Queueing Syst (2015) 80:71–103
DOI 10.1007/s11134-014-9428-4

A queueing model with independent arrivals, and its fluid and diffusion limits

Harsha Honnappa · Rahul Jain · Amy R. Ward

Received: 29 December 2013 / Revised: 29 October 2014 / Published online: 25 November 2014

Abstract We study a queueing model with ordered arrivals, which can be called the $\Delta_{(i)}/GI/1$ queue. Here, customers from a fixed, finite, population independently sample a time to arrive from some given distribution F, and enter the queue in order of the sampled arrival times. Thus, the arrival times are order statistics, and the inter-arrival times are differences of consecutive order statistics. They are served by a single server with independent and identically distributed service times, with general service distribution G. The discrete event model is analytically intractable. Thus, we develop fluid and diffusion limits for the performance metrics of the queue. The fluid limit of the queue length is observed to be a reflection of a 'fluid netput' process, while the diffusion limit is observed to be a function of a Brownian motion and a Brownian bridge process or 'diffusion netput' process. The diffusion limit can be seen as being reflected through the directional derivative of the Skorokhod regulator of the fluid netput process in the direction of the diffusion netput process. We also observe what may be interpreted as a sample path Little's Law. Sample path analysis reveals various operating regimes where the diffusion limit switches between a free diffusion,

H. Honnappa
Department of Electrical Engineering, University of Southern California, 3740 McClintock Ave.,
Los Angeles, CA 90089, USA
e-mail: honnappa@usc.edu

R. Jain (✉)
EE, CS and ISE Departments, University of Southern California, 3740 McClintock Ave., Los Angeles,
CA 90089, USA
e-mail: rahul.jain@usc.edu

A. R. Ward
Marshall School of Business, University of Southern California, Bridge Hall, Trousdale Parkway,
Los Angeles, CA 90089-0809, USA
e-mail: amyward@marshall.usc.edu

a reflected diffusion process, and the zero process, with possible discontinuities during regime switches. The weak convergence results are established in the M_1 topology.

Keywords Queueing models · Transient queueing systems · Fluid and diffusion limits · Distributional approximations · Directional derivatives · M_1 topology

Mathematics Subject Classification 60K25 · 68M20 · 90B22 · 60F17

1 Introduction

Most of modern queueing theory is concerned with scenarios where arrival and service processes are stationary and ergodic. Renewal traffic models are a standard assumption in queueing theory. This is mathematically convenient as it allows full use of the tools that renewal theory and ergodic theory provide. However, it is not true in some queueing scenarios. For example, in some queueing scenarios, each arriving customer takes an independent decision of when to arrive. When we assume that every arriving customer draws an arrival time from the same distribution, this does not lead to a renewal arrival process. Moreover, such a distribution may only have finite support, meaning that the system is transient. This scenario does not fit the standard, single server models in queueing theory such as $M/M/1$, $M/G/1$, etc.

There has been an interest in developing a theory for non-stationary queues [1–6]. However, in almost all of these models, the assumption of a non-homogeneous Poisson arrival and service process remains ubiquitous. Recent work in [7,8] relaxes these assumptions. However, all these models assume a queueing system that operates forever with an infinite population of customers and (possibly) a steady state (when arrival and service rates are cyclostationary).

In contrast, many queueing systems serve only a finite number of customers, the queueing system itself may operate only in a finite window of time, or a modeler is interested only in the transient behavior of the system. Scenarios where such behavior is apparent include queueing outside stores before new product launches, DMV or postal offices, lunch cafeterias etc., some call centers where customers take independent decisions of when to call and service time is finite (8 a.m.–5 p.m., for example), and even emergency departments of hospitals, where day-of-week effects strongly indicate that a manager would want to study the queueing dynamics on a single day. In communication networks, single file transfers such as a video streaming session and packet transmissions over a fixed interval of interest are examples of systems where a modeler may wish to study transient delay distributions.

In this paper, we study a *transitory* queueing model for such systems. Consider n customers who arrive into a single server queue. Each customer's time of arrival is modeled as an i.i.d. sample from a distribution F (restrictions on F will be stated later), and customers enter the queue in order of the sampled times. Service times are i.i.d. with distribution G. If $X_{(i)}$ is the ith order statistic from a sample of size n drawn from F and $\Delta_{(i)} := (X_{(i)} - X_{(i-1)})$, then, in Kendall's notation, this model can be called the $\Delta_{(i)}/GI/1$ *queueing model*.

The analysis of the discrete event model is quite difficult, in general. For instance, when the service process is Poisson, the Kolmogorov forward equations for the joint distribution of the queue length and cumulative arrival processes can be written down, but there is no easy way to obtain analytical solutions. In this paper, we develop fluid and diffusion approximations to the queue length process directly as the population size scales to infinity and the service rate is *accelerated* appropriately (to be defined). We also establish a sample path Little's Law that links the limit queue length and virtual waiting time processes under both fluid and diffusion limits.

To develop the fluid limits, we use the Glivenko–Cantelli theorem and the functional Strong Law of Large Numbers for renewal processes along with the Skorokhod reflection mapping theorem. We show that the fluid limit of the queue length process switches between 'overloaded,' 'underloaded,' and 'critically loaded' regimes as time progresses. The limiting diffusion for the queue length process is derived using a directional derivative reflection mapping lemma. The diffusion process approximation is a reflection of a Brownian bridge process that arises from the invariance principle related to the Kolmogorov–Smirnov statistic, combined with a Brownian motion that arises from the functional central limit theorem for renewal processes.

We also note that our diffusion process convergence results are in Skorokhod's M_1 topology on $\mathcal{D}_{\lim}[0,\infty)$, the space of functions that are right- or left-continuous at every point, and right-continuous at 0.

The rest of this paper is organized as follows. We start with a brief review of the existing literature related to this model. Section 2 presents the $\Delta_{(i)}/GI/1$ queueing model and some basic results about fluid and diffusion approximations to arrival and service processes. Section 3 develops fluid approximations to the queue length, busy-time, and virtual waiting time processes. In Sect. 4, we develop diffusion approximations to these processes. Section 5 develops waiting time approximations, as well as a sample path Little's law. Section 6 takes a closer look at the sample paths of the queue length process in various operating regimes. Section 7 presents some examples and simulations of queue length process. We then conclude in Sect. 8 with some remarks about potential future directions. In the appendix, we place proofs that are more technical in nature.

1.1 Related literature

The *form* of the diffusion and fluid approximations to the $\Delta_{(i)}/GI/1$ queue parallel that of the well studied $M_t/M_t/1$ model in the sense that (1) the fluid limit may switch between overloaded, underloaded, and critically loaded periods, and (2) the diffusion limit arises using a directional derivative for the Skorokhod reflection map. Approximations for the latter model were developed in [6], wherein the Poisson arrival and service processes are approximated sample pathwise by Gaussian processes on an accelerated time scale, by leveraging strong approximation results for Lévy processes. We, instead, prove a weak convergence by utilizing the Skorokhod almost sure representation theorem to establish the desired results. Another important difference is that our fluid and diffusion limits depend on empirical process theory (i.e., the Glivenko–Cantelli and Kolmogorov–Smirnov theorems), whereas such results are not relevant in [6].

There have been earlier attempts to understand ‘transitory’ behavior in queueing systems. In the late 1960s [1] (also [9]), Newell introduced queueing models with both time-varying arrival and service processes. He studied the Fokker–Planck (or heat) equation for the Gaussian process approximation to a general arrival process in various special cases on the arrival rate function. However, these approximations were not rigorously justified with a weak convergence result. In [10], Gaver et al. discuss several transitory demand queueing problems and propose a model similar to a $\Delta_{(i)}/M/1$ queue. In [11], Louchard considers a similar model to the $\Delta_{(i)}/GI/1$ queue. The analysis focuses on the local behavior of the queue, similar to the analyses of Newell [1]. The author only establishes local weak convergence to Gaussian processes at continuity points of the limit process. Our results, on the other hand, establish a single “process-level” convergence result over all time and, indeed, this is the main difficulty in the analysis.

2 Preliminaries

Notations Unless noted otherwise, all intervals of time are subsets of $[-T_0, \infty)$, for a given $-T_0 \leq 0$ (where $-T_0$ represents the time the first instant a user can arrive; without loss of generality, we assume that service starts at 0). Let $\mathcal{D}_{\lim} := \mathcal{D}_{\lim}[-T_0, \infty)$ be the space of functions $x : [-T_0, \infty) \to \mathbb{R}$ that are right-continuous at $-T_0$, and are either right- or left-continuous at every point $t > -T_0$. Note that this differs from the usual definition of the space $\mathcal{D}$ as the space of functions that are right-continuous with left limits (cádlág functions). We denote almost sure convergence by $\xrightarrow{a.s.}$ and weak convergence by $\Rightarrow$. (S, m) represents the metric space and metric of convergence. Thus, $X_n \xrightarrow{a.s.} X$ in $(\mathcal{D}_{\lim}, J_1)$ as $n \to \infty$ indicates that $X_n \in \mathcal{D}_{\lim}$ converges to $X \in \mathcal{D}_{\lim}$ in the (strong) J_1 topology almost surely. Similarly, $X_n \Rightarrow X$ in $(\mathcal{D}_{\lim}, J_1)$ as $n \to \infty$ indicates that $X_n \in \mathcal{D}_{\lim}$ converges weakly to $X \in \mathcal{D}_{\lim}$ in the (strong) J_1 topology. $(\mathcal{D}_{\lim}, M_1)$ indicates that the topology of convergence is the M_1 topology. When convergence is joint for a collection of random variables, we will either be working with *strong* M_1 (SM_1) topology or the *weak* J_1 (WJ_1) topology on the product space of the sample paths (see [12] for formal definitions of these spaces). $\bar{X}$ indicates a fluid-scaled or fluid limit process. $\hat{X}$ and $\tilde{X}$ are used to indicate diffusion-scaled and diffusion limit processes. We use $\circ$ to denote the composition of functions or processes. The indicator function is denoted by $\mathbf{1}_{\{\cdot\}}$ and the positive part operator by $(\cdot)_+$.

2.1 The queueing model

Consider a single server, infinite buffer queue that is non-preemptive, non-idling, and starts empty. Service follows a first-come-first-served (FCFS) schedule. Let n be the customer population size. Customers independently sample an arrival time $T_i, i = 1, \ldots, n$, from a common distribution function F assumed to have support $[-T_0, T] \subset \mathbb{R}$, where $T > 0$. For simplicity, we assume that F is absolutely continuous with a continuous density function. The customer entry times are the order statistics $T_{(1)} \leq$

$T_{(2)} \leq \ldots \leq T_{(n)}$ of the sampled arrival times. The arrival process is the cumulative number of customers that have arrived by time t:

$$A(t) := \sum_{i=1}^{n} \mathbf{1}_{\{T_i \leq t\}}, \tag{1}$$

where $\mathbf{1}_{\{\cdot\}}$ represents an indicator function.

Let $\{\nu_i, i \geq 1\}$ be a sequence of independent and identically distributed (i.i.d.) random variables, where ν_i represents the service time of the ith customer. Assume that the mean service time $\mathbb{E}\nu_i = 1/\mu < \infty$ and the variance of the service times $\mathrm{Var}(\nu_i) < \infty$, and that the associated CDF G has support $[0, \infty)$. Finally, also assume that the sequence is independent of the arrival times T_i, $i = 1, \ldots, n$. Thus, service starts at time $t = 0$. Let S be the service process, defined as a renewal counting process, so that

$$S(t) := \begin{cases} 0 & \forall t \in [-T_0, 0), \\ \sup\{m \geq 1 | V(m) \leq t\}, & \forall t \geq 0, \end{cases} \tag{2}$$

where

$$V(m) := \sum_{i=1}^{m} \nu_i$$

is the cumulative load from m jobs. Let $V(t) := \sum_{i=1}^{\lfloor t \rfloor} \nu_i$ be the offered load process.

The amount of time a customer arriving at time t has to wait for service is

$$Z(t) := V(A(t)) - B(t) - t\mathbf{1}_{\{t \leq 0\}}, \tag{3}$$

where

$$B(t) := \left(\int_0^t \mathbf{1}_{\{Q(s)>0\}} \, \mathrm{d}s \right) \mathbf{1}_{\{t \geq 0\}}, \quad \forall t \in [-T_0, \infty) \tag{4}$$

is the *busy time* process.

Note that this definition of the virtual waiting time varies slightly from the standard definition due to the fact that an arrival at time $t < 0$ before service starts has to wait an extra t units of time for service to start, which accounts for the $-t\mathbf{1}_{\{t \leq 0\}}$ term.

Let Q represent the queue length process, including both any customer in service and all waiting customers. This is defined in terms of the arrival and service processes as

$$Q(t) := A(t) - S(B(t)), \quad \forall t \in [-T_0, \infty), \tag{5}$$

where $B(t)$ is the busy time process.

Finally, the idle time process of the server is

$$I(t) := t\mathbf{1}_{\{t \geq 0\}} - B(t) = \left(\int_0^t \mathbf{1}_{\{Q(s)=0\}} \, \mathrm{d}s \right) \mathbf{1}_{\{t \geq 0\}} \quad \forall t \in [-T_0, \infty). \tag{6}$$

2.2 Basic results

We now present known functional strong law of large numbers (FSLLN) and functional central limit theorem (FCLT) or diffusion limits, for the arrival and service processes, as the population size n increases to ∞.

