure process over time scales that are of the same order as the second station's intrinsic "diffusion time scale" (that is of order $(1-\rho_2)^{-2}$). For example, this implies that when $c = 0$, the first station is likely to be consistently busy over the time scale $(1-\rho_2)^{-2}$, so that the queueing effects at the second station depends only on the variability characteristics of the first station's service times (and not on the variability characteristics of the arrival process to station (1). This occurs even though part (a) of Theorem 6 asserts that the long time scale behavior of station 1's departure process is determined by the arrival sequence's variability. (Recall that in equilibrium, the amount of work held at station 1 is of order $(1-\rho_1)^{-1}$.)

We start by noting that because the departures from station 1 are the arrivals to station 2,

$$\begin{aligned} \tilde{W}_{\rho_2}(n+1, x, y) \quad &= \quad \left[\tilde{W}_{\rho_2}(n, x, y) + \rho_2 \tilde{V}_n - (D_{\rho_1}(n+1, x) - D_{\rho_1}(n, x))\right]^+ \\ &= \quad \Big[\tilde{W}_{\rho_2}(n, x, y) + \rho_2 \tilde{V}_n - \chi_{n+1} - (W_{\rho_1}(n+1, x) - W_{\rho_1}(n, x)) \\ &\quad - \rho(V_{n+1} - V_n)\Big]^+ . \end{aligned} \tag{7.2}$$

Consequently, the triplet $(W_{\rho_1}(n, x), \tilde{W}_{\rho_2}(n, x, y), V_n)$ satisfies a stochastic recursion in n that ensures that the triplet forms an $\mathbb{R}_+^3$-valued Markov chain. (We have added V_n to the state in order to ensure that the "noise" sequence defining the recursion is independent across n.) According to [19], this Markov chain is aperiodic and positive Harris recurrent when $\rho_1 < 1$ and $\rho_2 < 1$. As such, it has a unique equilibrium distribution described by the triplet $(W_{\rho_1}(\infty), \tilde{W}_{\rho_2}(\infty), V_\infty)$, where $V_\infty \overset{\mathcal{D}}{=} V_0$. Furthermore, it is known that when $E\chi_1^{1/r} + EV_0^{1/r} < \infty$ for $r < 1/2$,

$$(1-\rho_1) W_{\rho_1}(\infty) \Rightarrow \mathscr{E} \tag{7.3}$$

as $\rho_1 \nearrow 1$, where $\mathscr{E}$ is exponentially distributed; see [1] p.287.

Our focus, in this section, is on the behavior of the second station when it is stable (in isolation), so that $\rho_2 < 1$. We start with the setting in which the entire network is stable, so that $\rho_1 < 1$ as well as $\rho_2 < 1$. Suppose then that $((W(n), \tilde{W}(n), V_n) : n \geq 0)$ is a stationary version of our Markov chain, so that $(W(n),\ \tilde{W}(n),\ V_n) \overset{\mathscr{D}}{=} (W_{\rho_1}(\infty),\ \tilde{W}_{\rho_2}(\infty),\ V_\infty)$. Also,

$$\begin{aligned} W(n) - W(n+1) &= W(n) - [W(n) + \rho_1 V_n - \chi_{n+1}]^+ \\ &\le W(n) - (W(n) + \rho_1 V_n - \chi_{n+1}) \\ &= \chi_{n+1} - \rho_1 V_n. \end{aligned} \tag{7.4}$$

Set $g(w,\ \tilde{w},\ v) = \tilde{w}^2$. Then (7.2) and (7.4) imply that

$$\begin{aligned} g(W(1),\ \tilde{W}(1),\ V_1) &= \left(\left[\tilde{W}(0) + \rho_2 \tilde{V}_0 - \chi_1 + W_0 - W_1 + \rho_1 V_0 - \rho_1 V_1\right]^+\right)^2 \\ &\le \left(\left[\tilde{W}(0) + \rho_2 \tilde{V}_0 - \chi_1 + \chi_1 - \rho_1 V_0 + \rho_1 V_0 - \rho_1 V_1\right]^+\right)^2 \\ &= \left(\left[\tilde{W}(0) + \rho_2 \tilde{V}_2 - \rho_1 V_1\right]^+\right)^2 \\ &\le \left(\tilde{W}(0) + \rho_2 \tilde{V}_0 - \rho_1 V_1\right)^2, \end{aligned}$$

