Queueing Syst (2013) 73:147–193
DOI 10.1007/s11134-012-9307-9

Fluid models of many-server queues with abandonment

Jiheng Zhang

Received: 20 December 2010 / Revised: 2 April 2012 / Published online: 17 May 2012

Abstract We study many-server queues with abandonment in which customers have general service and patience time distributions. The dynamics of the system are modeled using measure-valued processes, to keep track of the residual service and patience times of each customer. Deterministic fluid models are established to provide a first-order approximation for this model. The fluid model solution, which is proved to uniquely exist, serves as the fluid limit of the many-server queue, as the number of servers becomes large. Based on the fluid model solution, first-order approximations for various performance quantities are proposed.

Keywords Many-server queue · Abandonment · Measure valued process · Quality driven · Efficiency driven · Quality and efficiency driven

Mathematics Subject Classification Primary 60K25 · Secondary 68M20 · 90B22 · 68M07

1 Introduction

Recently, there has been a great interest in queues with a large number of servers, motivated by applications to telephone call centers. Since a customer can easily hang up after waiting for too long, abandonment is a non-negligible aspect in the study of many-server queues. In our study, a customer can leave the system (without getting service) once he/she has been waiting in queue for more than his patience time. A recent statistical study by Brown et al. [3] suggests that the exponential assumption on the service time distribution, in many cases, is not valid. In fact, the distribution of

J. Zhang (✉)
Department of Industrial Engineering and Logistic Management, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
e-mail: j.zhang@ust.hk

service times at call centers may be log-normal in some cases as shown in [3]. This emphasizes the need to look at the many-server model with generally distributed service and patience times.

In this paper we study many-server queues with general patience and service times. The queueing model is denoted by $G/GI/n+GI$. The G represents a general stationary arrival process. The first *GI* indicates that service times come from a sequences of independent and identically distributed (i.i.d.) random variables with a general distribution. The n denotes the number of homogeneous servers. There is an unlimited waiting space, called the buffer, where customers wait be served according to the first-come-first-served (FCFS) policy. Customers can choose to abandon if their patience times expire before their service starts. Again, the patience times are i.i.d. and with a general distribution (the *GI* after the '+' sign).

Useful insights can be obtained by considering a many-server queue in limit regimes where the number n of servers increases along the sequence with the arrival rate λ^n such that the traffic intensity

$$\rho^n = \frac{\lambda^n}{n\mu} \to \rho \quad \text{as } n \to \infty,$$

where μ is the service rate of a single server (in other words, the reciprocal of the mean service time), and $\rho \in [0, \infty)$. In our study, the limit ρ in the above need not to be less than 1. In fact, according to ρ, the limit regimes can be divided into three classes, i.e. *Efficiency-Driven* (ED) regime when $\rho > 1$, *Quality-and-Efficiency-Driven* (QED) regime when $\rho = 1$ and *Quality-Driven* (QD) regime when $\rho < 1$. The QED regime is also called the *Halfin–Whitt* regime due to the seminal work Halfin and Whitt [13]. With this motivation, we establish the fluid (also called law of large number) limit for the $G/GI/n+GI$ queue in all the ED, QED and QD limit regimes.

We show that the fluid model has an equilibrium, which yields approximations for various performance metrics. These fluid approximations work pretty well in the ED and QD regime where ρ is not that close to 1, as demonstrated in the numerical experiments of Whitt [31]. However, when ρ is very close (say within 5 %) to 1, the fluid approximations lose their accuracy and one could consider the more refined limit, the diffusion limit, in this case. The diffusion limit is not within the scope of the current paper.

In a system where multiple customers are processed simultaneously either by a single server via a sharing policy or by many servers such as the model we are studying in this paper, how to describe the system becomes an important issue. The number of customers in the system does not give much information since they may all have large remaining service times or all have small remaining service times, and this information can affect future evolution of the system. It will be nice to have a rich descriptor that can contain more information than just the headcount. So we choose to use finite Borel measures on $(0, \infty)$ to describe the system. At any time $t \geq 0$, in additional to recording the total number of customers in service (i.e. the number of busy servers), we record all the remaining service times using measure $\mathcal{Z}(t)$. For any Borel set $C \subset (0, \infty)$, $\mathcal{Z}(t)(C)$ indicates the number of customers in server with *remaining service time* belonging to C at that time. A similar idea applies for the remaining patience times. We first introduce the *virtual buffer*, which holds all the customers who have arrived but not yet scheduled to receive service according to the FCFS policy. We record all the remaining patience times for those in the virtual buffer using finite Borel measure $\mathcal{R}(t)$ on $\mathbb{R}=(-\infty,\infty)$. At time $t\geq 0$, $\mathcal{R}(t)(C)$ indicates the number of customers in the virtual buffer with *remaining patience time* belonging to the Borel set C. The descriptor $(\mathcal{R}(\cdot),\mathcal{Z}(\cdot))$ contains quite rich information in the sense that it reflects the residual service and patience times of all customers at time t, thus can help to write equations that reveal the dynamics of the system (cf. (2.4) and (2.5)). Also, traditional performance metrics can be recovered from the descriptor. For example, the actual number of customers in the buffer is