Let $A^n := A$ be the arrival process associated with the system having population size n. The fluid-scaled arrival process is $\bar{A}^n := \frac{A^n}{n}$. Next, consider an *accelerated* service process, where the service times (or, equivalently, the service rate) are scaled by the population size n, so that

$$S^n(t) := \begin{cases} 0 & \forall t \in [-T_0, 0), \\ \sup\left\{ m \geq 1 | \sum_{i=1}^{m} \frac{\nu_i}{n} \leq t \right\}, & \forall t \geq 0. \end{cases}$$

The fluid-scaled service process is $\bar{S}^n := \frac{S^n}{n}$. Also, the fluid-scaled offered load process is

$$\bar{V}^n(t) := \begin{cases} 0 & \forall t \in [-T_0, 0), \\ \sum_{i=1}^{\lfloor nt \rfloor} \nu_i^n, & \forall t \in [0, \infty). \end{cases} \tag{7}$$

Note that our assumption that $\nu_i, i \geq 1$ is an i.i.d. sequence implies that $S^n(t)$ is equivalent to the time-scaled process $S(nt)$ (where n is an arbitrary parameter that increases to infinity) used in the conventional heavy-traffic setting. Acceleration, however, provides a nice interpretation to our scaling that we conjecture can potentially be extended to non-i.i.d. settings. The following proposition establishes the fluid limits for these processes.

Proposition 1 *As $n \to \infty$,*

$$(\bar{A}^n(t), \bar{S}^n(t)\mathbf{1}_{t\geq 0}, \bar{V}^n(t)\mathbf{1}_{t\geq 0}) \xrightarrow{a.s.} (F(t), \quad \mu t \mathbf{1}_{\{t\geq 0\}}, \quad \frac{t}{\mu}\mathbf{1}_{\{t\geq 0\}}) \text{ in } (\mathcal{D}^3_{\lim}, WJ_1), \tag{8}$$

where $\mathcal{D}^3_{\lim}$ is the three-dimensional product space of sample paths.

Remarks The proof of Proposition 1 follows easily from standard results and we omit it. The fluid arrival process limit is given by the Glivenko–Cantelli Theorem (see [13]). The fluid limits of the service process and the offered load process follow from the functional strong law of large numbers for renewal processes (see [14]). Joint convergence is a consequence of the independence assumptions between the service times and arrival times.

Next, looking at the errors of the fluid-scaled arrival process around the fluid limit, the diffusion-scaled arrival process is

$$\hat{A}^n(t) := \sqrt{n}\Big(\bar{A}^n(t) - F(t)\Big) \quad \forall t \in [-T_0, \infty).$$

Similarly, the diffusion-scaled service and offered load processes are

$$\hat{S}^n(t) := \sqrt{n}\left(\bar{S}^n(t) - \mu t\right), \quad t \geq 0$$
$$\hat{V}^n(t) := \sqrt{n}\left(\bar{V}^n(t) - \frac{1}{\mu}t\right), \quad t \geq 0.$$

The following proposition presents the diffusion limits for these processes.

Proposition 2 *As $n \to \infty$,*

$$(\hat{A}^n, \hat{S}^n, \hat{V}^n) \Rightarrow \big(W^0 \circ F, \sigma\mu^{3/2} W \circ e, -\sigma\mu^{1/2} W \circ \frac{e}{\mu}\big) \textit{ in } (\mathcal{D}^3_{\lim}, WJ_1), \tag{9}$$

where W^0 is the standard Brownian bridge process and W is the standard Brownian motion process, both are mutually independent, and $e : [0,\infty) \to [0,\infty)$ is the identity map.

Remarks (1) The proof of this proposition follows easily from standard results: The FCLT limit for the diffusion-scaled arrival process, also called the empirical process, is a Brownian bridge by Donsker's Theorem (see Sects. 13 and 16 in [15]). Note that this limit also arises in the study of the invariance principle associated with the Kolmogorov–Smirnov statistic used to compare empirical distributions with candidate ones (see [12] for more detail). The limits for the diffusion-scaled service and offered work processes follow from the FCLT for renewal processes (see Sect. 16 in [15] and Chap. 5 in [14]). Joint convergence follows from independence.

(2) Our assumption that the support of F is compact is largely for technical reasons; viz., the Skorokhod topologies restrict weak convergence to compact intervals of the domain $[-T_0, \infty)$. Proving a diffusion approximation that holds for distributions with infinite support would require strong approximation results, and is beyond the scope of the current paper.

3 Fluid approximations

Following (5), the fluid-scaled queue length process is

$$\frac{Q^n(t)}{n} = \frac{1}{n}A^n(t) - \frac{1}{n}S^n(B^n(t)), \tag{10}$$

where $B^n(t)$ is the fluid-scaled version of the busy time process (4) defined as

$$B^n(t) := \left(\int_0^t \mathbf{1}_{\{Q^n(s)>0\}}\, \mathrm{d}s\right)\mathbf{1}_{\{t\geq 0\}}.$$

Next, add and subtract the functions $F(t)$, $\mu t\mathbf{1}_{\{t\geq 0\}}$ and $\mu B^n(t)$ to obtain

$$\frac{Q^n(t)}{n} := \left(\frac{A^n(t)}{n} - F(t)\right) - \left(\frac{S^n(B^n(t)}{n} - \mu B^n(t)\right) + \left(F(t) - \mu t\mathbf{1}_{\{t\geq 0\}}\right) + \mu I^n(t),$$

where $I^n(t) = t\mathbf{1}_{\{t\geq 0\}} - B^n(t)$ is the fluid-scaled idle time process. Thus, (10) is equivalently

$$\bar{Q}^n(t) := \frac{Q^n(t)}{n} = \bar{X}^n(t) + \mu I^n(t), \quad \forall t \in [-T_0, \infty), \tag{11}$$

where $\bar{X}^n(t)$ is

$$\bar{X}^n(t) := \left(\frac{A^n(t)}{n} - F(t)\right) - \left(\frac{S^n(B^n(t))}{n} - \mu B^n(t)\right) + (F(t) - \mu t\mathbf{1}_{\{t\geq 0\}}). \tag{12}$$

In preparation for the main theorem in this section, recall that the Skorokhod reflection map is a continuous functional $(\Phi, \Psi) : \mathcal{D}_{\lim} \to \mathcal{D}_{\lim} \times \mathcal{D}_{\lim}$ defined as $x \mapsto \Psi(x) := \sup_{-T_0\leq s\leq t}(-x(s))_+$, and $x \mapsto \Phi(x) := x + \Psi(x)$, $\forall x \in \mathcal{D}_{\lim}$. The continuity of the map with respect to the uniform topology on $\mathcal{D}_{\lim}$ follows from Theorem 3.1 in [16].

Theorem 1 (Fluid limit) *The pair $(\bar{Q}^n, \mu I^n)$ has a unique representation $(\Phi(\bar{X}^n), \Psi(\bar{X}^n))$ in terms of $\bar{X}^n$. Furthermore, as $n \to \infty$,*

$$(\bar{Q}^n, \mu I^n) \xrightarrow{a.s.} (\Phi(\bar{X}), \Psi(\bar{X})) \ \ in\ (\mathcal{D}_{\lim} \times \mathcal{D}_{\lim}, WJ_1),$$

where $\bar{X}(t) = (F(t) - \mu t\mathbf{1}_{\{t\geq 0\}})$.

Proof First note that $\bar{Q}^n(t) \geq 0$, $\forall t \in [-T_0, \infty)$. It is also true that $I^n(-T_0) = 0$ and $dI^n(t) \geq 0$, $\forall t \in [-T_0, \infty)$. By definition of $I^n(t)$, it follows that $\int_{-T_0}^{\infty} \bar{Q}^n(t)dI^n(t) = 0$. Thus, by the Skorokhod reflection mapping theorem (first proved in [17]), the joint process $(\bar{Q}^n(t), \mu I^n(t))$ has a unique reflection mapping representation in terms of $\bar{X}^n(t)$ as $(\Phi(\bar{X}^n), \Psi(\bar{X}^n))$.

Note that by definition of $B^n(t) \leq t$ and from Proposition 1, it follows that $\left(\frac{S^n \circ B^n}{n} - \mu B^n\right) \xrightarrow{a.s.} 0$ in $(\mathcal{D}_{\lim}, J_1)$. Using this and Proposition 1, it follows that $\bar{X}^n \xrightarrow{a.s.} \bar{X}$ in $(\mathcal{D}_{\lim}, J_1)$, where $\bar{X} := (F(t) - \mu t\mathbf{1}_{\{t\geq 0\}})$. Using the limit derived above and the continuous mapping theorem, it follows that

$$(\bar{Q}^n, \mu I^n) = (\Phi(\bar{X}^n), \Psi(\bar{X}^n)) \xrightarrow{a.s.} (\Phi(\bar{X}), \Psi(\bar{X})) \ \text{ in } (\mathcal{D}_{\lim} \times \mathcal{D}_{\lim}, WJ_1).$$

□

Remarks (1) $\bar{X}$ is the difference between the fluid limits of the arrival and service processes, and is often referred to as the fluid limit of the *netput* process.

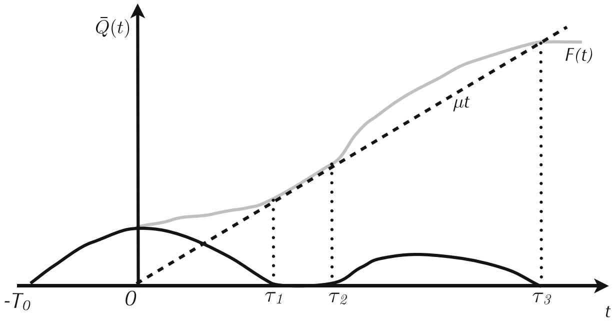

Fig. 1 An example of a $\Delta_{(i)}/GI/1$ queue that will undergo multiple "regime changes." The fluid queue length process is positive on $[-T_0, \tau_0)$ and $[\tau_2, \tau_3)$, and 0 on $[\tau_0, \tau_2)$ and $[\tau_3, \infty)$

(2) Theorem 1 shows that the fluid limit of the queue length process is

$$\bar{Q}(t) = (F(t) - \mu t \mathbf{1}_{\{t \geq 0\}}) + \sup_{-T_0 \leq s \leq t} (-(F(s) - \mu s \mathbf{1}_{\{s \geq 0\}}))_+, \quad \forall t \in [-T_0, \infty).$$

$\bar{Q}$ can be interpreted as the sum of the fluid netput process and the amount of fluid service capacity lost from the system. As it will be seen below, the time instants where the regulator term $\sup_{-T_0 \leq s \leq t}(-(F(s) - \mu s \mathbf{1}_{\{s \geq 0\}}))_+$ increases are precisely where the queue idles.

(3) Figure 1 depicts an example queue length process in the fluid limit, and its dependence on the arrival distribution F and service rate μ. Note that the process switches between being positive and zero, during the time the server operates. We will investigate this behavior in detail in Sect. 6. Without formally defining the terms, intuitively it should be clear that on $[-T_0, \tau_0)$ and $[\tau_2, \tau_3)$, the queue is 'overloaded,' while on the intervals $[\tau_0, \tau_1)$ and $[\tau_3 \infty)$, it is 'underloaded.'

Next, consider the busy time process. It is interesting to observe that B^n does not converge to the identity process, in contrast to the conventional heavy-traffic approximation setting.

Corollary 1 *As* $n \to \infty$,

$$B^n \xrightarrow{a.s.} \bar{B} \text{ in } (\mathcal{D}_{\lim}, J_1) \tag{13}$$

where $\bar{B}(t) := t\mathbf{1}_{\{t \geq 0\}} - \frac{1}{\mu}\Psi(\bar{X}(t))$, $\forall t \in [-T_0, \infty)$.

Proof By definition, we have $B^n(t) = t\mathbf{1}_{\{t \geq 0\}} - I^n(t)$. This can be rewritten as $B^n(t) = t\mathbf{1}_{\{t \geq 0\}} - I^n(t)$. Using Theorem 1, the claim then follows. □

Note that $\bar{B}(t) = 0$ for all $t \leq 0$, as $\Psi(\bar{X})(t) = 0$ on that interval.

4 Diffusion approximations

In this section, we assume F is absolutely continuous in order to establish the desired limit result. As noted before, this is mainly for simplicity of the analysis.

4.1 Queue length process

Define the *diffusion-scaled queue length process* as

$$\frac{Q^n(t)}{\sqrt{n}} := \frac{A^n(t)}{\sqrt{n}} - \frac{S^n(B^n(t))}{\sqrt{n}}, \quad \forall t \in [-T_0, \infty) \tag{14}$$

Introducing the terms $\sqrt{n}\mu t\mathbf{1}_{\{t\geq 0\}}$, $\sqrt{n}F(t)$, and $\sqrt{n}\mu B^n(t)$, we have

$$\begin{aligned}\frac{Q^n(t)}{\sqrt{n}} = \left(\frac{A^n(t)}{\sqrt{n}} - \sqrt{n}F(t)\right) - \left(\frac{S^n(B^n(t))}{\sqrt{n}} - \sqrt{n}\mu B^n(t)\right)\\ + \sqrt{n}(F(t) - \mu t\mathbf{1}_{\{t\geq 0\}}) + \sqrt{n}\mu(t\mathbf{1}_{\{t\geq 0\}} - B^n(t)).\end{aligned}$$

Recalling the definition of the idle time process $Q^n/\sqrt{n}$ is

$$\frac{Q^n}{\sqrt{n}} = \hat{X}^n + \sqrt{n}\bar{X} + \sqrt{n}\mu I^n, \tag{15}$$

where

$$\begin{aligned}\hat{X}^n(t) &:= \left(\frac{A^n(t)}{\sqrt{n}} - \sqrt{n}F(t)\right) - \left(\frac{S^n(B^n(t))}{\sqrt{n}} - \sqrt{n}\mu B^n(t)\right)\\ &= \hat{A}^n(t) - \hat{S}^n(B^n(t)), \quad \forall t \in [-T_0, \infty).\end{aligned} \tag{16}$$

Recall from Theorem 1 that $\bar{X}(t) = (F(t) - \mu t\mathbf{1}_{t\geq 0})$ is the fluid netput process. Lemma 1 below proves a diffusion approximation to the diffusion-scale refinement $\hat{X}^n(t)$ as an immediate consequence of Proposition 2.