so that if $f(w,\ \tilde{w},\ v) \overset{\Delta}{=} \tilde{w}$, then

$$\begin{aligned} E\left[g(W(1),\ \tilde{W}(1),\ V_1) \,\middle|\, W(0),\ \tilde{W}(0),\ V_0\right] \quad \le &\quad \tilde{W}(0)^2 + 2(\rho_2 - \rho_1)\alpha \tilde{W}(0) \\ &+ E(\rho_2 \tilde{V}_0 - \rho_1 V_1)^2 \\ = &\quad g(W(0),\ \tilde{W}(0),\ V_0) \\ &- 2|\rho_1 - \rho_2| \alpha f(W(0),\ \tilde{W}(0),\ V_0) \\ &+ E(\rho_2 \tilde{V}_0 - \rho_1 V_1)^2, \end{aligned} \tag{7.5}$$

provided that $\rho_1 > \rho_2$. According to [11], (7.5) implies that

$$E\tilde{W}_{\rho_2}(\infty) \le \frac{E(\rho_2 \tilde{V}_0 - \rho_1 V_1)^2}{2|\rho_1 - \rho_2|\alpha}, \tag{7.6}$$

when $\rho_1 > \rho_2$.

Suppose that $c = 0$ in (7.1). Then, (7.6) implies that $\tilde{W}_{\rho_2}(\infty) = O((1-\rho_2)^{-1})$ as $\rho_1 \nearrow 1$. If N_0 is the index of the first customer to experience no waiting at station 1, then (7.3) implies that $(1-\rho_1)^2 N_0$ converges in distribution to a positive rv, from which it is evident that $(1-\rho_2)^2 N_0 \to \infty$ in probability as $\rho_1 \nearrow 1$. Hence, for any $t > 0$ and (measurable) $C \in \mathbb{R}_+$,

$$\begin{aligned} P(\tilde{W}_{\rho_2}(\infty) \in C) &= P(\tilde{W}_{\rho_2}(n) \in C) + o(1) \\ &= P(\tilde{W}_{\rho_2}(n) \in C,\ N_0 > n) + o(1), \end{aligned}$$

provided that $n = \lfloor t/(1-\rho_2)^2 \rfloor$ for $t > 0$. If $\tilde{N}_0$ is the index of the first customer to experience no waiting time at station 2, then

$$\begin{aligned} P(\tilde{W}_{\rho_2}(\infty) \in C) &= P(\tilde{W}_{\rho_2}(n) \in C,\ \tilde{N}_0 \le n < N_0) + O(P(\tilde{N}_0 > n)) + o(1) \\ &= E(P(\tilde{W}_{\rho_2}(n) \in C,\ N_0 > n|\tilde{N}_0)\mathbb{I}(\tilde{N}_0 \le n)) + O(P(\tilde{N}_0 > n)) + o(1). \end{aligned} \tag{7.7}$$

On $\{\tilde{N}_0 = k\}$ with $k \le n$,

$$\begin{aligned} P(\tilde{W}_{\rho_2}(n) \in C,\ N_0 > n \mid \tilde{N}_0 = k) &= P(\tilde{W}_{\rho_2}(n-k) \in C \mid \tilde{W}_{\rho_2}(0) = 0) + o(1) \\ &= P\left(\max_{0 \le j \le n-k} \tilde{S}_j \in C\right) + o(1), \end{aligned}$$

where

$$\tilde{S}_j = \sum_{\ell=0}^{j-1}(\rho_2 \tilde{V}_\ell - \rho_1 V_j).$$

Since it is well known (see, for example, [1] p.287) that

$$(1-\rho_2)\max_{j\geq 0} \tilde{S}_j \Rightarrow \frac{\mathrm{Var}(\tilde{V}_0 + V_0)}{2\alpha}\mathscr{E}(1),$$

where $\mathscr{E}(1)$ is an exponential rv with mean 1, our next theorem follows from (7.6) and (7.7) (upon choosing t arbitrarily large).