$$Q(t)=\mathcal{R}(t)\big((0,\infty)\big)\quad \text{for all } t\geq 0,$$

since a customer with negative or zero remaining patience time has already abandoned. More details will be discussed when we rigorously introduce the mathematical model in Sect. 2. In the literature, another descriptor that keeps track of the ages of customers in service and the ages of customers in waiting have been used, e.g. [17, 31]. In almost all systems, the age information of customers is observable. In other words, how long the customers have been in the system is recorded or can be found out. This is not the case for tracking the residuals, though in some special systems it is possible to observe the residual. However, from the system perspective, what we really care about is not the measure-valued process. Ultimately, we care about performance measures such as waiting time, queue length, etc. Modeling using measure to keep a rich information of either the "age" or the "residual" is for the purpose of analysis. Both age and residual descriptions of the system often results in the same steady state insights. In this paper we focus on residual processes only.

The framework of using measure-valued process has been successfully applied to study models where multiple customers are processed at the same time. Existing work includes Gromoll and Kruk [10], Gromoll, Puha and Williams [11] and Gromoll, Robert and Zwart [12], to name a few. Most of this work is on the processor sharing queue and related models where there is no waiting buffer. Recently, Zhang, Dai and Zwart [33, 34] applied the measure-valued process to study the limited processor sharing queue, where only limited number of customers can be served at any given time with extra customers waiting in a buffer. Some techniques in this paper closely follow from those developed in [33]. There has been a large literature on many-server queue and related models since the seminal work by Halfin and Whitt [13]. But there are not many successes with the case where the service time distribution is allowed to be non-exponential. One exception is the work of Reed [28], in which fluid and diffusion limits of the customer-count process of many server queues (without abandonment) are established where few assumptions beyond a first moment are placed on the service time distribution. Later, Puhalskii and Reed [26] extend the aforementioned results to allow noncritical loading, generally distributed service times, and general initial conditions. Jelenković et al. [15] study the many-server queue with deterministic service times; Garmarnik and Momčilović [8] study the model with lattice-valued service times; Puhalskii and Reiman [27] study the model with phase-type service time distributions. Mandelbaum and Momčilović [21] study the virtual waiting time processes, and Kaspi and Ramanan [18] study the fluid limit of measure-valued processes for many-server queues with general service times. Recently, [19] characterized the diffusion limit of the many-server queueing system via a stochastic partial differential equation.

For the many-server queue with abandonment, a version of the fluid model has been established as a conjecture in Whitt [30], where a lot of insights were demonstrated, which help greatly in our work. Recently, Kang and Ramanan also worked on the same topic and summarized their result in [17]. Although we focus on the same topic, our work uses different methodology from that in [17] and requires fewer assumptions on both the service time and the patience time distributions. From the modeling aspect, our approach mainly based on tracking the "residual" processes, while [17] tracks the "age" processes for studying the queueing model. As far as we understand in [17], modeling based on "ages" facilities the application of martingale techniques. However, to apply the martingale, there need to be some compensators which involve the hazard rate functions of the distributions. Thus, the distribution functions must have a density and there are some additional conditions on the hazard rate functions. By tracking the "residual", we can avoid using martingales and have a simpler representation of the dynamics of the limiting process. This not only simplifies the analysis, but also requires weaker assumptions. In this work, the only assumption on the service time distribution is continuity and the assumption on the patience time distribution is Lipchitz continuity. Since we have a simple representation of the fluid model, the existence of the solution to the fluid model is proved directly from the fluid dynamic equations (3.1)–(3.4) without invoking stochastic limit, while the existence is proved by showing each limit of the fluid scaled stochastic processes satisfies the fluid equations (3.8)–(3.16) in [17]. The ability to separate the analysis of fluid model from stochastic limit also help to explicitly characterize the equilibrium state of the fluid model, which is consistent with the one proposed in [30]. As demonstrated in [30], the equilibrium state yields approximation formulas for various performance measures of the stochastic model in the ED regime. In addition, we also verify at the end of this paper (cf. Sect. 6) that our fluid model is consistent with the special case where both service and patience times are exponentially distributed, as established in Whitt [30] for the ED regime, Garnet et al. [9] for QED regime and Pang and Whitt [24] and Puhalskii [25] for all three regimes.

Additional work on many-server queues with abandonment includes Dai, He and Tezcan [5] for phase-type service time distributions and exponential patience time distribution; Zeltyn and Mandelbaum [32] for exponential service time distribution and general patience time distributions; Mandelbaum and Momčilović [22] for both general service time distribution and general patience time distribution. The difference between our work and [22] is that we study the fluid limit of measure-valued processes in all three regimes, and [22] studies the diffusion limit of customer-count processes and virtual waiting processes in the QED regime. By assuming a convenient initial condition, [22] does not require a detailed fluid model analysis.