Lemma 1 *As $n \to \infty$,*

$$\hat{X}^n \Rightarrow \hat{X} := W^0 \circ F - \sigma\mu^{3/2} W \circ \bar{B} \;\; in\; (\mathcal{D}_{\lim}, J_1) \tag{17}$$

where $\bar{B}$ is defined in (13), and W^0 and W are independent standard Brownian bridge and standard Brownian motion, respectively.

Proof First note that $B^n(t) \leq t, \forall t \in [0, \infty)$, implying that $S^n \circ B^n \in \mathcal{D}_{\lim}$. Using Proposition 2, Corollary 1 and the random time change theorem (see, for example, Sect. 17 of [15]), it follows that $\sqrt{n}\left(\frac{S^n \circ B^n}{n} - \mu B^n\right) \Rightarrow \sigma\mu^{3/2} W \circ \bar{B}$. Now, it follows from Proposition 2 that $\hat{X}^n \Rightarrow \hat{X}(t) := W^0 \circ F - \sigma\mu^{3/2} W \circ \bar{B}$, thus concluding the proof. □

Remarks Note that using a classical time change (see, for example, [18]), it is possible to see that the Brownian bridge is equal in distribution to a time-changed Brownian motion, and $\hat{X}$ is equal in distribution to a stochastic integral

$$\hat{X}(t) \overset{d}{=} \begin{cases} \int_{-T_0}^{t} \sqrt{g'(s)} d\tilde{W}_s, & \forall t \in [-T_0, T] \\ -\sigma\mu^{3/2} W(\bar{B}(\tau^* \vee T)), & \forall t > \tau^* \vee T \end{cases}, \tag{18}$$

where $g(t) = F(t)(1-F(t)) + \sigma^2\mu^3\bar{B}(t)$, $\tilde{W}$ is a standard Brownian motion process, $\tau^* := \frac{1}{\mu}$ and $\vee$ is the max operator. Thus, the process $\hat{X}$ can also be interpreted as a time-changed Brownian motion on the interval $[-T_0, T]$, and its sample path is a constant on (T, ∞).

In the rest of this section, we will use Skorokhod's almost sure representation theorem [17,19], and replace the random processes above that converge in distribution by those defined on a common probability space that have the same distribution as the original processes and converge almost surely. The requirements for the almost sure representation are mild; it is sufficient that the underlying topological space is Polish (a separable and complete metric space). We note without proof that the space $\mathcal{D}_{\lim}$, as defined in this paper, is Polish when endowed with the M_1 topology. This conclusion follows from [12]. The authors in [6] also point out that [20] has a more general proof of this fact.

We conclude that we can replace the weak convergence in (9) by

$$(\hat{A}^n, \hat{S}^n, \hat{V}^n) \xrightarrow{a.s.} \left(W^0 \circ F, \sigma\mu^{3/2}W, -\sigma\mu^{1/2}W \circ \frac{h}{\mu} \right) \text{ in } (\mathcal{D}_{\lim}, J_1),$$

where, abusing notation, we use the same letters as our original processes. Thus, Lemma 1 implies that

$$\hat{X}^n \xrightarrow{a.s.} \hat{X} \text{ in } (\mathcal{D}_{\lim}, J_1), \quad \text{as } n \to \infty.$$

The FCLT to the queue length process relies on the directional derivative of the Skorokhod reflection map (Φ, Ψ), defined as

$$\sup_{\nabla_t^{\bar{X}}}(-y)(t) = \lim_{n\to\infty} \Psi(\sqrt{n}x + y)(t) - \sqrt{n}\Psi(x)(t), \tag{19}$$

pointwise in $\mathcal{D}_{\lim}$, where $x \in \mathcal{C}$ and $y \in \mathcal{C}$, and $\nabla_t^x = \{-T_0 \le s \le t | x(s) = -\Psi(x)(t)\}$, is a correspondence of points upto time t where the fluid netput process achieves an infimum. We can now state and prove our main limit theorem. Let $\tilde{Y}^n := \sqrt{n}\mu I^n - \sqrt{n}\Psi(\bar{X})$.

Theorem 2 (Diffusion limit) *The pair $(\hat{Q}^n, \tilde{Y}^n)$ has a unique representation in terms of $\hat{X}^n$ and $\sqrt{n}\bar{X}$ given by $\big(\Phi(\hat{X}^n + \sqrt{n}\bar{X}) - \sqrt{n}\bar{Q}, \Psi(\hat{X}^n + \sqrt{n}\bar{X}) - \sqrt{n}\Psi(\bar{X})\big)$, where $\bar{Q} = \bar{X} + \Psi(\bar{X})$ is the fluid limit of the queue length process. Furthermore, as $n \to \infty$*

$$(\hat{Q}^n, \tilde{Y}^n) \Rightarrow (\hat{X} + \tilde{Y}, \tilde{Y}) \text{ in } (\mathcal{D}_{\lim} \times \mathcal{D}_{\lim}, SM_1),$$

where $\hat{X}(t) = W^0(F(t)) - \sigma\mu^{3/2}W(\bar{B}(t))$, and $\tilde{Y}(t) = \max_{s\in\nabla_t^{\bar{X}}}(-\hat{X}(s))$ $\forall t \in [-T_0, \infty)$, and SM_1 is the strong *M_1 topology on the product space $\mathcal{D}_{\lim} \times \mathcal{D}_{\lim}$.*

Proof First, using (15), it follows by the Skorokhod reflection mapping theorem that

$$\left(\frac{Q^n}{\sqrt{n}}, \sqrt{n}\mu I^n\right) = \left(\Phi(\hat{X}^n + \sqrt{n}\bar{X}), \Psi(\hat{X}^n + \sqrt{n}\bar{X})\right). \tag{20}$$

This implies that $\hat{Q}^n = \frac{Q^n}{\sqrt{n}} - \sqrt{n}\bar{Q} = \Phi(\hat{X}^n + \sqrt{n}\bar{X}) - \sqrt{n}\bar{Q}$. Using the fact that $\bar{Q} = \bar{X} + \Psi(\bar{X})$ and $\Phi(x) = x + \Psi(x)$ for any $x \in \mathcal{D}_{\text{lim}}$, it follows that

$$\begin{aligned}\hat{Q}^n &= \hat{X}^n + \sqrt{n}\bar{X} + \Psi(\hat{X}^n + \sqrt{n}\bar{X}) - \sqrt{n}(\bar{X} + \Psi(\bar{X})),\\ &= \hat{X}^n + \Psi(\hat{X}^n + \sqrt{n}\bar{X}) - \sqrt{n}\Psi(\bar{X}).\end{aligned} \tag{21}$$

Next, from the expression for $\sqrt{n}\mu I^n$ in (20) it follows that $\tilde{Y}^n = \Psi(\hat{X}^n + \sqrt{n}\bar{X}) - \sqrt{n}\Psi(\bar{X})$, implying that $\hat{Q}^n = \hat{X}^n + \tilde{Y}^n$. The limit result now follows by use of the following directional derivative reflection mapping lemma which is adapted from Lemma 5.2 in [6], and whose proof can be found in the Appendix. □

Lemma 2 (Directional derivative reflection mapping lemma) *Let x and y be real-valued continuous functions on $[0, \infty)$, and $\Psi(z)(t) = \sup_{0 \le s \le t}(-z(s))$, for any process $z \in \mathcal{D}_{\text{lim}}$. Let $\{y_n\} \subset \mathcal{D}_{\text{lim}}$ be a sequence of functions such that $y_n \overset{a.s.}{\to} y$ as $n \to \infty$. Then, with respect to Skorokhod's M_1 topology, $\tilde{y}_n := \Psi(\sqrt{n}x + y_n) - \sqrt{n}\Psi(x) \longrightarrow \tilde{y} := \sup_{s \in \nabla_t^x}(-y(s))$ as $n \to \infty$, where $\nabla_t^x = \{0 \le s \le t | x(s) = -\Psi(x)(t)\}$.*

Observe that $\tilde{Y}_n$ is exactly in the form of $\tilde{y}_n$ defined in the lemma above. Lemma 1 and Lemma 2 together imply that $\tilde{Y}_n \xrightarrow{a.s.} \tilde{Y} := \max_{s \in \nabla_{\cdot}^{\bar{X}}}(-\hat{X}(s))$ in $(\mathcal{D}_{\text{lim}}, M_1)$. It follows that $\hat{Q}^n = \hat{X}^n + \tilde{Y}^n \xrightarrow{a.s.} \hat{X} + \max_{s \in \nabla_{\cdot}^{\bar{X}}}(-\hat{X}(s))$ in $(\mathcal{D}_{\text{lim}}, M_1)$.

It remains to prove that $\hat{Q}^n$ and $\tilde{Y}^n$ converge jointly in the strong M_1, or SM_1, topology. Notice that the joint process can be written as

$$\begin{pmatrix}\hat{Q}^n\\ \tilde{Y}^n\end{pmatrix} = \begin{pmatrix}\hat{X}^n\\ 0\end{pmatrix} + \begin{pmatrix}\Psi(\hat{X}^n + \sqrt{n}\bar{X}) - \sqrt{n}\Psi(\bar{X})\\ \Psi(\hat{X}^n + \sqrt{n}\bar{X}) - \sqrt{n}\Psi(\bar{X})\end{pmatrix}.$$

The first term on the right-hand side converges to

$$\hat{\mathbf{X}} := \begin{pmatrix}\hat{X}\\ 0\end{pmatrix}$$

almost surely in $(\mathcal{D}_{\text{lim}} \times \mathcal{D}_{\text{lim}}, SM_1)$ by Theorem 12.6.1 of [12], as $\hat{X}$ is continuous. The second term converges to

$$\tilde{\mathbf{Y}} := \begin{pmatrix}\tilde{Y}\\ \tilde{Y}\end{pmatrix}$$

almost surely in $(\mathcal{D}_{\text{lim}} \times \mathcal{D}_{\text{lim}}, SM_1)$. Now, by definition, $\hat{\mathbf{X}}$ is a continuous process and does not share any discontinuity points with $\tilde{\mathbf{Y}}$. Therefore, by Corollary 12.7.1 of [12], the addition operator is continuous, implying that

$$\begin{pmatrix}\hat{Q}^n\\ \tilde{Y}^n\end{pmatrix} \overset{a.s.}{\to} \begin{pmatrix}\hat{X} + \tilde{Y}\\ \tilde{Y}\end{pmatrix}$$

in $(\mathcal{D}_{\lim} \times \mathcal{D}_{\lim}, SM_1)$. Finally, the weak convergence is a direct implication of the almost sure convergence result, thus concluding the proof.

Remarks (1) Observe that the diffusion limit to the queue length process is a function of a Brownian bridge and a Brownian motion. This is significantly different from the usual limits obtained in a heavy-traffic or large population approximation to a single server queue. For instance, in the $G/GI/1$ queue, one would expect a reflected Brownian motion in the heavy-traffic setting. In [6], it was shown that the diffusion limit process to the $M_t/M_t/1$ queue is a time-changed Brownian motion $W(\int \lambda(s)\mathrm{d}s + \int \mu(s)\,\mathrm{d}s)$, where $\lambda(s)$ is the time inhomogeneous rate of arrival process and $\mu(s)$ is that of the service process, reflected through the directional derivative reflection map used in Lemma 2. There are very few examples of heavy-traffic limits involving a diffusion that is a function of a Brownian bridge and a Brownian motion process. However, there have been some results in other queueing models where a Brownian bridge arises in the limit. In [21], for instance, a Brownian bridge limit arises in the study of a many-server queue in the Halfin–Whitt regime.

(2) We noted in the remarks after Theorem 1 that the fluid limit can change between being positive and zero in the arrival interval for a completely general F. One can then expect the diffusion limit to change as well, and switch between being a 'free' diffusion, a reflected diffusion, and a zero process. This is indeed the case. Figure 2 illustrates this for the example in Figure 1. Note that $\forall t \in [-T_0, \tau_1)$, $\Psi(\bar{X})(t) = -\bar{X}(-T_0)$, implying that the set $\nabla_t^{\bar{X}}$ is a singleton. On the other hand, at τ_1, $\nabla_t^{\bar{X}} = \{-T_0, \tau_1\}$. For $t \in (\tau_1, \tau_2]$, $\Psi(\bar{X})(t) = 0 = \bar{X}(t)$, implying that $\nabla_t^{\bar{X}} = (\tau_1, t]$. On (τ_2, τ_3), $\Psi(\bar{X})(t) = 0$, but $\bar{X}(t) > 0$, so that $\nabla_t^{\bar{X}} = (\tau_1, \tau_2]$. Finally, the fluid queue length becomes zero when the fluid service process exceeds the fluid arrival process in $[\tau_3, \infty)$, implying that $\Psi(\bar{X})(t) = -(F(t) - \mu t) > 0$. It can be seen that $\nabla_t^{\bar{X}} = \{t\}$ in this case.

Recall from Corollary 1 that B^n converges to a continuous process $\bar{B}$ as $n \to \infty$. Define the diffusion-scaled busy time process as

$$\hat{B}^n := \sqrt{n}(\bar{B} - B^n). \tag{22}$$

Note that from the definitions of $B^n(t)$ and $\bar{B}(t)$, it follows that $\hat{B}^n(t) = 0,\ \forall t < 0$. The diffusion limit for this process is given as follows.

Corollary 2 *The diffusion-scaled busy time weakly converges to a regulated diffusion process:* $\hat{B}^n \Rightarrow \hat{B} := \frac{1}{\mu} \max_{s \in \nabla_\cdot^{\bar{X}}}(-\hat{X}(s))$, *in* $(\mathcal{D}_{\lim}, M_1)$ *as* $n \to \infty$.