Theorem 9. *If $c = 0$ with $\rho_1 \nearrow 1$, then*

$$(1-\rho_2)\tilde{W}_{\rho_2}(\infty) \Rightarrow \frac{\mathrm{Var}(\tilde{V}_0 + V_0)}{2\alpha}\mathscr{E}(1)$$

as $\rho_1 \nearrow 1$.

The key point in Theorem 9 is that the equilibrium distribution at the second station has no dependence on the inter-arrival distribution in this setting, as predicted by the analysis earlier in this section.

We turn next to the case where $c \in (0, \infty]$, so that the second station holds a significant fraction of the network's total work in equilibrium. We recall that if the network starts empty (so that $W_{\rho_1}(0) = \tilde{W}_{\rho_2}(0) = 0$), then the nth departure time $D_{\rho_1}(n)$ at station 1 is given by

$$\begin{aligned} D_{\rho_1}(n) &= A_n + W_{\rho_1}(n) + \rho_1 V_n \\ &= \rho_1 \sum_{j=0}^{n} V_j + \max_{1\leq j\leq n}\left[A_j - \rho_1\sum_{i=0}^{j-1} V_i\right]^+ \\ &= \max_{0\leq j\leq n}\left[\rho_1 \sum_{i=j}^{n} V_i + A_j\right]^+ . \end{aligned}$$

Similarly, because the jth arrival time to station 2 is $D_{\rho_1}(j)$, it follows that the nth departure time $\tilde{D}_{\rho_2}(n)$ from station 2 is given by

$$\begin{aligned} \tilde{D}_{\rho_2}(n) &= \max_{0\leq k\leq n}\left[D_{\rho_1}(k) + \rho_2\sum_{i=k}^{n}\tilde{V}_i\right]^+ \\ &= \max_{0\leq k\leq n}\left[\max_{0\leq j\leq k}\left[\rho_1\sum_{i=j}^{k} V_i + A_j\right]^+ + \rho_2\sum_{\ell=k}^{n}\tilde{V}_\ell\right]^+ \\ &= \max_{0\leq j\leq k\leq n}\left[A_j + \rho_1\sum_{i=j}^{k} V_i + \rho_2\sum_{\ell=k}^{n}\tilde{V}_\ell\right]^+ . \end{aligned} \tag{7.8}$$

The departure time representation (7.8) for tandem networks is well known; see [13] and [24]. On the other hand,

$$\begin{aligned} \tilde{D}_{\rho_2}(n) &= D_{\rho_1}(n) + \tilde{W}_{\rho_2}(n) + \rho_2 \tilde{V}_n \\ &= A_n + W_{\rho_1}(n) + \rho_1 V_n + \tilde{W}_{\rho_2}(n) + \rho_2\tilde{V}_n, \end{aligned} \tag{7.9}$$

so the total network time-in-system for customer n is given by

$$\begin{aligned}\tilde{T}(n) &= \tilde{D}_{\rho_2}(n) - A_n \\ &= \max_{0\le j\le k\le n}\left[A_j - A_n + \rho_1\sum_{i=j}^{k} V_i + \rho_2\sum_{\ell=k}^{n}\tilde{V}_\ell\right]^+ .\end{aligned}$$