The paper is organized as follows: We begin in Sect. 2 by formulating the mathematical model of the $G/GI/n+GI$ queue. The dynamics of the system are clearly described by modeling with measure-valued processes; see (2.4) and (2.5). The main results, including a characterization of the fluid model and the convergence of the stochastic processes underlying the $G/GI/n+GI$ queue to the fluid model solution are stated in Sect. 3. In Sect. 4 we explore the fluid model and give proofs of all the results on the fluid model. Section 5 is devoted to establishing the convergence of stochastic processes, which includes the proof of pre-compactness and the characterization of the limit as the fluid model solution.

1.1 Notation

The following notation will be used throughout. Let $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{R}$ denote the set of natural numbers, integers and real numbers, respectively. Let $\mathbb{R}_+ = [0,\infty)$. For $a, b \in \mathbb{R}$, write a^+ for the positive part of a, $\lfloor a \rfloor$ for the integer part, $\lceil a \rceil$ for $\lfloor a \rfloor + 1$, $a \vee b$ for the maximum, and $a \wedge b$ for the minimum. For any $A \subset \mathbb{R}$, denote $\mathscr{B}(A)$ the collection of all Borel subsets which are subsets of A.

Let $\mathbf{M}$ denote the set of all non-negative finite Borel measures on $\mathbb{R}$, and $\mathbf{M}_+$ denote the set of all non-negative finite Borel measures on $(0,\infty)$. To simplify the notation, let us take the convention that for any Borel set $A \subset \mathbb{R}$, $\nu(A \cap (-\infty, 0]) = 0$ for any $\nu \in \mathbf{M}_+$. Also, by this convention, $\mathbf{M}_+$ is embedded as a subspace of $\mathbf{M}$. For $\nu_1, \nu_2 \in \mathbf{M}$, the Prohorov metric is defined to be

$$\begin{aligned}\mathbf{d}[\nu_1,\nu_2] = \inf\bigl\{&\epsilon > 0 : \nu_1(A) \le \nu_2\bigl(A^\epsilon\bigr) + \epsilon \text{ and} \\ &\nu_2(A) \le \nu_1\bigl(A^\epsilon\bigr) + \epsilon \text{ for all closed Borel set } A \subset \mathbb{R}\bigr\}, \qquad (1.1)\end{aligned}$$

where $A^\epsilon = \{b \in \mathbb{R} : \inf_{a\in A} |a-b| < \epsilon\}$. This is the metric that induces the topology of weak convergence of finite Borel measures. (See Sect. 6 in [2].) For any Borel measurable function $g : \mathbb{R} \to \mathbb{R}$, the integration of this function with respect to the measure $\nu \in \mathbf{M}$ is denoted by $\langle g, \nu\rangle$. We denote the zero measure in $\mathbf{M}$ by $\mathbf{0}$.

Let $\mathbf{M}_+ \times \mathbf{M}$ denote the Cartesian product. There are a number of ways to define the metric on the product space. For convenience we define the metric to be the maximum of the Prohorov metric between each component. With a little abuse of notation, we still use $\mathbf{d}$ to denote this metric, i.e.

$$\mathbf{d}\bigl[(\mu_1,\nu_1),(\mu_2,\nu_2)\bigr] = \max\bigl(\mathbf{d}[\mu_1,\mu_2], \mathbf{d}[\nu_1,\nu_2]\bigr)$$

for any $(\mu_1,\nu_1), (\mu_2,\nu_2) \in \mathbf{M}_+ \times \mathbf{M}$.

Let $(\mathbf{E},\pi)$ be a general metric space. We consider the space $\mathbf{D}$ of all right-continuous $\mathbf{E}$-valued functions with finite left limits defined either on a finite interval $[0,T]$ or the infinite interval $[0,\infty)$. We refer to the space as $\mathbf{D}([0,T],\mathbf{E})$ or $\mathbf{D}([0,\infty),\mathbf{E})$ depending upon the function domain. The space $\mathbf{D}$ is also known as the space of *càdlàg* functions. For $g(\cdot), g'(\cdot) \in \mathbf{D}([0,T],\mathbf{E})$, the uniform metric is defined as

$$\upsilon_T\bigl[g,g'\bigr] = \sup_{0\le t\le T} \pi\bigl[g(t), g'(t)\bigr]. \qquad (1.2)$$

However, a more useful metric we will use is the following Skorohod J_1 metric:

$$\varrho_T\bigl[g,g'\bigr] = \inf_{f\in\Lambda_T} \bigl(\|f\|_T^\circ \vee \upsilon_T\bigl[g, g'\circ f\bigr]\bigr), \qquad (1.3)$$

where $g \circ f(t) = g(f(t))$ for $t \geq 0$ and Λ_T is the set of strictly increasing and continuous mapping of $[0, T]$ onto itself and

$$\|f\|_T^\circ = \sup_{0 \leq s < t \leq T} \left| \log \frac{f(t) - f(s)}{t - s} \right|.$$

If $g(\cdot)$ and $g'(\cdot)$ are in the space $\mathbf{D}([0, \infty), \mathbf{E})$, the Skorohod J_1 metric is defined as

$$\varrho[g, g'] = \int_0^\infty e^{-T} \big(\varrho_T[g, g'] \wedge 1\big)\, dT. \tag{1.4}$$

By saying convergence in the space $\mathbf{D}$, we mean the convergence under the Skorohod J_1 topology, which is the topology induced by the Skorohod J_1 metric [7].