Proof Recall that $B^n(t) = t\mathbf{1}_{\{t \geq 0\}} - I^n(t)$. Substituting this and $\bar{B}$ from (13) in the definition of $\hat{B}^n$, and rearranging the expression, we obtain $\hat{B}^n = \frac{1}{\mu}\tilde{Y}^n$. A simple application of Theorem 2 then provides the necessary conclusion. □

Observe that $B^n(t)$ is approximated in distribution by $\hat{B}$ as $B^n(t) \overset{d}{\approx} \bar{B}(t) - \frac{1}{\sqrt{n}}\hat{B}(t)$, where $Y \overset{d}{\approx} X$ is defined to be $\mathbb{P}(Y \leq x) \approx \mathbb{P}(X \leq x)$, and the approximation is rigorously supported by an appropriate weak convergence result.

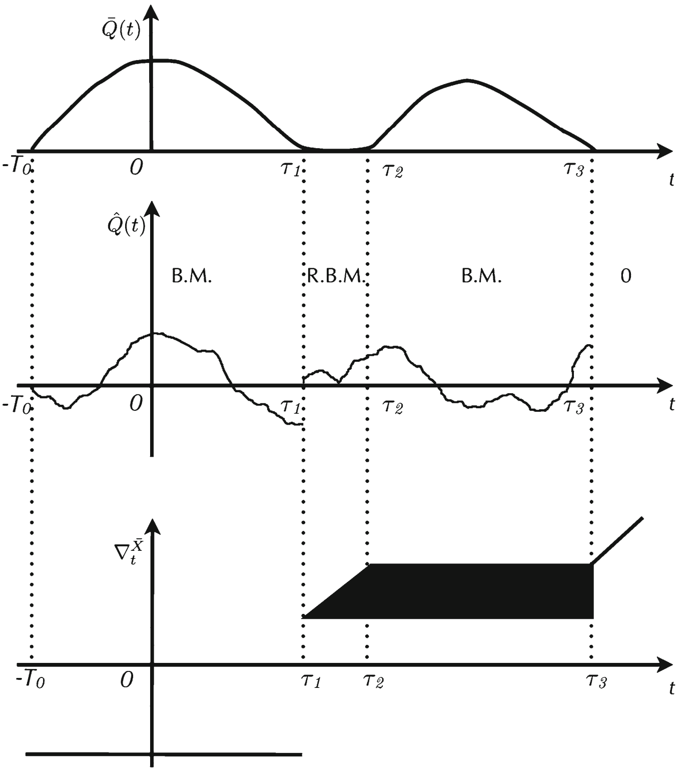

Fig. 2 An example of a $\Delta_{(i)}/GI/1$ queue that will undergo multiple "regime changes." The diffusion limit switches between a free Brownian motion (BM), a reflected Brownian motion (RBM), and the zero process

The case of uniform F on $[-T_0, T]$ is instructive, and it can be seen that on $[-T_0, \tau)$, the queue length in the fluid limit is positive. However, as the server starts at time 0, the only interesting sub-interval of $[-T_0, \tau)$ is $[0, \tau)$. Using the appropriate definitions, note that $\bar{B}(t) = t$ and $\hat{B}(t) = 0$ for all $t \in [0, \tau)$, implying that $B^n(t) = t$ approximately, though in the non-asymptotic regime $B^n(t)$ may be strictly smaller than t. On the other hand, the fluid queue length is zero in (τ, ∞) and it follows from definition of $\Psi(\bar{X})$ that $\bar{B}(t) = t - \frac{1}{\mu}(-\bar{X}(t)) = \frac{1}{\mu}F(t)$ for $t \in (\tau, \infty)$. Substituting this expression together with that of $\hat{B}$, and expanding $\hat{X}$, we see that

$$B^n(t) \overset{d}{\approx} t + \frac{1}{\mu}(\bar{X}(t) + \frac{1}{\sqrt{n}}\hat{X}(t)) \overset{d}{=} \frac{1}{\mu}\left(F(t) + \frac{1}{\sqrt{n}}W^0(F(t)) - \sigma\mu W(F(t))\right),$$

where the second $\overset{d}{=}$ is due to the fact that we used the Brownian motion scaling property. Note that this depends on the arrival distribution F alone. In the fluid limit of the busy time process, we see that $\bar{B}(t) = F(t)/\mu$ which is the fraction of time from the interval $[0, t]$ that the queue has spent serving.

5 Waiting time and the sample path Little's Law

Little's Law is a fundamental tenet of queueing theory that provides immediate insight into the operation of a queue. While the standard Little's Law relates averages, in this section, we prove a large population asymptotic functional relationship that holds on sample paths of the queue length and workload approximations. One may also view this 'sample path Little's Law' as parallel to a snapshot principle in the conventional heavy-traffic setting.

First, the accelerated or fluid-scaled virtual waiting time process is $Z^n(t) = V^n\big(n\big(\frac{A^n(t)}{n}\big)\big) - B^n(t) - t\mathbf{1}_{\{t\le 0\}}$, $\forall t \in [-T_0, \infty)$.

Proposition 3 (Fluid Little's Law) *The fluid-scale workload process is asymptotically related to the queue length fluid limit as $n \to \infty$: $Z^n \xrightarrow{a.s.} \bar{Z} := \bar{Q}/\mu - e$ in $(\mathcal{D}_{\text{lim}}, J_1)$, where $e : \mathbb{R} \to [0, \infty)$ is defined as $e(t) := t\mathbf{1}_{\{t\le 0\}}$ $\forall t \in \mathbb{R}$.*

Proof First note that $Z^n(t)$ can be rewritten as $Z^n(t) = V^n\big(n\big(\frac{A^n(t)}{n}\big)\big) - \frac{1}{\mu}\frac{A^n(t)}{n} + \big(\frac{1}{\mu}\frac{A^n(t)}{n} - t\mathbf{1}_{\{t\le 0\}} - B^n(t)\big)$. Proposition 1 implies that $\bar{V}^n(t) \xrightarrow{a.s.} t/\mu$ in $(\mathcal{D}_{\text{lim}}, J_1)$. Now, using the random time change theorem (Theorem 5.3 in [14]) and setting $h = A^n/n$ it follows that, as $n \to \infty$, $\big(V^n \circ A^n - \frac{1}{\mu}\frac{A^n}{n}\big) \xrightarrow{a.s.} 0$ in $(\mathcal{D}_{\text{lim}}, J_1)$. Using Proposition 1 and Corollary 1, substituting for $\bar{B}(t)$, we have $\bar{Z}(t) = \frac{1}{\mu}\bar{Q}(t) - t\mathbf{1}_{\{t\le 0\}}$.□

Remarks The term $e(t) = t\mathbf{1}_{\{t\le 0\}}$ accounts for the fact that an arrival at time $t < 0$ would require $-t$ time units for service to start. Now, consider the diffusion-scale virtual waiting time process given by $\hat{Z}^n(t) = \sqrt{n}(Z^n(t) - \bar{Z}(t))$ $\forall t \in [-T_0, \infty)$. Proposition 4 below proves a diffusion approximation to $\hat{Z}^n$ and relates the sample paths of the limit process to that of $\hat{Q}$.

Proposition 4 (Diffusion Little's Law) *The diffusion-scaled virtual waiting time process satisfies an FCLT in the limit as $n \to \infty$: $\hat{Z}^n \Rightarrow \hat{Z} := \frac{1}{\mu}\hat{Q} + \sigma\mu^{1/2}W \circ \bar{B} - \sigma\mu^{1/2}W \circ F$ in $(\mathcal{D}_{\text{lim}}, M_1)$.*

Proof Expanding the definition of $\hat{Z}^n(t)$ and introducing the term $\frac{1}{\mu}\frac{A^n(t)}{n}$, we obtain $\hat{Z}^n(t) = \sqrt{n}\big(V^n(A^n(t)) - \frac{1}{\mu}\frac{A^n(t)}{n} + \frac{1}{\mu}\frac{A^n(t)}{n} - \frac{F(t)}{\mu} + \bar{B}(t) - B^n(t)\big)$. Using the random time change theorem (Sect. 17 of [15]), Proposition 1 and Proposition 2

$$\sqrt{n}\left(V^n \circ A^n - \frac{1}{\mu}\frac{A^n}{n}\right) \Rightarrow -\sigma\mu^{1/2}W \circ \frac{F}{\mu} \text{ in } (\mathcal{D}_{lim}, J_1). \tag{23}$$

Finally, using this fact, Proposition 2 and Corollary 2, it follows that $\hat{Z}^n \Rightarrow \hat{Z} = \sigma\mu^{1/2}W \circ \frac{F}{\mu} + \frac{1}{\mu}W^0 \circ F + \hat{B}$ in $(\mathcal{D}_{\text{lim}}, M_1)$.

Note that W and W^0 are independent processes. Adding and subtracting the process $\sigma\mu^{1/2}W \circ \bar{B}$, where W is the Brownian motion in (23), we obtain $\hat{Z} = \frac{1}{\mu}\hat{Q} + \big(\sigma\mu^{1/2}W \circ \bar{B} - \sigma\mu^{1/2}W \circ \frac{F}{\mu}\big)$. □

Remarks (1) The limit process in Proposition 4 is equal to

$$\hat{Z}(t) = \frac{1}{\mu}\hat{Q}(t) - \sigma\mu^{1/2} W\Big(\frac{\bar{Q}(t)}{\mu}\Big). \tag{24}$$

Interestingly, the extra diffusion term is non-zero only when the fluid limit of the queue length process is positive, indicating that it arises from temporal variations in the operating regimes of the queue. To see this, note that the variance of the diffusion term is $\sigma^2\mu\,\mathbb{E}\big|W(\bar{B}(t)) - W\big(\frac{F(t)}{\mu}\big)\big|^2 = \sigma^2\mu\big(\bar{B}(t) + \frac{F(t)}{\mu} - 2\bar{B}(t)\wedge\frac{F(t)}{\mu}\big)$, where $x \wedge y := \min(x, y)$. Clearly, the expression on the right-hand side changes depending upon the ratio of the number of users arrived to the number served in the fluid regime at time t. It follows that

$$\sigma^2\mu\,\mathbb{E}\bigg|W(\bar{B}(t)) - W\bigg(\frac{F(t)}{\mu}\bigg)\bigg|^2 = \begin{cases} \sigma^2\mu\bigg(\frac{F(t)}{\mu} - \bar{B}(t)\bigg), & \frac{F(t)}{\mu\bar{B}(t)} > 1 \\ \sigma^2\mu\bigg(\bar{B}(t) - \frac{F(t)}{\mu}\bigg), & \frac{F(t)}{\mu\bar{B}(t)} \le 1. \end{cases}$$

It is easy to see that the first condition above, $F(t)/(\mu\bar{B}(t)) > 1$, implies $\bar{Q}(t)/\mu > 0$. The second condition, $F(t)/(\mu\bar{B}(t)) \le 1$, implies $\bar{Q}(t) = 0$. This in turn implies $(F(t) - \mu t\mathbf{1}_{\{t\ge 0\}}) + \Psi(F(t) - \mu t\mathbf{1}_{\{t\ge 0\}}) = 0$. Rearranging this expression, it follows that $F(t) = \mu t\mathbf{1}_{\{t\ge 0\}} - \Psi(F(t) - \mu t\mathbf{1}_{\{t\ge 0\}})$.

Now, using the definition of $\bar{B}$ from (13) we have $F(t)/(\mu\bar{B}(t)) = 1$. It follows that the diffusion term is equal in distribution to the following (time-changed) Brownian Motion

$$\begin{aligned}&\sigma\mu^{1/2}\bigg(W(\bar{B}(t)) - W\bigg(\frac{F(t)}{\mu}\bigg)\bigg) \\ &\overset{d}{=} \begin{cases} \sigma\mu^{1/2}W\bigg(\frac{F(t)}{\mu} - \bar{B}(t)\bigg) = \sigma\mu^{1/2}W\bigg(\frac{\bar{Q}(t)}{\mu}\bigg), & \bar{Q}(t) > 0 \\ \sigma\mu^{1/2}W\bigg(\bar{B}(t) - \frac{F(t)}{\mu}\bigg) = 0, & \bar{Q}(t) = 0. \end{cases}\end{aligned}$$

This leads to expression (24).

(2) We note that $\hat{Z}$ can be interpreted as a sample path Little's Law in the diffusion limit. This result is useful because it provides a sample path relationship between the workload and current queue state. Note that the FCLT of the workload process in a $G/GI/1$ queue (see Chap. 6 of [14] for details) with arrival rate λ and service rate μ has the form $\tilde{Z}(t) = \frac{1}{\mu}\hat{Q}(t) + \sigma\mu^{1/2}(W((\rho \wedge 1)t) - W(\rho t))$, where $\rho = \lambda/\mu$ is the traffic intensity function for the $G/GI/1$ queue, and this is similar to $\hat{Z}$. The extra diffusion term in (24) captures the variation of the workload, as the (fluid) queue transitions between various operating states (see Sect. 6 for more details on these states).

(3) Another interpretation of the term $\sigma\mu^{1/2}W(\bar{Q}(t)/\mu)$ is that it is in fact the diffusion limit to the service backlog at time t, and the variation in the backlog at each

point in time is captured in the term $\hat{Q}/\mu$. Suppose that $f(t) < \mu$ then the fluid queue length process is zero and the server will idle, and the zero state is recurrent for the queue length process. The workload in the system (for most of the time when $f(t) < \mu$) should be 0. On the other hand, if $F(t) = \mu t$, so that the fluid queue length is zero but the server does not idle, it is reasonable to expect that the virtual waiting time is zero for an arrival at time t. However, there is a non-zero probability of the queue being backlogged at time t, and this fact is captured in the term $\hat{Q}/\mu$.

6 Queue regimes and states

As noted in Sect. 4, the diffusion limit for the queue length process is piecewise continuous, with discontinuity points determined by the fluid limit. Indeed, the discontinuity points are precisely where the fluid limit switches between being 'overloaded' and either 'underloaded' and/or 'critically loaded.' We now provide formal definitions of these notions, in terms of the fluid limit arrival and service processes.

We also characterize the sample path of the queue length limit process, and the points at which it has discontinuities. Developments in this section follow the study of the directional derivative limit process in [6]. However, the limit processes and the setting of our model are different, as our limit process is a function of a tied down Gaussian process, while in [6] the limit process is a function of a standard Brownian motion. Thus, where necessary, we prove some of the facts about the sample paths.