Since $(\chi_1, \ldots, \chi_n) \overset{\mathscr{D}}{=} (\chi_n, \ldots, \chi_1)$, $(V_0, \ldots, V_n) \overset{\mathscr{D}}{=} (V_n, \ldots, V_0)$, and $(\tilde{V}_0, \ldots, \tilde{V}_n) \overset{\mathscr{D}}{=} (\tilde{V}_n, \ldots, \tilde{V}_0)$, it follows that

$$\begin{aligned}(W_{\rho_1}(n),\ \tilde{T}(n)) &\overset{\mathscr{D}}{=} \Bigg(\max_{0\le k\le n}\sum_{i=k}^{n-1}[\rho_1 V_{n-i-1} - \chi_{n-i}], \\ &\qquad \max_{0\le j\le k\le n}\left[-\sum_{\ell=j+1}^{n}\chi_{n+1-\ell} + \rho_1\sum_{i=j}^{k} V_{n-i} + \rho_2\sum_{\ell=k}^{n}\tilde{V}_{n-\ell}\right]\Bigg) \\ &= \Bigg(\max_{0\le k\le n}\sum_{\ell=1}^{n-k}[\rho_1 V_{\ell-1} - \chi_\ell], \\ &\qquad \max_{0\le j\le k\le n}\left[-\sum_{\ell=1}^{n-j}\chi_{n+1-\ell} + \rho_1\sum_{s=n-k}^{n-j} V_s + \rho_2\sum_{r=0}^{n-k}\tilde{V}_r\right]\Bigg) \\ &= \Bigg(\max_{0\le k\le n}\sum_{\ell=1}^{k}[\rho_1 V_{\ell-1} - \chi_\ell], \\ &\qquad\quad \max_{0\le j\le k\le n}\left[\rho_2\sum_{r=0}^{j}\tilde{V}_r + \rho_1\sum_{s=j}^{k} V_s - \sum_{\ell=1}^{k}\chi_\ell\right]\Bigg) \\ &\overset{\Delta}{=} (M_1(n),\ M_2(n)).\end{aligned}$$

Clearly,

$$(M_1(n),\ M_2(n)) \Rightarrow (M_1(\infty),\ M_2(\infty)) \ \text{ a.s.} \tag{7.10}$$

as $n \to \infty$. Furthermore, because

$$\frac{1}{n}\sum_{\ell=1}^{k}\chi_\ell \to \alpha \ \text{ a.s.},\ \frac{1}{n}\sum_{\ell=1}^{n} V_\ell \to \alpha \ \text{ a.s.},\ \frac{1}{n}\sum_{\ell=1}^{n}\tilde{V}_\ell \to \alpha \ \text{ a.s.},$$

and ρ_1, $\rho_2 < 1$, $M_1(\infty)$ and $M_2(\infty)$ are finite-valued.

If $E\chi_1^{1/r} + EV_0^{1/r} + E\tilde{V}_0^{1/r} < \infty$ for $r < 1/2$, then [16] show that we may assume the existence of independent zero mean Brownian motions for which

$$\sum_{j=0}^{n} V_j = n\alpha + Z_1(n) + o(n^r) \ \text{ a.s.},$$

$$A_n = n\alpha + Z_2(n) + o(n^r) \ \text{ a.s.},$$

and

$$\sum_{j=0}^{n}\tilde{V}_j = n\alpha + Z_3(n) + o(n^r) \ \text{ a.s.}$$

as $n \to \infty$.

As a result,

$$(1-\rho_2)(M_1(\lfloor t/(1-\rho_2)^2\rfloor),\ M_2(\lfloor t/(1-\rho_2)^2\rfloor)) = (M_1'(t),\ M_2'(t)) + o(t^r(1-\rho_2)^{1-2r}) \text{ a.s.,} \tag{7.11}$$

where

$$M_1'(t) = \max_{0\le s\le t}\left[-\left(\frac{1-\rho_1}{1-\rho_2}\right)\alpha s + (1-\rho_2)\rho_1 Z_1\left(\frac{s}{(1-\rho_2)^2}\right) - (1-\rho_2)Z_2\left(\frac{s}{(1-\rho_2)^2}\right)\right]$$

and

$$\begin{aligned} M_2'(t) \quad = \quad & \max_{0\le s\le u\le t}\left[-\alpha s + (1-\rho_2)\rho_2 Z_3\left(\frac{s}{(1-\rho_2)^2}\right) - \left(\frac{1-\rho_1}{1-\rho_2}\right)\alpha(u-s)\right. \\ & \left. + \ (1-\rho_2)\rho_1\left(Z_1\left(\frac{u}{(1-\rho_2)^2}\right) - Z_1\left(\frac{s}{(1-\rho_2)^2}\right)\right) - (1-\rho_2)Z_2\left(\frac{u}{(1-\rho_2)^2}\right)\right]. \end{aligned}$$