We use "$\rightarrow$" to denote the convergence in the metric space $(\mathbf{E}, \pi)$, and use "$\Rightarrow$" to denote the convergence in distribution of random variables taking value in the metric space $(\mathbf{E}, \pi)$.

2 Stochastic model

In this section we first describe the $G/GI/n+GI$ queueing system and then introduce a pair of measure-valued processes that capture the dynamics of the system.

There are n identical servers in the system. Customers arrive according to a general stationary arrival process (the initial G) with arrival rate λ. Let a_i denote the arrival time of the ith arriving customer, $i = 1, 2, \ldots$. An arriving customer enters service immediately upon arrival if there is a server available. If all n servers are busy, the arriving customer waits in a buffer, which has infinite capacity. Customers are served in the order of their arrival by the first available server. Waiting customers may also elect to abandon. We assume that each customer has a random patience time. A customer abandons the system once the time he has waited in the buffer exceeds his patience time. Once a customer starts his service, the customer remains until the service is completed. There are no retrials; abandoning customers leave without affecting future arrivals.

The two GIs in the notation mean that the service times and patience times come from two independent sequences of i.i.d. random variables; these two sequences are assumed to be independent of the arrival process. Let u_i and v_i denote the patience and service time of the ith arriving customer, $i = 1, 2, \ldots$. In many applications such as telephone call centers, customers cannot see the queue (the case of invisible queues, cf. [23]), thus do not know the experience of other customers. This provides some justifications for us to assume that the patience times are i.i.d. Denote $F(\cdot)$ and $G(\cdot)$ the distributions for the patience and service times, respectively.

To describe the system using measure-valued process, we first introduce the notion of *virtual buffer*. For the real physical buffer, a customer joins upon arrival (if all servers are busy) and leaves upon starting service or his waiting time exceeding his patience time. The difference for the virtual buffer is that a customer does not leave immediately upon his waiting time exceeding his patience time. He stays in the virtual buffer but will be tagged as "abandoned". The virtual buffer is also served based on

FCFS. When a customer in the virtual buffer is about to be admitted into service, the system checks whether he has been tagged as "abandoned" or not. If yes, the system discards this customer (now, this abandoned customer leaves the virtual buffer) and picks the next customer in the virtual buffer and performs the same check. Otherwise, the system admits the customer into service. At any time $t \geq 0$, $\mathcal{R}(t)$ denotes the random measure in $\mathbf{M}$ such that $\mathcal{R}(t)(C)$ is the number of customers in the virtual buffer with remaining patience time in $C \in \mathscr{B}(\mathbb{R})$. Note that this way of modeling requires the measure $\mathcal{R}(\cdot)$ to be defined on $\mathbb{R}$, not just $(0, \infty)$. In fact, the remaining patience time being non-positive serves as the tag for indicating a customer being abandoned. It is clear that

$$Q(t) = \mathcal{R}(t)\big((0, \infty)\big) \quad \text{and} \quad R(t) = \mathcal{R}(t)(\mathbb{R}) \tag{2.1}$$

represent the number of customers waiting in the real buffer and the number of customers in the virtual buffer, respectively.

We also use a measure to describe the servers. At any time $t \geq 0$, $\mathcal{Z}(t)$ denotes a measure in $\mathbf{M}_+$ such that $\mathcal{Z}(t)(C)$ is the number of customers in service with remaining service time in $C \in \mathscr{B}((0, \infty))$. Differently from the virtual buffer, the servers only hold customers with positive remaining service times, so we only care about the subsets in $(0, \infty)$. The quantity

$$Z(t) = \mathcal{Z}(t)\big((0, \infty)\big), \tag{2.2}$$

represents the number of customers in service at any time $t \geq 0$.

The measure-valued (taking value in $\mathbf{M} \times \mathbf{M}_+$) stochastic process $(\mathcal{R}(\cdot), \mathcal{Z}(\cdot))$ serves as the descriptor for the $G/GI/n{+}GI$ queueing model. Before we use it to describe the dynamics of the system, let us first talk about the initial condition, since the system is allowed to be non-empty initially. The initial state specifies $R(0)$, the number of customers in the virtual buffer as well as their remaining patience times u_i and service times v_i, $i = 1 - R(0), 2 - R(0), \ldots, 0$. The initial state also specifies $Z(0)$, the number of customers in service as well as their remaining service times v_i, $i = 1 - R(0) - Z(0), \ldots, -R(0)$. Briefly, the initial customers are given negative index, in order not to conflict with the index of arriving customers. Those initial customers in the buffer are also assumed to have i.i.d. service times with distribution $G(\cdot)$. For each $t \geq 0$, denote $E(t)$ the number of customers that has arrived during the time interval $(0, t]$. Arriving customers are indexed by $1, 2, \ldots$ according to the order of their arrival. From this way of indexing customers, it is clear that the index of the head-of-the-line customer in the virtual buffer at time $t \geq 0$ is $B(t) + 1$, where