6.1 Regimes of $\bar{Q}$

It is useful to characterize the state of a queue in terms of a "traffic intensity" measure. For instance, in the case of a $G/G/1$ queue, the traffic intensity is well defined as the ratio of the arrival rate to the service rate. This definition is inappropriate for the $\Delta_{(i)}/GI/1$ queue, as these systems can be time varying. In [2], a traffic intensity function for the $M_t/M_t/1$ queue with arrival rate $\lambda(\cdot)$ and service rate $\mu(\cdot)$ was introduced as the continuous function

$$\rho^*(t) := \sup_{0 \le r \le t} \frac{\int_r^t \lambda(u)du}{\int_r^t \mu(u)du}, \quad t > 0.$$

Note that ρ^* follows from the *pre-limit* model describing the arrival and service processes in the $M_t/M_t/1$ queue.

For the $\Delta_{(i)}/GI/1$ queue, we define the traffic intensity in terms of the fluid limit:

$$\rho(t) := \begin{cases} \infty, & \forall t \in [-T_0, 0] \\ \sup_{0 \le r \le t} \frac{F(t)-F(r)}{\mu(t-r)}, & \forall t \in [0, \tilde{T}] \\ 0, & \forall t > \tilde{T}, \end{cases} \tag{25}$$

where $\tilde{T} := \inf\{t > 0 | F(t) = 1 \text{ and } \bar{Q}(t) = 0\}$. Note that we define the traffic intensity to be ∞ in the interval $[-T_0, 0]$ as there is no service, but there can be fluid arrivals.

For example, with F uniform over $[-T_0, T]$, ρ can be shown to be

$$\rho(t) = \frac{t \wedge T}{t} \frac{1}{\mu(T + T_0)}, \quad \forall t \in [0, \tilde{T}].$$

Note that ρ is continuous in time. Now, consider the following obvious definitions of the operating regimes of the fluid $\Delta_{(i)}/GI/1$ queue.

Definition 1 *(Operating regimes)* The $\Delta_{(i)}/GI/1$ queue is (at time t)

(1) *overloaded* if $\rho(t) > 1$.
(2) *critically loaded* if $\rho(t) = 1$.
(3) *underloaded* if $\rho(t) < 1$.

The operating regimes can also be referenced in terms of the process $\bar{Q}$, which in many instances is more intuitive. The following lemma presents this equivalence.

Lemma 3 *The $\Delta_{(i)}/GI/1$ queue is*

(1) *Overloaded at time t if $\bar{Q}(t) > 0$.*
(2) *Critically loaded at time t if $\bar{Q}(t) = 0$, $\bar{X}(t) = \Psi(\bar{X})(t)$, and there exists an $r < t$ such that $\Psi(\bar{X})(t) = \Psi(\bar{X})(s)$ for all $s \in [r, t]$.*
(3) *Underloaded at time t if $\bar{Q} = 0$, $\bar{X}(t) = \Psi(\bar{X})(t)$, and there exists an $r < t$ such that $\Psi(\bar{X})(t) > \Psi(\bar{X})(s)$ for all $s \in (r, t)$.*

The proof of the lemma is in the appendix. Figure 3 shows an example of the various operating regimes with the displayed arrival time distribution F and service rate $\mu > 1/T$. Here, BB refers to a Brownian Bridge process and BM refers to a Brownian motion process. Theorem 2 proved a diffusion limit to the standardized queue length process, and we have shown that

$$Q^n \overset{d}{\approx} L_n \bar{Q} + \sqrt{L_n} \hat{Q}.$$

As noted in the remarks after Theorem 2, the queue length process switches between being a ‘free’ diffusion *BB+BM* (when the fluid limit model is overloaded), to a ‘reflected’ diffusion *R(BB+BM)* (when the fluid limit model is critically loaded) and to a ‘zero’ process *0* (when the fluid limit model is underloaded).

Notice that these regimes correspond to those of a time homogeneous $G/G/1$ queue. However, since the queue length fluid limit in the $\Delta_{(i)}/GI/1$ queue can also vary with time, we also identify the following “finer” operating states; this is analogous to the $M_t/M_t/1$ queue, as demonstrated in [6]. In particular, these states are useful in studying the approximation to the distribution of the queue length process on local time scales. We also note that Louchard [11] identified some of these operating regimes in his analysis. The definitions below formalize the intuitive presentation in [11].

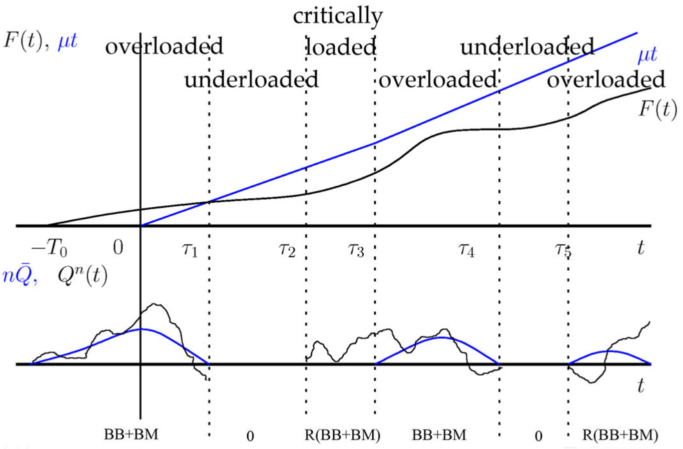

Fig. 3 An illustration of the various operating regimes of a transitory queueing model. Here, we consider the i.i.d. sampling $\Delta_{(i)}$ model

Definition 2 *(Operating states)* A transitory queue is at

(i) End of overloading at time t if $\rho(t) = 1$ and there exists an open interval (a, t) or (t, a) such that $\rho(r) > 1$ for all r in that interval.
(ii) Onset of critical loading at time t if $\rho(t) = 1$ and there exists a sequence $\lambda_n \uparrow t$ such that $\rho(\lambda_n) < 1$ for all n.
(iii) End of critical loading at time t if $\rho(t) = 1$, and there exists a sequence $\lambda_n \uparrow t$ such that $\rho(\lambda_n) = 1$ for all n and a sequence $\gamma_n \downarrow t$ such that $\rho(\gamma_n) < 1$ for all n.
(iv) Middle of critical loading at time t if $\rho(t) = 1$, and t is in an open interval (a, b), such that $\sup_{t\in(a,b)} \rho(t) \geq 1$ and there exists a sequence $\lambda_n \uparrow t$ such that $\rho(\lambda_n) = 1$ for all n.

We illustrate how the limit process can be used to approximate the queue length distribution of the *exact* (pre-limit) model. Our goal is to study this *distributional approximation* as $\bar{Q}$ and $\hat{Q}$ vary through the various operating regimes and states as defined above.

Theorem 3 (Distributional approximations) *The queue length can be approximated in the various operating regimes as follows.*

(i) *Overloaded state. Let $t \in (t^*, \tau)$ be a time instant of overloading in the overloaded interval, where $t^* := \sup \nabla_t^{\bar{X}}$ and $\tau := \inf\{s > t^* | \rho(s) = 1\}$. Then*

$$\frac{Q^n(t)}{\sqrt{n}} - \sqrt{n}(F(t) - F(t^*) - \mu(t - t^*)) \Rightarrow \hat{X}(t) + X^*, \text{ as } n \to \infty$$

where $X^* := \sup_{s\in\nabla_{t^*}^{\bar{X}}}(-\hat{X}(s))$. Further, $\tilde{Z}_t^n := \sqrt{n}(F(t) - F(t^*) - \mu(t - t^*)) + \hat{X}(t) + X^*$ is the strong solution to the stochastic differential equation

$d\tilde{Z}^n_t = \sqrt{n}(f(t) - \mu)dt + \sqrt{g^{'}(t)}dW_t \quad \forall t \in (t^*, \tau)$ with initial condition $\tilde{Z}_{t^*} = \hat{X}(t^*) - X^*$, where $g(t) = F(t)(1 - F(t)) + \sigma^2\mu^3\bar{B}(t)$.

(ii) *Underloaded state. If t is a point of underloading, i.e., if $\rho(t) < 1$, then $\frac{Q^n(t)}{\sqrt{n}} \Rightarrow 0$, as $n \to \infty$.*

(iii) *Middle and End of critically loaded state. An open set of the domain (t^*, τ) is a critically loaded interval, where t^* is a point in the onset of critically loaded state and τ a point at the end of critically loaded state, as defined in Definition 2. For any $t \in (t^*, \tau)$, let $u = t - t^*$ and we have, as $n \to \infty$,*

$$\frac{Q^n(t)}{\sqrt{n}} \Rightarrow (\hat{X}(t) + \sup_{0\le s\le u} (-\hat{X}(s))),$$

where $\hat{X}(u) \overset{d}{=} \hat{X}(t) - \hat{X}(t^*)$, and $\hat{X}(t) \overset{d}{=} \int_{-T_0}^{t} \sqrt{g^{'}(s)}dW_s$.

(iv) *End of overloading state. Let t be a point of end of overloading. Then, for all $\tau > 0$*

$$\frac{Q^n(t - \frac{\tau}{\sqrt{n}})}{\sqrt{n}} \Rightarrow \left(\hat{X}(t) + \left(\sup_{s\in\nabla_t^{\bar{X}}\setminus\{t\}} (-\hat{X}(s))\right) - (f(t) - \mu)\tau\right)_+, \text{ as } n \to \infty,$$

where $f(t)$ is the density function associated with the fluid limit F.

The proof is relegated to the appendix.

Remarks (1) *Overloaded regime* (i) In this case the approximate distribution is Gaussian with mean $F(t) - \mu t$. However, the variance is affected by the fact that the queue may have idled in the past. Recall that the variance is $g(t) = F(t)(1 - F(t)) + \sigma^2\mu^3\bar{B}(t)$, where from Corollary 1

$$\bar{B}(t) = \mathbf{1}_{\{t\ge 0\}} - \frac{1}{\mu}\Psi(\bar{X})(t).$$

(ii) We note that this result is analogous to case 5 of Sect. 4 in [11]. However, in [11], the author notes that no reflection needs be applied in an overloaded sub-interval, and proceeds to derive the limit process (in this interval alone) as $W^0 \circ F(t) - \sigma\mu^{3/2}W(t)$. This is not entirely accurate as the starting state of the process in each new interval of overloading must be factored into the approximation. That is, while $\nabla_t^{\bar{X}}$ is fixed for all t in an overloaded sub-interval, the value $\sup_{s\in\nabla_t^{\bar{X}}}(-\hat{X}(s))$ provides the starting state for the diffusion in such an interval.

(2) *Critically loaded regime* The queue length process in the critically loaded regime is approximated by a driftless reflected process, with continuous sample paths, with starting state $\hat{X}(t^*)$. By the definition of a critically loaded state $\rho(t) = 1$ at all such points and $\nabla_t^{\bar{X}}$ "accumulates" the points of critical loading, as t evolves through the critically loaded interval. It follows that the set $\nabla_t^{\bar{X}}$ is the interval $(t^*, t]$.

(3) *End of overloading regime* As noted in the definition, a point t is one of end of overloading if the traffic intensity is 1 at t, and is strictly greater than 1 at all points to the left of it. Here, we are primarily interested in the rate at which the queue empties out asymptotically as overloading ends. Consider a sequence of τ_n defined as a sequence of times at which the queue in the nth system first empties out, and define $v := t - \frac{\tau_n}{\sqrt{n}}$. Then, from Theorem 3

$$\tau_n = \sqrt{n}(t - v) \Rightarrow \frac{\hat{X}(t) + \sup_{s \in \nabla_t^{\bar{X}} \setminus \{t\}}(-\hat{X}(s))}{f(t) - \mu}$$

Thus, it can be seen that the time at which the queue empties out converges to a Gaussian random variable. A similar conclusion was drawn in [11] and in [6] for the $M_t/M_t/1$ queue.

6.2 Sample paths

We now characterize a typical sample path of the limit process $\hat{Q}$.

Proposition 5 *The process $\hat{Q}$ is upper-semicontinuous almost surely.*

The following proposition summarizes where discontinuities occur in $\hat{Q}$. We note that this is also part of Theorem 3.1 of [6]. Since the proof follows that in [6], we omit it.

Proposition 6 *$\hat{Q}$ is discontinuous at time t, with a non-zero probability, if and only if t is the end-point of overloading or critical loading. The set of such points is nowhere dense.*

Remarks (1) We note that the queue length limit sample paths for the $M_t/M_t/1$ model are also upper-semicontinuous as shown in Theorem 3.1 of [6]. There the sequence of converging processes was shown to be monotone, which easily leads to upper-semicontinuity by Dini's Theorem. As this monotonicity property does not hold for the corresponding processes in the $\Delta_{(i)}/GI/1$ model, we argue that the sample path is upper-semicontinuous directly from the characterization of the points of continuity and discontinuity in the domain of the sample path.

(2) The intuition for the regime switching behavior proved in is easy to see in the case of a uniform arrival distribution with early-bird arrivals, such that the service rate is greater than the value of the density function. Here, the (fluid) queue is overloaded on the interval $[-T_0, \tau)$ with the singleton set $\nabla_t^{\bar{X}} = \{-T_0\}$, and underloaded on the interval (τ, ∞) with the singleton set $\nabla_t^{\bar{X}} = \{t\}$. At τ itself, there are two points in the set $\nabla_t^{\bar{X}} = \{-T_0, \tau\}$. Thus, there is a discontinuity due to the fact that the set $\nabla_t^{\bar{X}}$ changes from being a singleton on the interval $[-T_0, \tau)$ to $\{-T_0, \tau\}$ at τ.