Furthermore, by virtue of the law of the iterated logarithm for Brownian motion and the negative drift, it is evident that when $c \in (0,\ \infty]$,

$$(1-\rho_2)(M_1(\infty) - M_1(t/(1-\rho_2)^\kappa),\ M_2(\infty) - M_2(t/(1-\rho_2)^\kappa)) \to 0 \text{ a.s.} \tag{7.12}$$

as $\rho_2 \nearrow 1$ whenever $\kappa > 2$.

Finally, note that the scaling properties of Brownian motion imply that

$$(M_1'(t),\ M_2'(t)) \overset{\mathscr{D}}{=} (M_1''(t),\ M_2''(t)),$$

where

$$M_1''(t) = \max_{0\le s\le t}\left[-\left(\frac{1-\rho_1}{1-\rho_2}\right)\alpha s + \rho_1 Z_1(s) - Z_2(s)\right] \tag{7.13}$$

and

$$M_2''(t) = \max_{0\le s\le u\le t}\left[-\alpha s + \rho_2 Z_3(s) - \left(\frac{1-\rho_1}{1-\rho_2}\right)\alpha(u-s) + \rho_1(Z_1(u) - Z_1(s)) - Z_2(u)\right].$$

As a consequence of (7.10) through (7.12), we find that if we choose $\kappa \in (2,\ 1/r)$, then

$$\begin{aligned} (1-\rho_2)(M_1(\infty) - M_2(\infty)) &= (1-\rho_2)(M_1(\lfloor t/(1-\rho_2)^\kappa\rfloor),\ M_2(\lfloor t/(1-\rho_2)^\kappa\rfloor)) + o(1) \text{ a.s.} \\ &= (1-\rho_2)(M_1'(\lfloor t/(1-\rho_2)^\kappa\rfloor),\ M_2'(\lfloor t/(1-\rho_2)^\kappa\rfloor)) + o(t^r(1-\rho_2)^{1-\kappa r}) \\ &\overset{\mathscr{D}}{=} (M_1''(t/(1-\rho_2)^{\kappa-2}),\ M_2''(t/(1-\rho_2)^{\kappa-2})) + o_p(1) \\ &\Rightarrow \Big(\max_{s\ge 0}[-c\alpha s + Z_1(s) - Z_2(s)], \\ &\qquad \max_{0\le s\le u}[-\alpha s + Z_3(s) - c\alpha(u-s) + Z_1(u) - Z_1(s) - Z_2(u)]\Big) \qquad (7.14) \\ &\overset{\Delta}{=} (\Lambda_1,\ \Lambda_2) \end{aligned}$$

as $\rho_2 \nearrow 1$ when $c \in (0,\ \infty)$. On the other hand, when $c = \infty$, the drift terms converge to $-\infty$, so that

$$(1-\rho_2)(M_1(\infty),\ M_2(\infty)) \Rightarrow \left(0,\ \max_{s\ge 0}[-\alpha s + Z_3(s) - Z_2(s)]\right). \tag{7.15}$$

The maximum rv appearing on the right hand side is the equilibrium rv associated with an RBM having drift $-\alpha$ and variance parameter $\operatorname{Var}\chi_1 + \operatorname{Var}\tilde{V}_0$, and hence is exponentially distributed with mean $(\operatorname{Var}\chi_1 + \operatorname{Var}\tilde{V}_0)/(2\alpha)$.

In view of (7.8) and (7.9), we have therefore established the following result.