$$B(t) = E(t) - R(t). \tag{2.3}$$

In other words, $B(t)$ can be interpreted as the number of customers in the virtual buffer who could have entered the service by time t. It should be pointed out that not all of them really get the service. At the time a customer in the virtual buffer is allowed to enter service, the system will check whether he has abandon the system (i.e. remaining patience time is less than 0). The customer can only be admitted into service if he has not abandoned the system at that time. Denote by τ_i the time when the ith job starts *service* for all $i \geq 1 - R(0)$. For $i < 0$, a_i may be a negative number indicating how long the ith customer had been there by time 0. We will impose some conditions on a_i's with $i < 0$ later on. Let δ_x and $\delta_{(x,y)}$ denote the Dirac point measure at $x \in \mathbb{R}$ and $(x, y) \in \mathbb{R}^2$, respectively. Denote $C + x = \{c + x : x \in C\}$ for any subset $C \subset \mathbb{R}$. For any subsets $C, C' \subset \mathbb{R}$, let $C \times C'$ denote the Cartesian product. Using the Dirac measure and the above introduced notations, the evolution of the virtual buffer and the servers can be captured by the following *stochastic dynamic equations*:

$$\mathcal{R}(t)(C) = \sum_{i=1+B(t)}^{E(t)} \delta_{u_i}(C + t - a_i), \quad \text{for all } C \in \mathscr{B}(\mathbb{R}), \tag{2.4}$$

$$\begin{aligned} \mathcal{Z}(t)(C) = & \sum_{i=1-R(0)-Z(0)}^{-R(0)} \delta_{v_i}(C + t) \\ & + \sum_{i=1-R(0)}^{B(t)} \delta_{(u_i,v_i)}(C_0 + \tau_i - a_i) \times (C + t - \tau_i), \\ & \text{for all } C \in \mathscr{B}\big((0, \infty)\big), \end{aligned} \tag{2.5}$$

for all $t \geq 0$. Note that the buffer size $Q(t)$ can be recovered from $\mathcal{R}(t)$ via (2.1). We can see from the second term on the right hand side of (2.5) that customers whose waiting time $\tau_i - a_i$ is longer than their patience time u_i do not actually enter the service.

Denote the total number of customers in the system by

$$X(t) = Q(t) + Z(t) \quad \text{for all } t \geq 0. \tag{2.6}$$

The following *non-idling constraints* must be satisfied at any time $t \geq 0$:

$$Q(t) = \big(X(t) - n\big)^+, \tag{2.7}$$

$$Z(t) = \big(X(t) \wedge n\big), \tag{2.8}$$

where n, as introduced above, denotes the number of servers in the system.

3 Main results

The main results of this paper contain two parts. The first part is a characterization of the fluid model, including the existence and uniqueness of the fluid model solution, and the equilibrium of the fluid model; these results are summarized in Sect. 3.1. The second part is the convergence of the stochastic processes to the fluid model solution; this result is stated in Sect. 3.2.

3.1 Fluid model

To study the stochastic model, we introduce a deterministic fluid model. To simplify notations, let $F^c(\cdot)$ denote the complement of the patience time distribution $F(\cdot)$, i.e.

$F^c(x) = 1 - F(x)$ for all $x \in \mathbb{R}$; the complement of the service time distribution, denoted by $G^c(\cdot)$, is defined in the same way. For the remaining of the paper, let $C_x = (x, \infty)$. We introduce the following *fluid dynamic equations*:

$$\bar{\mathcal{R}}(t)(C_x) = \lambda \int_{t-\frac{\bar{R}(t)}{\lambda}}^{t} F^c(x+t-s)\,ds, \quad t \geq 0,\ x \in \mathbb{R}, \tag{3.1}$$

$$\bar{\mathcal{Z}}(t)(C_x) = \bar{\mathcal{Z}}(0)(C_x + t) + \int_0^t F^c\left(\frac{\bar{R}(s)}{\lambda}\right) G^c(x+t-s)\,d\bar{B}(s),$$
$$t \geq 0,\ x \in (0, \infty), \tag{3.2}$$

where

$$\bar{R}(s) = \bar{\mathcal{R}}(s)(\mathbb{R}) \quad \text{and} \quad \bar{B}(s) = \lambda s - \bar{R}(s).$$

Here, all the time dependent quantities are assumed to be right-continuous on $[0, \infty)$ and to have left limits in $(0, \infty)$. The integral $\int_0^t g(s)\,d\bar{B}(s)$ is interpreted as the Lebesgue–Stieltjes integral on the interval $(0, t]$.