7 Examples and simulations

We illustrate the queue length process approximations with uniform and exponential arrival time distributions. The former is interesting, as the uniform distribution emerges as the mean field equilibrium arrival profile when arriving users are strategic about when they enter the queue in order to minimize their delay through the queue; see [22,23]. The exponential distribution case serves to illustrate the fact that many of the conclusions of our theorems can be carried over to infinite support arrival time distributions, though the limit results remain to be fully justified.

7.1 Uniform arrival distribution

The uniform arrival case is particularly simple and illustrates the discontinuities in the limit processes. Recall that $\nabla_t^{\bar{X}}$ is a correspondence that maps each time t to the set of points (up to t) at which the fluid netput process is equal to its infimum at t.

Corollary 3 *Let F be the uniform distribution on $[-T_0, T]$, where $-T_0 < 0$. Then,*

$$\hat{Q}(t) = \begin{cases} W^0(F(t)) - \sigma\mu^{\frac{3}{2}} W(t) & \forall t \in [-T_0, \tau) \\ (W^0(F(\tau)) - \sigma\mu^{\frac{3}{2}} W(\tau)) + (-(W^0(F(\tau)) - \sigma\mu^{\frac{3}{2}} W(\tau)))_+ & t = \tau \\ 0, & \forall t \in (\tau, \infty), \end{cases}$$

where $\tau = \{-T_0 \leq t < \infty \mid F(t) = \mu t\}$.

Proof Recall from Theorem 2 that $\hat{Q} = \hat{X} + \sup_{s \in \nabla_{\cdot}^{\bar{X}}}(-\hat{X})$ where $\hat{X} = W^0 \circ F - \sigma\mu^{\frac{3}{2}} W \circ \bar{B}$, and $\bar{B}$ is the fluid busy time process. Now, using the definition of $\nabla_t^{\bar{X}}$, it is easy to deduce that in this case, we have

$$\nabla_t^{\bar{X}} = \begin{cases} \{-T_0\} & \forall t \in [-T_0, \tau), \\ \{-T_0, \tau\} & t = \tau, \\ \{t\} & \forall t \in (\tau, \infty). \end{cases}$$

Further, Corollary 1 yields

$$\bar{B}(t) = \begin{cases} t & \forall t \in [-T_0, \tau], \\ 0 & \forall t \in (\tau, \infty). \end{cases}$$

Using these facts, the conclusion follows by substitution. □

The time τ can be interpreted as the first time that the fluid service process catches up with the fluid arrival process. For a uniform F, there is at most one such point, but in general, there can be many such points.

Remarks (1) A useful way to interpret the discontinuity at τ in Corollary 3 is to consider the process on the two sub-intervals separately and try to "patch" them together. If $\hat{Q}(\tau-) = \hat{X}(\tau) = \hat{Q}(\tau) > 0$, we should expect a free diffusion path

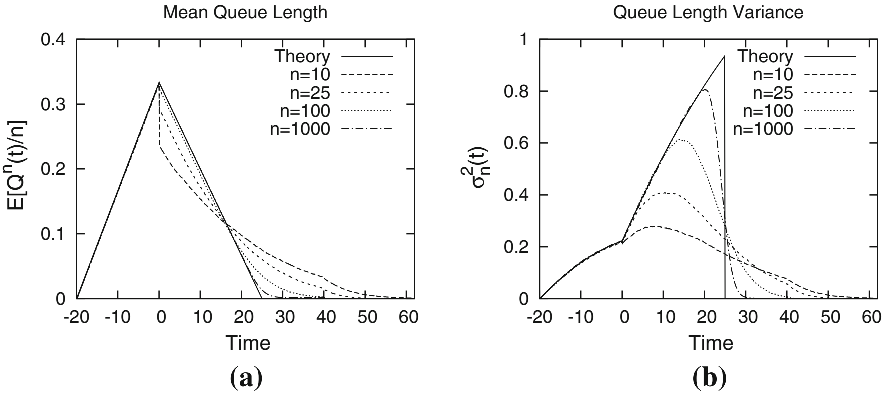

Fig. 4 Typical sample paths, mean and variance envelopes of the queue length process for F uniform over $[-20, 40]$, and exponentially distributed service times with rate $\mu = 0.03$. **a** Sample queue length process mean for $n = 10, 25, 100, 1000$, averaged over 10,000 simulation runs. **b** Sample queue length process variance for $n = 10, 25, 100, 1000$, averaged over 10,000 simulation runs

on the interval $[-T_0, \tau]$, and a reflected process such that the path is 0 on (τ, ∞). Furthermore, $\hat{Q}(\tau)$ becomes the "starting state" for the process on the interval (τ, ∞), and the reflection operator is applied an instant after τ. On the other hand, if $\hat{Q}(\tau-) = \hat{X}(\tau-) \leq 0$, we have a free diffusion on $[-T_0, \tau)$ and the zero process on $[\tau, \infty)$, i.e., the process drops to zero at τ. Thus, $\hat{Q}(\tau-)$ provides the starting conditions for the new "regime" of the diffusion, as the process transitions from $[-T_0, \tau)$ to (τ, ∞).

(2) We note that in [11], a diffusion approximation to the queue length process is derived independently for different operating regimes, and as such does not involve the directional derivative reflection map. These limit results have not been "patched" together to obtain a "process-level" convergence result, which is precisely where the mathematical challenges lie.

Note that the nature of the discontinuity at $\hat{Q}(\tau)$ depends on the the sign of $\hat{X}(\tau)$. Following [6], it can be shown t is a point of *right-discontinuity* for a function $x \in \mathcal{D}_{\lim}$ if x is left-continuous at t, and $x(t-) > x(t+)$. On the other hand, t is a point of *left-discontinuity* if x is right-continuous at t, and $x(t+) > x(t-)$.

Corollary 4 *Let F be the uniform distribution over $[-T_0, T]$, where $T_0 > 0$, and $\tau = \{-T_0 \leq t < \infty | F(t) = \mu t \mathbf{1}_{\{t \geq 0\}}\}$. Then, for the process $\hat{Q}$ in Corollary 3, we have*

(i) $[-T_0, \tau) \cup (\tau, \infty)$ *are points of continuity.*
(ii) τ *is a point of right-discontinuity, when* $\hat{X}(\tau) \geq 0$.
(iii) τ *is point of left-discontinuity, when* $\hat{X}(\tau) < 0$.

The proof is available in the Appendix.

Simulations can provide insight into the accuracy of the approximations for various population sizes. Consider a uniform arrival distribution over the interval $[-20, 40]$, with service times i.i.d. and exponentially distributed with parameter $\mu = 0.03$. Figures 4a, b show the sample mean and the sample variance of the (scaled) queue length process for $n = 10, 25, 100, 1000$ over 10,000 sample runs. Note that as n increases, the sample mean approaches the fluid limit, and the sample variance approaches the theoretical variance of the queue length process. For the given F, the latter quantity is

$$\sigma^2(t) = \begin{cases} F(t)(1 - F(t)) & \forall t \in [-T_0, 0] \\ F(t)(1 - F(t)) + \sigma^2\mu^3 t & \forall t \in (0, \tau) \\ 0 & \forall t > \tau. \end{cases}$$

Observe from Fig. 4a that even for small n, the sample mean is quite close to the fluid limit for $t < 0$. However, once queueing dynamics come into play, the fluid limit is a good approximation only for $n = 100$ or larger. A similar effect is manifest for the diffusion limit as well: once service starts, and queueing dynamics come into play, the diffusion limit becomes a reasonably good approximation only for $n = 1000$ or larger.

7.2 Exponential arrival distribution

Assume F is an exponential distribution function with parameter $\lambda > 0$, so that $F(t) = 1 - e^{-\lambda t}$ and $-T_0 = 0$. Keep in mind that this is unlike the $M/GI/1$ queue where the exponential distribution models the inter-arrival times. Recall that the limit results in Theorems 1 and 2 are proved on compact sets of the domain $[-T_0, \infty)$. Therefore, the limit does not hold simultaneously at all points in the support of F, and proving the FSLLN and FCLT for infinite support distributions is beyond the scope of the current paper. However, observe that the queue length fluid model can be conjectured to be

(i) If $\mu \geq \lambda$, then $\bar{Q}(t) = 0 \; \forall t \in [0, \infty)$.
(ii) If $\mu < \lambda$, then

$$\bar{Q}(t) = \begin{cases} (1 - e^{-\lambda t} - \mu t) & \forall t \in [0, \tau) \\ 0 & \forall t \geq \tau, \end{cases}$$

where $\tau := \inf\{t \geq 0 | F(t) = \mu t\}$ is the last instant and the fluid queue length is positive (also known as the *makespan*). To see this, recall the definition of $\bar{Q}(t)$ and notice that if $\mu \geq \lambda$ then $\lambda e^{-\lambda t} \leq \mu, \; \forall t > 0$. This implies that the queue is underloaded, as defined in Sect. 6.1. On the other hand, if $\mu < \lambda$, the system shifts from overload to underload, per our definition in Sect. 6.1. It can be shown that $\tau = \frac{1}{\lambda}\mathscr{W}\left(-\frac{\lambda}{\mu}e^{-\frac{\lambda}{\mu}}\right) + \frac{1}{\mu}$, where $\mathscr{W}(\cdot)$ is the Lambert W function. To see this, recall that it is the first (strictly positive) solution to $e^{-\lambda t} = 1 - \mu t$. Substituting in $-x = -\lambda t + \frac{\lambda}{\mu}$, we have $xe^x = -\frac{\lambda}{\mu}e^{-\frac{\lambda}{\mu}}$. It is well known that this is the defining equation for the Lambert W function $\mathscr{W}$, implying that $x = \mathscr{W}\left(-\frac{\lambda}{\mu}e^{-\frac{\lambda}{\mu}}\right)$. Substituting back for t, we obtain the expression for τ.

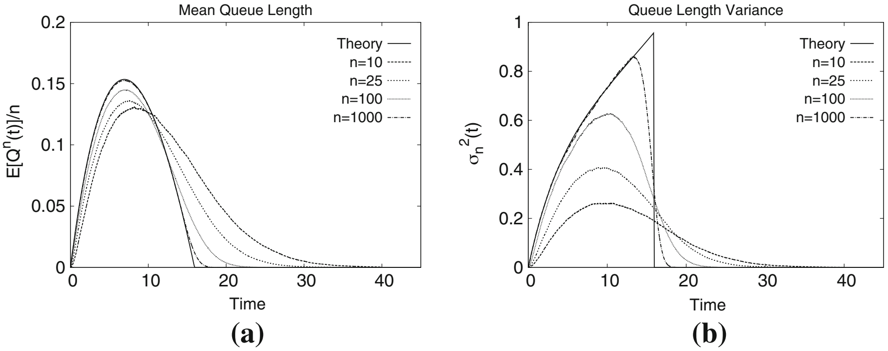

Fig. 5 Typical sample paths, mean and variance envelopes of the queue length process for F exponential with parameter $\lambda = 0.1$ and exponentially distributed service times with mean rate $\mu = 0.05$. **a** Sample mean queue length for $n = 10, 25, 100, 1000$, averaged over 30 simulation runs. **b** Sample variance for $n = 10, 25, 100, 1000$, averaged over 30 simulation runs

The fluid model allows us to conjecture the corresponding diffusion refinement. Let $\hat{Q}$ be the queue length diffusion model. Then,

(i) If $\mu \geq \lambda$, then $\hat{Q}(t) = 0 \; \forall t \in [0, \infty)$.
(ii) If $\mu < \lambda$, then

$$\hat{Q}(t) = \begin{cases} W^0(F(t)) - \sigma\mu^{\frac{3}{2}}W(t) & \forall t \in [0,\tau) \\ (W^0(F(t)) - \sigma\mu^{\frac{3}{2}}W(t)) + (-W^0(F(t)) + \sigma\mu^{\frac{3}{2}}W(t))_+ & t = \tau \\ 0 & \forall t \in (\tau, \infty). \end{cases}$$

The "proof" of this is straightforward. Part (i) follows from the fact that the fluid model is underloaded under the same condition. Part (ii) follows from the reasoning in the proof of Corollary 3. A little algebra shows that the variance curve of the diffusion approximation $\hat{Q}$ when $\mu < \lambda$ is given by

$$\sigma^2(t) = \begin{cases} F(t)(1 - F(t)) + \sigma^2\mu^3 t & \forall t \in [0,\tau), \\ 0 & \forall t > \tau. \end{cases}$$

Let us consider a specific example, where $\lambda = 0.1$ and $\mu = 0.05$, in which case it can be verified that $\tau = 15.9362$. Figure 5a shows that for even low values of n, the fluid limit is a very good approximation to the observed mean queue length. Similarly, Figure 5b shows that the variance of the diffusion limit is a reasonable approximation to the variance of the queue length in the (accelerated) discrete event system.

We also note a very interesting connection between random graph theory and the $\Delta_{(i)}/GI/1$ queue, brought to our notice by J.S.H. van Leeuwaarden in a personal communication. Specifically, he has shown that the excursions of the queue length process in the discrete event system, observed at the departure times of jobs, also measure the size of the connected components of a random graph with n vertices. [24] shows that in the "large graph" limit (i.e., as $n \to \infty$), the connected components in a (nearly) critical Erdös-Rényi random graph (see [25] for details on these terms) can be related to the excursions of a Brownian motion on a parabola by a weak convergence limit result linking the two. This type of result is also intimately connected with the question of the final size of an *epidemic* in a critical random graph, see [26,27] where the distribution of the final size in a critical susceptible–infected–recovered (SIR) epidemic model is studied. Using a Taylor series expansion on the fluid limit of the queue length, it can be shown that for small t and ignoring terms of order 3 and higher, the diffusion approximation is a Brownian excursion on a parabola. This connection with the $\Delta_{(i)}/GI/1$ queue might provide a new framework to study the final size distribution of other epidemic models in the critical regime.