Theorem 10. *Under the conditions of this section,*

$$(1-\rho_2)(W_{\rho_1}(\infty),\ \tilde{W}_{\rho_2}(\infty)) \Rightarrow (\Lambda_1,\ \Lambda_2-\Lambda_1)$$

as $\rho_2 \nearrow 1$ *when* $c \in (0,\ \infty)$*, and*

$$(1-\rho_2)(W_{\rho_1}(\infty),\ \tilde{W}_{\rho_2}(\infty)) \Rightarrow \left(0,\ \frac{\mathrm{Var}\,\chi_1+\mathrm{Var}\,\tilde{V}_0}{2\alpha}\mathscr{E}(1)\right)$$

as $\rho_2 \nearrow 1$ *with* $c=\infty$.

Remark 5. This result is essentially that derived by Harrison [13] under different assumptions. His result does not discuss the case $c=\infty$ (nor does it deal with $c=0$).

Our last result describes the behavior of the second queue when the system as a whole is unstable, but the second station is stable. In other words, we consider here the case where $\rho_1 \geq 1$, but $\rho_2 < 1$. In this context, there is no stationary distribution for $(W_{\rho_1}(n,\ x),\ \tilde{W}_{\rho_2}(n,\ x,\ y),\ V_n)$. Nevertheless, there is a limiting distribution for $\tilde{W}_{\rho_2}(n,\ x,\ y)$, in the sense that

$$\tilde{W}_{\rho_2}(n,\ x,\ y) \Rightarrow \tilde{W}_{\rho_2}(\infty)$$

as $n \to \infty$, because when n is large relative to $(1-\rho_2)^{-2}$, the second queue will empty with high probability prior to n, as in the proof of Theorem 9. Furthermore, over the time span that elapses between the emptying time and n, the first queue will never empty and its inter-departure times will match the successive service times at the first queue. Hence,

$$\tilde{W}_{\rho_2}(n,\ x,\ y) \Rightarrow \max_{j\geq 0}\left[\sum_{i=1}^{j}(\rho_2\tilde{V}_{i-1}-\rho_1 V_i)\right] \tag{7.16}$$

as $n \to \infty$. The limiting rv appearing in (7.16) is, of course, the equilibrium waiting time of a FIFO single-server queue with inter-arrival times given by the $\rho_1 V_i$'s and service times given by the $\rho_2\tilde{V}_j$'s.

By applying the heavy-traffic limit theorem (e.g. Asmussen (2003) p.287) to the right-hand side of (7.16), we obtain our final result.

Theorem 11. *Under the conditions of this section,*

$$(\rho_1-\rho_2)\tilde{W}_{\rho_2}(\infty) \Rightarrow \frac{\mathrm{Var}\,V_0+\mathrm{Var}\,\tilde{V}_0}{2\alpha}\mathscr{E}(1)$$

when $\rho_1 = 1$ *or* $\rho_1 \searrow 1$ *and* $\rho_2 \nearrow 1$.

The theorems of this section make clear that none of the equilibrium or limiting distributions identified when the downstream station is stable depend upon the BRAVO variance parameter. In other words, the BRAVO variance reduction that holds for the $M/M/1$ queue departure process apparently does not manifest itself in the behavior of the downstream queue, no matter how close are ρ_1 and ρ_2 to 1.

These theorems implicitly take into account the departure process variability of the first station in equilibrium in the two station tandem setting. Whitt and You [28] compute the stationary departure variability in a network with feedback, in both the Brownian and pre-limit settings. In contrast, our theory focuses on the implications of departure variability on downstream stations.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

It is an honor to contribute this paper to the special issue in the memory of Larry Shepp. Professor Shepp was an inspirational figure who made major contributions to many areas of probability and applied probability. In particular, his ability to generate novel, creative, and deep contributions across so many different problem areas was a singular and impressive accomplishment. One of the many areas to which he contributed is the subject of this paper, namely diffusion limits for queueing networks.