Here are some intuitive explanations about the fluid dynamic equations. They are not meant to serve as rigorous analysis (which will be presented in Sect. 4), but just to facilitate readers to gain some intuitive understanding. Suppose a unit amount of “fluid customers” arrives at time s. The proportion of the fluid amount with patience time larger than x is $F^c(x)$. Observing the system at time t, $t - s$ amount of time has passed, thus the proportion of the fluid mount with remaining patience time large than x becomes $F^c(x + t - s)$. The integration in (3.1) starts from $t - \frac{\bar{R}(t)}{t}$ because the “oldest” customer in the virtual buffer arrives at that time. For (3.2), note that the amount of fluid with remaining service time larger than x consists of two parts. The first part is the initial fluid customers in service, only those whose remaining service time larger than $x + t$ at time 0 will have remaining service time larger than x at time t. The second part contains all those who joined the service during time $[0, t]$. Since at time s, those who are about to join the service have been waiting for $\bar{R}(s)/\lambda$, $F(\bar{R}(s)/\lambda)$ fraction of them have already abandoned the system. The rest $F^c(\bar{R}(s)/\lambda)$ go ahead to get the service. View the system at time t, $t - s$ amount of time has passed, so the proportion of the fluid amount with remaining service time larger than s is $G^c(x + t - s)$. The integration is with respect to $d\bar{B}(s)$ because that is the rate of customers moving from the virtual buffer to service.

The quantities $\bar{Q}(\cdot)$, $\bar{Z}(\cdot)$ and $\bar{X}(\cdot)$ are defined in the same way as their stochastic counterparts in (2.1), (2.2), and (2.6). The following non-idling constraints must be satisfied for all $t \geq 0$:

$$\bar{Q}(t) = \big(\bar{X}(t) - 1\big)^+, \tag{3.3}$$

$$\bar{Z}(t) = \big(\bar{X}(t) \wedge 1\big). \tag{3.4}$$

The fluid dynamic equations (3.1) and (3.2) and the non-idling constraints (3.3) and (3.4) define a *fluid model*, which is denoted by (λ, F, G).

Denote $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0) = (\bar{\mathcal{R}}(0), \bar{\mathcal{Z}}(0))$ the initial condition of the fluid model. For the convenience of notations, also denote $\bar{Q}_0 = \bar{Q}(0)$, $\bar{Z}_0 = \bar{Z}(0)$ and $\bar{X}_0 = \bar{Q}_0 + \bar{Z}_0$.

We need to require that the initial condition satisfy the dynamic equations and the non-idling constraints, i.e.

$$\bar{\mathcal{R}}_0(C_x) = \lambda \int_0^{\frac{\bar{R}_0}{\lambda}} F^c(x+s)\,ds, \quad x \in \mathbb{R}, \tag{3.5}$$

$$\bar{Q}_0 = (\bar{X}_0 - 1)^+, \tag{3.6}$$

$$\bar{Z}_0 = (\bar{X}_0 \wedge 1). \tag{3.7}$$

We also require that

$$\bar{\mathcal{Z}}_0\bigl(\{0\}\bigr) = 0, \tag{3.8}$$

which means that nobody with remaining service time 0 occupies a server. We call any element $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0) \in \mathbf{M} \times \mathbf{M}_+$ a *valid* initial condition if it satisfies (3.5)–(3.8).

We call $(\bar{\mathcal{R}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \in \mathbf{D}([0,\infty), \mathbf{M} \times \mathbf{M}_+)$ a solution to the fluid model (λ, F, G) with a valid initial condition $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$ if it satisfies the fluid dynamic equations (3.1) and (3.2) and the non-idling constraints (3.3) and (3.4).

Denote by μ the reciprocal of first moment of the service time distribution $G(\cdot)$. Let

$$M_F = \inf\bigl\{x \ge 0 : F(x) = 1\bigr\}. \tag{3.9}$$

By the right-continuity of distribution functions, it is clear that $F(x) < 1$ for all $x < M_F$ and $F(x) = 1$ for all $x \ge M_F$.

Theorem 3.1 (Existence and Uniqueness) *Assume the service time distribution $G(\cdot)$ and its mean $1/\mu$ satisfy that*

$$G(\cdot) \textit{ is continuous}, \tag{3.10}$$

and

$$0 < \mu < \infty. \tag{3.11}$$

Assume that the patience time distribution $F(\cdot)$ satisfies that

$$F(\cdot) \textit{ is Lipschitz continuous}. \tag{3.12}$$

There exists a unique solution to the fluid model (λ, F, G) for any valid initial condition $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$.

The above theorem provides the foundation to further study the fluid model. A key property is that the fluid model has an equilibrium state. An equilibrium state is defined as follows.