8 Conclusions and future work

In this paper, we introduce a bespoke single-server queueing model, which we call the $\Delta_{(i)}/GI/1$ queue, to model systems that are purely transient in nature, and thus serve a finite population of customers. We develop pathwise asymptotic fluid and diffusion approximations to the system performance metrics as the population size is increased to infinity. These approximations are unlike the conventional heavy-traffic limits, but are closer in spirit to the uniform acceleration approximations to the $M_t/M_t/1$ queue.

Our original motivation for introducing the $\Delta_{(i)}/GI/1$ model came from the 'concert arrival game,' a game of arrival timing introduced in [22]. Customers choose to arrive at a queue to minimize a linear cost functional that depends on the waiting time and the number of people who have already arrived. In the fluid limit, the Nash equilibrium arrival profile was shown to be a uniform distribution function. An important question of interest is whether the equilibrium derived from the fluid model approximates in any way the equilibrium of the finite population 'concert arrival game.'

Our next step is to take the diffusion approximations for the $\Delta_{(i)}/GI/1$ queue model, and revisit the 'concert arrival game' problem. In [22], the assumption is that the queue lengths are unobservable. Our diffusion approximations can now allow us to study other situations where the queue length is fully or partially observable. In the spirit of mean field game theory, this could be understood to be a 'diffusion field game theory.'

An important direction to take this research would be to study transitory queueing models with non-stationary service processes. For instance, customers arriving closer to the end of day may experience shorter service times. We conjecture that the limit results will be interesting but non-trivial to establish.

Finally, it would be interesting to test empirically for how to fit the distribution F, that characterizes the arrival pattern, to data. Then, it would be possible to use the wait time predictions suggested by the $\Delta_{(i)}/GI/1$ model to make capacity sizing recommendations. This would also allow us to compare the performance of the $\Delta_{(i)}/GI/1$ to the more common $GI/GI/1$ model in various application contexts.

Appendix

Proof of Lemma 2

Rewrite $\tilde{y}_n$ as $\tilde{y}_n = (\Psi(\sqrt{n}x + y_n) - \Psi(\sqrt{n}x + y)) - (\Psi(\sqrt{n}x + y) - \sqrt{n}\Psi(x))$. Now, using the fact that the Skorokhod reflection map is Lipschitz continuous under the uniform metric (see Lemma 13.4.1 and Theorem 13.4.1 of [12]), we have $(\Psi(\sqrt{n}x + y_n) - \Psi(\sqrt{n}x + y)) \leq \|y_n - y\|$, where $\|\cdot\|$ is the uniform metric. It follows that $\tilde{y}_n \leq \|y_n - y\| + (\Psi(\sqrt{n}x + y) - \sqrt{n}\Psi(x))$. Now, by Theorem 9.5.1 of [19], we know that as $n \to \infty$

$$(\Psi(\sqrt{n}x + y) - \sqrt{n}\Psi(x)) \overset{a.s.}{\to} \tilde{y}, \text{ in } (\mathcal{D}_{\lim}, M_1).$$

Using this result, and the fact that by hypothesis y_n converges to y in $(\mathcal{D}_{lim}, J_1)$, we have $\tilde{y}_n \overset{a.s.}{\to} \tilde{y}$, in $(\mathcal{D}_{\lim}, M_1)$. □

Proof of Lemma 3

First, suppose $\bar{Q}(t) > 0$. It follows that $\bar{F}(t) - \mu t > \inf_{-T_0 \leq s \leq t}(\bar{F}(s) - \mu s) = w$, where the latter equality follows because the queue starts empty at time 0, and the fluid netput is positive before time 0 (Note that we ignore the positive part operator in the definition of Ψ, as the systems starts empty at time $-T_0$). Now, let $t^* = \sup\{0 \leq s \leq t | (\bar{F}(s) - \mu s) = \inf_{0 \leq s \leq t}(\bar{F}(s) - \mu s)\}$ be the point at which the infimum is achieved, on the right- hand side. It follows that $\bar{F}(t) - \mu t > \bar{F}(t^*) - \mu t^*$, in turn yielding

$$\rho(t) = \sup_{0 \leq s \leq t} \frac{\bar{F}(t) - \bar{F}(s)}{\mu(t-s)} > 1.$$

Next, suppose $\bar{Q}(t) = 0$, $\bar{X}(t) = \Psi(\bar{X})(t)$ and there exists an $r < t$ such that $\Psi(\bar{X})(t) = \Psi(\bar{X})(s)$ for all $s \in [r, t]$. It follows that $\bar{F}(t) - \mu t = -\sup_{-T_0 \leq s \leq t}(-(\bar{F}(s) - \mu s))$, implying there exists a point $r^* \in [0, t]$ such that $\bar{F}(t) - \mu t = \bar{F}(r^*) - \mu r^*$. This, in turn, implies that

$$\sup_{0 \leq s \leq t} \frac{\bar{F}(t) - \bar{F}(s)}{\mu(t-r)} \geq \frac{\bar{F}(t) - \bar{F}(r^*)}{\mu(t-r^*)} = 1.$$

However, a simple contradiction argument shows that

$$\sup_{0 \leq s \leq t} \frac{\bar{F}(t) - \bar{F}(s)}{\mu(t-r)} > 1$$

is impossible, implying that

$$\sup_{0 \leq s \leq t} \frac{\bar{F}(t) - \bar{F}(s)}{\mu(t-r)} = 1.$$

Finally, consider case (iii). We have $\forall r < t$,

$$-(\bar{F}(t) - \mu t) = \sup_{-T_0 \le s \le t} (-(\bar{F}(s) - \mu s)) > \sup_{-T_0 \le s \le r} (-(\bar{F}(s) - \mu s)).$$

It follows that $-(\bar{F}(t) - \mu t) > -(\bar{F}(r) - \mu r)$,
implying

$$1 > \frac{\bar{F}(t) - \bar{F}(r)}{\mu(t - r)} \ \forall r \in [0, t).$$

□

Proof of Theorem 3

(i) Overloaded regime

Proof First, note that τ is the first instant of an end of overloading phase, and the current overloaded phase ends at τ. In the overloaded state $\bar{Q}(t) > 0$, implying that $\Psi(\bar{X})(t)$ is a constant. Using the definition of $\nabla_t^{\bar{X}}$, it follows that $\Psi(\bar{X})(t) = -\bar{X}(t^*)$, and $\bar{Q}(t) = \bar{X}(t) - \bar{X}(t^*) = (\bar{F}(t) - \bar{F}(t^*) - \mu(t - t^*))$. Next, from Theorem 2, it is obvious that $\frac{Q^n(t)}{\sqrt{n}} \overset{d}{\approx} \tilde{Z}_t^n$.

Next, from Remark 1 after Lemma 1, $\hat{X}(t) - \hat{X}(t^*) = \int_{t^*}^{t} \sqrt{g^{'}(s)}\, \mathrm{d}W_s$, which can be seen to be a diffusion process that starts from 0 at t^*. Noting that $\nabla_t^{\bar{X}}$ does not change on the interval (t^*, τ), it follows that $X^* = \sup_{s \in \nabla_{t^*}^{\bar{X}}} \{-\hat{X}(s))$ is a fixed random variable, and $\tilde{Z}_t^n$ has an initial condition $\tilde{Z}_{t^*}^n = \hat{X}(t^*) - X^*$. It is straightforward to see that $\tilde{Z}_n^{\cdot}$ is the strong solution to the mentioned SDE, since it is adapted to the filtration generated by W. □

(ii) Underloaded regime
This result is immediate from the definition of the limit processes.

(iii) Middle and end of critically loaded state

Proof For any $t \in (t^*, \tau)$ we have $\bar{Q}(t) = 0$. From the weak convergence result in Theorem 2, we have $Q^n(t) \overset{d}{\approx} n\bar{Q}(t) + \sqrt{n}\hat{Q}(t)$, and expanding the definition of $\hat{Q}$, it follows that $Q^n(t) \overset{d}{\approx} \sqrt{n}(\hat{X}(s) + \sup_{s \in \nabla_t^{\bar{X}}}(-\hat{X}(s)))$. Using the fact that $\Psi(\bar{X})(t) = w = -\bar{X}(t)\ \forall\ t \in (t^*, \tau)$ in a critically loaded regime, it follows that $\nabla_t^{\bar{X}} = (t^*, t]$ for $t \in (t^*, \tau)$. Thus, we have $Q^n(t) \overset{d}{\approx} \sqrt{n}(\hat{X}(s) + \sup_{t^* < s \le t}(-\hat{X}(s)))$. Let $u = t - t^*$. Then, after a change of variables, we obtain $Q^n(u + t^*) \overset{d}{\approx} \sqrt{n}(\hat{X}(u + t^*) + \sup_{0 \le s < u}(-\hat{X}(s)))$.

Since W^0 is a Brownian Bridge process, the strong Markov property of Brownian motion ([18]) implies that $\hat{X}(u + t^*) - \hat{X}(t^*) = \hat{X}(u)$. Substituting this into the expression above we see that we have $Q^n(u + t^*) = Q^n(u) + \hat{X}(t^*)$, where $\hat{X}(t^*)$ is the starting state of the process in the middle-of-critically loaded state. A simple change of variables will provide the desired result. A similar argument will hold for the end of critical loading state as well. □

(iv) End of overloading state

Proof By definition for any $\tau > 0$, $t - \frac{\tau}{\sqrt{n}}$ is a point of overloading. Therefore,

$$\begin{aligned}\frac{Q^n\left(t - \frac{\tau}{\sqrt{n}}\right)}{\sqrt{n}} &= \hat{X}^n\Big(t - \frac{\tau}{\sqrt{n}}\Big) + \sqrt{n}\Big(F(t) - \frac{\tau}{\sqrt{n}}\Big) - \mu\Big(t - \frac{\tau}{\sqrt{n}}\Big) \\ &\quad + \Psi(\hat{X}^n + \sqrt{n}\bar{X})\Big(t - \frac{\tau}{\sqrt{n}}\Big) - \sqrt{n}\Psi(\bar{X})\Big(t - \frac{\tau}{\sqrt{n}}\Big).\end{aligned}$$

Without loss of generality, we assume that service started when the queue was in the overloaded state, so that $\Psi(\bar{X})\left(t - \frac{\tau}{\sqrt{n}}\right) = 0$. Now, using the fact the derivative f exists, the mean value theorem implies the existence of a point $\tilde{t} \in \left[t - \frac{\tau}{\sqrt{n}}, t\right]$ such that $F\left(t - \frac{\tau}{\sqrt{n}}\right) = F(t) - f(\tilde{t})\frac{\tau}{\sqrt{n}}$. Adding and subtracting the term $f(t)\tau/\sqrt{n}$ to the expression above, we have

$$F\Big(t - \frac{\tau}{\sqrt{n}}\Big) = F(t) - f(t)\frac{\tau}{\sqrt{n}} + f(t)\frac{\tau}{\sqrt{n}} - f(\tilde{t})\frac{\tau}{\sqrt{n}}.$$

Substituting this into the expression for Q^n above, and introducing the term $\hat{X}^n(t)$, we obtain

$$\begin{aligned}\frac{Q^n\Big(t - \frac{\tau}{\sqrt{n}}\Big)}{\sqrt{n}} &= \hat{X}^n\Big(t - \frac{\tau}{\sqrt{n}}\Big) - \hat{X}^n(t) + \hat{X}^n(t) + \sqrt{n}(F(t) - \mu t) - (f(t) - \mu)\tau \\ &\quad + \Psi(\hat{X}^n + \sqrt{n}\bar{X})\Big(t - \frac{\tau}{\sqrt{n}}\Big) + (f(t) - f(\tilde{t}))\frac{\tau}{\sqrt{n}}.\end{aligned}$$

Now, using Lemma 1 and the continuity of the limit process we see that $\hat{X}^n(t - \frac{\tau}{\sqrt{n}}) - \hat{X}^n(t) \Rightarrow 0$. Further, since f is bounded by virtue of being defined on a finite interval, we have $\tau(f(t) - f(\tilde{t}))/\sqrt{n} \to \infty$ as $n \to \infty$. Next, consider the term $\hat{Z}(t) := \hat{X}^n(t) + \sqrt{n}(F(t) - \mu t) + \Psi(\hat{X}^n + \sqrt{n}\bar{X})\left(t - \frac{t}{\sqrt{n}}\right)$.