References

[1] S. Asmussen, Applied Probability and Queues, second ed., Springer-Verlag, New York, 2003.
[2] P.J. Burke, The output of a queueing system, Oper. Res. 4 (1956) 699–704.
[3] W. Chang, Output distribution of a single channel queue, Oper. Res. 11 (1963) 620–623.
[4] H. Chen, D.D. Yao, Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization, Springer-Verlag, New York, 2001.
[5] E. Csáki, M. Csörgő, On additive functionals of Markov chains, J. Theor. Probab. 8 (1995) 905–919.
[6] D.J. Daley, The correlation structure of the output process of some single server queueing systems, Ann. Math. Stat. 39 (1968) 1007–1019.
[7] D.J. Daley, Queueing output processes, Adv. Appl. Probab. 8 (1976) 395–415.
[8] U. Einmahl, Extensions of results of Komlós, Major, and Tusnády to the multivariate case, J. Multivariate Anal. 28 (1) (1989) 20–68.
[9] P. Finch, The output process of the queueing system M/G/1, J. R. Stat. Soc. Ser. B Stat. Methodol. 21 (1959) 375–380.
[10] P.W. Glynn, R.J. Wang, On the rate of convergence to equilibrium for reflected Brownian motion, Queueing Syst. 89 (2018) 165–197.
[11] P.W. Glynn, A. Zeevi, Bounding stationary expectations of Markov processes, in: Institute of Mathematical Statistics Collections: Markov Processes and Related Topics: A Festschrift for Thomas G. Kurtz. Vol. 4, 2008, pp. 195–214.
[12] A.A. Hanbali, M. Mandjes, Y. Nazarathy, W. Whitt, The asymptotic variance of departures in critically loaded queues, Adv. Appl. Probab. 43 (2011) 243–263.
[13] J.M. Harrison, The heavy traffic approximation for single server queues in series, J. Appl. Probab. 10 (1973) 613–629.
[14] J.M. Harrison, Brownian Models of Performance and Control, Cambridge University Press, Cambridge, 2013.
[15] J.M. Harrison, L.A. Shepp, A tandem storage system and its diffusion limit, Stoch. Process. their Appl. 16 (3) (1984) 257–274.
[16] J. Komlós, P. Major, G. Tusnády, An approximation of partial sums of independent rv's and the sample df. II, Z. Wahrscheinlichkeitstheor. Verwandte Geb. 34 (1976) 33–58.
[17] F. Merlevède, E. Rio, Strong approximation for additive functionals of geometrically ergodic Markov chains, Electron. J. Probab. 20 (2015) 1–27.
[18] Y. Nazarathy, G. Weiss, The asymptotic variance rate of the output process of finite capacity birth-death queues, Queueing Syst. 59 (2008) 135–156.
[19] E. Nummelin, A conservation property for general GI/G/1 queues with an application to tandem queues, Adv. Appl. Probab. 11 (3) (1979) 660–672.
[20] A.D. Polyanin, Handbook of Linear Partial Differential Equations for Engineers and Scientists, Chapman and Hall/CRC, 2002.
[21] P. Protter, Stochastic Integration and Differential Equations, Springer-Verlag, Berlin Heidelberg, 1990.
[22] E. Reich, Departure processes (with discussion), in: W.L. Smith, W.E. Wilkinson (Eds.), Congestion Theory, 1965, pp. 439–459.

[23] L.C.G. Rogers, D. Williams, Diffusions, Markov Processes, and Martingales (Volume 1: Foundations), second ed., Cambridge University Press, Cambridge, 2000.

[24] S.V. Tembe, R.W. Wolff, The optimal order of service in tandem queues, Oper. Res. 22 (4) (1974) 824–832.

[25] T.L. Vlach, R.L. Disney, The departure process from the GI/G/1 queue, J. Appl. Probab. 6 (1969) 704–707.

[26] W. Whitt, The queueing network analyzer, Bell Syst. Tech. J. 62 (9) (1983) 2779–2815.

[27] W. Whitt, W. You, Heavy-traffic limit of the GI/GI/1 stationary departure process and its variance function, Stoch. Syst. 8 (2) (2018) 143–165.

[28] W. Whitt, W. You, Heavy-traffic limits for stationary network flows, Queueing Syst. 95 (2020) 53–68.