Definition 3.1 An element $(\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty) \in \mathbf{M} \times \mathbf{M}_+$ is called an *equilibrium state* for the fluid model (λ, F, G) if the solution to the fluid model with a valid initial condition $(\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$ satisfies

$$\bigl(\bar{\mathcal{R}}(t), \bar{\mathcal{Z}}(t)\bigr) = (\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty) \quad \text{for all } t \ge 0.$$

This definition says that if a fluid model solution starts from an equilibrium state, it will never change in the future. To present the result about equilibrium state, we need to introduce some more notation. For the service time distribution function $G(\cdot)$ on $\mathbb{R}_+$, the associated *equilibrium* distribution is given by

$$G_e(x) = \mu \int_0^x G^c(y)\, dy, \quad \textit{for all } x \geq 0.$$

Theorem 3.2 *Assume the conditions in Theorem 3.1. The state $(\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$ is an equilibrium state of the fluid model (λ, F, G) if and only if it satisfies*

$$\bar{\mathcal{R}}_\infty(C_x) = \lambda \int_0^w F^c(x+s)\, ds, \quad x \in \mathbb{R}, \tag{3.13}$$

$$\bar{\mathcal{Z}}_\infty(C_x) = \min(\rho, 1)\bigl[1 - G_e(x)\bigr], \quad x \in (0, \infty), \tag{3.14}$$

where w is a solution to the equation

$$F(w) = \max\biggl(\frac{\rho - 1}{\rho}, 0\biggr). \tag{3.15}$$

Remark 3.1 If (3.15) has multiple solutions, then the equilibrium is not unique (any solution w gives an equilibrium). If the equation has a unique solution (for example when $F(\cdot)$ is strictly increasing), then the equilibrium state is unique.

The quantity w is interpreted to be the *virtual* waiting time for an arriving customer. If his patience time exceeds w, he will not abandon. Thus, the probability of his abandonment is given by $F(w)$, which is equal to $(\rho - 1)/\rho$ when $\rho > 1$; the latter quantity is the fraction of traffic that has to be discarded due to the overloading. From (3.13), $\bar{\mathcal{R}}_\infty(C_x) = \lambda w$ for $x \leq -w$. Thus, the average number of customers in the virtual buffer is

$$\bar{R}_\infty = \bar{\mathcal{R}}_\infty(\mathbb{R}) = \lambda w,$$

which is consistent with Little's law. From (3.14), the average number of busy servers is

$$\bar{Z}_\infty = \bar{\mathcal{Z}}_\infty\bigl((0, \infty)\bigr) = \min(\rho, 1).$$

If $\rho > 1$, then one expects all servers to be busy, whereas, if $\rho < 1$, there will be no abandonment (on the fluid scaling) so that the fraction of busy servers will be ρ. These observations and interpretations were first made by Whitt [31], where approximation formulas based on a conjectured fluid model were also given, and were compared with extensive simulation results. The approximation formulas derived from our fluid model is consistent with those formulas in Whitt [31].

3.2 Convergence of stochastic models

We consider a sequence of queueing systems indexed by the number of servers n, with $n \to \infty$. Each model is defined in the same way as in Sect. 2. The arrival rate of each model is assumed be to proportional to n. To distinguish models with different indices, quantities of the nth model are accompanied with superscript n. Each model may be defined on a different probability space $(\Omega^n, \mathcal{F}^n, \mathbb{P}^n)$. Our results concern the asymptotic behavior of the descriptors under the *fluid* scaling, which is defined by

$$\bar{\mathcal{R}}^n(t) = \frac{1}{n}\mathcal{R}^n(t), \qquad \bar{\mathcal{Z}}^n(t) = \frac{1}{n}\mathcal{Z}^n(t), \tag{3.16}$$

for all $t \geq 0$. The fluid scaling for the arrival process $E^n(\cdot)$ is defined in the same way, i.e.

$$\bar{E}^n(t) = \frac{1}{n}E^n(t),$$

for all $t \geq 0$. We assume that

$$\bar{E}^n(\cdot) \Rightarrow \lambda \cdot \quad \text{as } n \to \infty. \tag{3.17}$$

Since the limit is deterministic, the convergence in distribution in (3.17) is equivalent to convergence in probability; namely, for each $T > 0$ and each $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}^n\Big(\sup_{0\leq t\leq T}\big|\bar{E}^n(t) - \lambda t\big| > \epsilon\Big) = 0.$$

Denote ν_F^n and ν_G^n the probability measures corresponding to the patience time distribution F^n and the service time distribution G^n, respectively. Assume that as $n \to \infty$,

$$\nu_F^n \to \nu_F, \qquad \nu_G^n \to \nu_G, \tag{3.18}$$

where ν_F and ν_G are some probability measures associated with distribution functions F and G, the convergence is in the Prohorov metric as defined in Sect. 1.1. Also, the following initial condition will be assumed:

$$\big(\bar{\mathcal{R}}^n(0), \bar{\mathcal{Z}}^n(0)\big) \quad \Rightarrow \quad (\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0) \quad \text{as } n \to \infty, \tag{3.19}$$

where, almost surely, $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$ is a valid initial condition and

$$\bar{\mathcal{R}}_0 \text{ and } \bar{\mathcal{Z}}_0 \text{ has no atoms.} \tag{3.20}$$

Theorem 3.3 *In addition to the assumptions* (3.10)–(3.12) *in Theorem* 3.1, *if the sequence of many-server queues satisfies* (3.17)–(3.20), *then*

$$\big(\bar{\mathcal{R}}^n(\cdot), \bar{\mathcal{Z}}^n(\cdot)\big) \quad \Rightarrow \quad \big(\bar{\mathcal{R}}(\cdot), \bar{\mathcal{Z}}(\cdot)\big) \quad \textit{as } n \to \infty,$$

where, almost surely, $(\bar{\mathcal{R}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ *is the unique solution to the fluid model* (λ, F, G) *with initial condition* $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$.