Let $\delta > 0$ be sufficiently small, so that the following decomposition of the expression above holds:

$$\begin{aligned}\hat{Z}^n(t) = &\sup_{-T_0 \le s < t-\delta} (\hat{X}^n(t) + \sqrt{n}(F(t) - \mu t) - \hat{X}^n(s) - \sqrt{n}\bar{X}(s)) \\ &\vee \sup_{t-\delta \le s \le t - \frac{\tau}{\sqrt{n}}} (\hat{X}^n(t) + \sqrt{n}(F(t) - \mu t) - \hat{X}^n(s) - \sqrt{n}\bar{X}(s)).\end{aligned}$$

Let $t^* = \sup\{\nabla_t^{\bar{X}} \backslash \{t\}\}$. Consider the first term on the RHS above, and call it $\hat{Z}_1^n(t)$. Since the queue is overloaded before t no points are "added" to the correspondence

$\nabla_t^{\bar{X}}$; it follows from the definition of an end of overloading point that $(F(t)-\mu t) = -\Psi(\bar{X})(t) \equiv -\Psi(\bar{X})(t^*+\delta)$. This, in turn, provides $\hat{Z}_1^n(t) = \hat{X}^n(t) + \Psi(\hat{X} + \sqrt{n}\bar{X})(t^*+\delta) - \sqrt{n}\Psi(\bar{X})(t^*+\delta)$. Using Lemma 2, it follows that $\hat{Z}_1^n(t) \Rightarrow \hat{X}(t) + \sup_{s\in\nabla_t^{\bar{X}}\setminus\{t\}}(-\hat{X}(s))$ as $n\to\infty$, followed by letting $\delta\to 0$. Next, consider the second term

$$\begin{aligned}\hat{Z}_2^n(t) &= \sup_{t-\delta\le s\le t-\frac{\tau}{\sqrt{n}}} (\hat{X}^n(t) + \sqrt{n}(F(t)-\mu t) - \hat{X}^n(s) - \sqrt{n}\bar{X}(s))\\ &\le \sup_{t-\delta\le s\le t-\frac{\tau}{\sqrt{n}}} (\hat{X}^n(t) - \hat{X}^n(s)) + \sup_{t-\delta\le s\le t-\frac{\tau}{\sqrt{n}}} \sqrt{n}(\bar{X}(t)-\bar{X}(s))\\ &\le \sup_{t-\delta\le s\le t} (\hat{X}^n(t) - \hat{X}^n(s)) + \sup_{t-\delta\le s\le t-\frac{\tau}{\sqrt{n}}} \sqrt{n}(\bar{X}(t)-\bar{X}(s)).\end{aligned}$$

For large n, as the queue is overloaded at $t-\frac{\tau}{\sqrt{n}}$, it follows that

$$\hat{Z}_2^n(t) \le \sup_{t-\delta\le s\le t}(\hat{X}(t)-\hat{X}(s)) + \sqrt{n}\Big(\bar{X}(t) - \bar{X}\Big(t-\frac{\tau}{\sqrt{n}}\Big)\Big).$$

Again, by the mean value theorem,

$$\begin{aligned}\sqrt{n}\Big(\bar{X}(t) - \bar{X}\Big(t-\frac{\tau}{\sqrt{n}}\Big)\Big) &= \sqrt{n}\Big(F(t) - F\Big(t-\frac{\tau}{\sqrt{n}}\Big) - \mu\frac{\tau}{\sqrt{n}}\Big)\\ &= \sqrt{n}(f(t)-\mu)\frac{\tau}{\sqrt{n}} + (f(t)-f(\tilde{t}))\tau,\end{aligned}$$

where $\tilde{t}\in[t-\frac{\tau}{\sqrt{n}}, t]$. Since, $\tilde{t}\to t$ as $n\to\infty$, by the (right) continuity of f, it follows that $f(t)-f(\tilde{t})\to 0$ as $n\to\infty$. Then it follows by an application of Lemma 1 (and using Skorokhod's almost sure representation) that $\overline{\lim}_{n\to\infty}\hat{Z}_2^n(t) \le \hat{X}(t) + \sup_{t-\delta\le s\le t}(-\hat{X}(s)) + (f(t)-\mu)\tau$. On the other hand, for a lower bound, using the mean value theorem again, we have $\hat{Z}_2^n(t) \ge \hat{X}^n(t) - \hat{X}^n(t-\frac{\tau}{\sqrt{n}}) + (f(t)-\mu)\tau + (f(t)-f(\tilde{t}))\tau$. Once again, using the continuity of f, the almost sure representation theorem and Lemma 1, and noting the continuity of the limit process $\hat{X}$, we have

$$\underline{\lim}_{n\to\infty}\hat{Z}_2^n(t) \ge (f(t)-\mu)\tau \ a.s.$$

Now, using the limits derived for $\hat{Z}_1^n$ and $\hat{Z}_2^n$, it follows that

$$\begin{aligned}&\frac{Q^n(t-\frac{\tau}{\sqrt{n}})}{\sqrt{n}} \Longrightarrow -(f(t)-\mu)\tau + \sup_{s\in\nabla_t^{\bar{X}}\setminus\{t\}}(\hat{X}(t)-\hat{X}(s)) \vee (f(t)-\mu)\tau\\ &= \left(\hat{X}(t) + \sup_{s\in\nabla_t^{\bar{X}}\setminus\{t\}}(-\hat{X}(s)) - (f(t)-\mu)\tau\right)_+.\end{aligned}$$

□

Proof of Proposition 5

The proof is a consequence of the following lemma, which consolidates Lemmas 6.5, 6.6, and 6.7 in [6]. The lemma characterizes the points of discontinuity (and continuity) of the process $\tilde{Y}(t) = \sup_{s\in\nabla_t^{\bar{X}}}(-\hat{X}(s))$ in relation to the correspondence $\nabla_t^{\bar{X}}$. We do not prove these conditions, but direct the interested reader to [6].

Lemma 4 *A point $t \in [-T_0, \infty)$ is characterized as follows.*

(i) *Continuity Conditions. The following are equivalent:*
  (1) *t is a continuity point.*
  (2) $t \in \nabla_t^{\bar{X}} = \{t\}$, or $t \notin \nabla_t^{\bar{X}}$, or $t \in \nabla_t^{\bar{X}} \neq \{t\}$, *and t is not isolated in* $\nabla_t^{\bar{X}}$ *and* $\nabla_t^{\bar{X}} \subseteq \nabla_u^{\bar{X}}$ *for some* $u > t$.

(ii) *Right-discontinuity Conditions. The following are equivalent:*
  (1) *t is a point of right-discontinuity.*
  (2) *$t \in \nabla_t^{\bar{X}} \neq \{t\}$ and $\nabla_u^{\bar{X}} \subseteq (t, u]\ \forall\, u > r$.*
  (3) $\tilde{Y}(t) = \tilde{Y}(t-) > \tilde{Y}(t+) = -\hat{X}(t)$.

(iii) *Left-discontinuity Conditions. The following are equivalent:*
  (1) *t is a point of left-discontinuity.*
  (2) *$t \in \nabla_t^{\bar{X}} \neq \{t\}$ and t is isolated in $\nabla_t^{\bar{X}}$.*
  (3) $\tilde{Y}(t) = \tilde{Y}(t+) = -\hat{X}(t) > \tilde{Y}(t-)$.

A point of right-discontinuity can be seen to be left-continuous, coupled with an ordering on the right and left limits, such that $\tilde{Y}(t-) > \tilde{Y}(t+)$. Similarly, a point of left-discontinuity is right-continuous, and the limits are ordered such that $\tilde{Y}(t+) > \tilde{Y}(t-)$. Using these definitions, we proceed to prove the upper-semicontinuity of the limit process.

Proof (Proposition 5) By definition, $\hat{X}$ is continuous, and it suffices to check that a sample path of the component $\tilde{Y}(t) = \sup_{s\in\nabla_t^{\bar{X}}}(-\hat{X}(s))$ is upper-semicontinuous. To see this, consider the pullback of the level set $\tilde{Y}^{-1}[a, \infty) = \{t \in [-T_0, \infty)|\tilde{Y}(t) \geq a\}$. It suffices to check that this is a closed set; see [28]. Let $\{\tau_n\} \subseteq \{t \in [-T_0, \infty)|\tilde{Y}(t) \geq a\}$ be a sequence of points such that $\tau_n \to \tau$ as $n \to \infty$, where $\tau \in [-T_0, \infty)$ is an arbitrary point in the domain of $\tilde{Y}$. Thus, if $\epsilon > 0$, then there exists an $n_0 \in \mathbb{N}$ such that $\forall\, n \geq n_0, \epsilon \geq \tau - \tau_n \geq -\epsilon$. If τ is a continuity point, then the conclusion is obvious. On the other hand, suppose that τ is a left-discontinuity point. By part (iii) of Lemma 4, it follows that $\tilde{Y}(\tau-) < \tilde{Y}(\tau+) = \tilde{Y}(\tau)$. By the definition of a left-discontinuity, there exits an interval $[t^*, \tau)$, where $t^* = \sup \nabla_\tau^{\bar{X}} \backslash \{\tau\}$, on which $\tilde{Y}$ is (locally) continuous. Fix $\delta > 0$, then there exists an $\eta > 0$ such that if $\geq -\eta$, then $\delta \geq \tilde{Y}(\tau-) - \tilde{Y}(t) \geq -\delta$. If ϵ is small enough, then there exists n_0 such that $\forall\, n \geq n_0, \tau - \tau_n > -\eta$. It follows that $\delta \geq \tilde{Y}(\tau_n) - \tilde{Y}(\tau-) \geq a - \tilde{Y}(\tau-)$, implying that $\tilde{Y}(\tau-) \geq a - \delta$. Since δ is arbitrary, it follows that $\tilde{Y}(\tau-) \geq a$, in turn implying that $\tilde{Y}(\tau) \geq 0$. Thus, $\tau \in \tilde{Y}^{-1}[a, \infty)$. Next, suppose that τ is a right-discontinuity point. Then, from part (ii) of Lemma 4, we have $\tilde{Y}(\tau) = \tilde{Y}(\tau-) < \tilde{Y}(\tau+)$. Furthermore, for any $u > \tau$, we have $\nabla_u^{\bar{X}} \subseteq (\tau, u]$ implying that these are continuity points (by part (i) of Lemma 4). Using an argument similar to that for a left-discontinuity, on points to the right of τ, it follows that $\tilde{Y}(\tau) \geq a$. This implies that the pullback set $\tilde{Y}^{-1}[a, \infty)$ is closed. As $\{\tau_n\}$ is an arbitrary sequence in $\tilde{Y}^{-1}[a, \infty)$ it is necessarily true that $\tilde{Y}$ is upper-semicontinuous. □

Proof of Corollary 4

The proof of the corollary depends on Lemma 4 above.

Proof (Corollary 4) Recall that $\hat{Q} = \hat{X} + \tilde{Y}$, where $\tilde{Y}(t) = \sup_{s \in \nabla_t^{\bar{X}}}(-\hat{X}(s))$. The proof of (i) follows directly from part (i) of Lemma 4. Next, recall from the proof of Corollary 3 that $\nabla_\tau^{\bar{X}} = \{-T_0, \tau\}$. Thus, τ is isolated in the set and it follows that part (iii) of Lemma 4 is satisfied. On the other hand, recall that $\nabla_t^{\bar{X}} = \{t\} \subset (\tau, t]$, $\forall t > \tau$, and τ can also be a point of right-discontinuity, by part (ii) of Lemma 4. Thus, τ is one or the other depending on the path of $\hat{X}$. If $\hat{X}(\tau) < 0$ then $\tilde{Y}(\tau+) = \tilde{Y}(\tau) > \tilde{Y}(\tau-)$ and τ is a point of left-discontinuity. Otherwise, if $\hat{X}(\tau) \geq 0$, then s$\tilde{Y}(\tau) = \tilde{Y}(\tau-) = 0 > \tilde{Y}(\tau+)$ and τ is a point of right-discontinuity. □

References

1. Newell, G.F.: Queues with time-dependent arrival rates I, II and III. J. Appl. Probab., 5:436–451 (I); 579–590 (II); 591–606 (III) (1968)
2. Massey, W.A.: Non-stationary queues. PhD thesis. Stanford University (1981)
3. Keller, J.B.: Time-dependent queues. SIAM Rev. **24**, 401–412 (1982)
4. Massey, W.A.: Asymptotic analysis of the time dependent M/M/1 queue. Math. Oper. Res. **10**, 305–327 (1985)
5. Hall, R.W.: Queueing Methods: For Services and Manufacturing. Prentice Hall, Englewood Cliffs (1990)
6. Mandelbaum, A., Massey, W.A.: Strong approximations for time-dependent queues. Math. Oper. Res. **20**(1), 33–64 (1995)
7. Liu, Y., Whitt, W.: A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading. Oper. Res. Lett. **40**(5), 307–312 (2012)
8. Liu, Y., Whitt, W.: The $G_t/GI/s_t + GI$ many-server fluid queue. Queueing Syst. **71**(4), 405–444 (2012)
9. Newell, G.F.: Applications of Queueing Theory, 2nd edn. Chapman and Hall Ltd., New York (1982)
10. Gaver, D.P., Lehorsky, J.P., Perlas, M.: Service systems with transitory demand. In: Logistics, vol. 1 (1975)
11. Louchard, G.: Large finite population queueing systems. The single-server model. Stoch. Proc. Appl. **53**(1), 117–145 (1994)
12. Whitt, W.: Stochastic Process Limits. Springer, New York (2001)
13. Durrett, R.: Probability: Theory and Examples, 4th edn. Cambridge University Press, Cambridge (2010)
14. Chen, H., Yao, D.D.: Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization. Springer, New York (2001)
15. Billingsley, P.: Convergence of Probability Measures. Wiley, New York (1968)
16. Mandelbaum, A., Ramanan, K.: Directional derivatives of oblique reflection maps. Math. Oper. Res. **35**(3), 527 (2010)
17. Skorokhod, A.V.: Limit theorems for stochastic processes. Theory Probab. Appl., **1**(3), 261–290 (1956)
18. Karatzas, I., Shreve, S.E.: Brownian Motion and Stochastic Calculus. Springer, New York (1991)
19. Whitt, W.: Internet Supplement To Stochastic Process Limits. Springer, New York (2001)
20. Pomarede, J.L.: A Unified Approach via Graphs to Skorohod's Topologies on the Function Space D PhD thesis. Yale University, New Haven (1976)
21. Puhalskii, A.A., Reed, J.E.: On many-server queues in heavy traffic. Ann. Appl. Probab. **20**(1), 129–195 (2010)

22. Jain, R., Juneja, S., Shimkin, N.: The concert queueing game: to wait or to be late. Discr. Event Dyn. Syst. **21**(1), 103–134 (2011)
23. Honnappa, H. Jain, R.: Strategic arrivals into queueing networks: the network concert queueing game. Oper. Res. (2013)
24. Aldous, D.: Brownian excursions, critical random graphs and the multiplicative coalescent. Ann. Probab. **25**, 812–854 (1997)
25. Durrett, R.: Random Graph Dynamics. Cambridge University Press, Cambridge (2007)
26. Martin-Löf, Anders: The final size of a nearly critical epidemic, and the first passage time of a wiener process to a parabolic barrier. J. Appl. Probab. **35**(3), 671–682 (1998)
27. Van der Hofstad, R., Janssen, A.J.E.M., van Leeuwaarden, J.S.H.: Critical epidemics, random graphs, and brownian motion with a parabolic drift. Adv. Appl. Probab. **42**(4), 1187–1206 (2010)
28. Rudin, W.: Real and Complex Analysis. McGraw-Hill, New York (2006)