Corollary 3.1 *Under the same assumption as Theorem 3.3, as $n \to \infty$, the fluid scaled process $\bar{X}^n(\cdot)$ converges weakly to the solution to the following equation*:

$$\bar{X}(t) = \zeta_0(t) + \rho \int_0^t H\big(\big(\bar{X}(t-s)-1\big)^+\big)\, dG_e(s) + \int_0^t \big(\bar{X}(t-s)-1\big)^+ \, dG(s). \tag{3.21}$$

4 Properties of the fluid model

In this section we analyze the proposed fluid model and establish some basic properties of the fluid model solution. The proof of Theorem 3.1 for existence and uniqueness and the proof of Theorem 3.2 for characterization of the equilibrium will be presented in Sect. 4.1 and Sect. 4.2, respectively.

4.1 Existence and uniqueness of fluid model solutions

We first present some calculus on the fluid dynamic equations (3.1) and (3.2), which define the fluid model. It follows from (3.1) that

$$\bar{Q}(t) = \bar{\mathcal{R}}(t)(C_0) = \lambda \int_{t-\frac{\bar{R}(t)}{\lambda}}^{t} F^c(t-s)\, ds = \lambda \int_0^{\frac{\bar{R}(t)}{\lambda}} F^c(s)\, ds.$$

Let

$$F_d(x) = \int_0^x \big[1 - F(y)\big]\, dy \quad \text{for all } x \geq 0.$$

Note that the density of $F_d(\cdot)$ is not scaled by the mean of $F(\cdot)$. Thus, this is not exactly the equilibrium distribution associated with $F(\cdot)$. In fact, we do not need the mean

$$N_F = \int_0^\infty \big[1 - F(y)\big]\, dy \tag{4.1}$$

to be finite. Now we have

$$\frac{\bar{Q}(t)}{\lambda} = F_d\bigg(\frac{\bar{R}(t)}{\lambda}\bigg). \tag{4.2}$$

It follows from (3.2) that

$$\begin{aligned}\bar{Z}(t) &= \bar{\mathcal{Z}}(t)(C_0)\\ &= \bar{\mathcal{Z}}_0(C_0+t) + \lambda \int_0^t F^c\bigg(\frac{\bar{R}(s)}{\lambda}\bigg) G^c(t-s)\, ds\\ &\quad - \int_0^t F^c\bigg(\frac{\bar{R}(s)}{\lambda}\bigg) G^c(t-s)\, d\bar{R}(s).\end{aligned}$$

Note that, by (4.2), $d\bar{Q}(s) = F^c(\frac{\bar{R}(s)}{\lambda})\,d\bar{R}(s)$. So

$$\bar{Z}(t) = \bar{\mathcal{Z}}_0(C_0 + t) + \frac{\lambda}{\mu}\int_0^t F^c\left(\frac{\bar{R}(s)}{\lambda}\right) dG_e(t-s) - \int_0^t G^c(t-s)\,d\bar{Q}(s).$$

Performing change of variable and integration by parts, we have

$$\begin{aligned}\bar{Z}(t) = {}& \bar{\mathcal{Z}}_0(C_t) + \frac{\lambda}{\mu}\int_0^t F^c\left(\frac{\bar{R}(t-s)}{\lambda}\right) dG_e(s) \\ & - \bar{Q}(t)G^c(0) + \bar{Q}(0)G^c(t) + \int_0^t \bar{Q}(t-s)\,dG(s). \end{aligned} \tag{4.3}$$

We wish to represent the term $F^c(\frac{\bar{R}(\cdot)}{\lambda})$ using $\bar{Q}(\cdot)$. Recall M_F and N_F, which are defined in (3.9) and (4.1), respectively. It is clear that $F_d(x)$ is strictly monotone for $x \in [0, M_F)$. Thus, $F_d^{-1}(y)$ is well defined for each $y \in [0, N_F)$. We define $F_d^{-1}(y) = M_F$ for all $y \geq N_F$. Thus, (4.2) implies that

$$F^c\left(\frac{\bar{R}(t)}{\lambda}\right) = F^c\left(F_d^{-1}\left(\frac{\bar{Q}(t)}{\lambda}\right)\right). \tag{4.4}$$

Note that $G^c(0) = 1$ by assumption (3.10). Combining (3.3), (3.4), (4.3), and (4.4), we obtain

$$\begin{aligned}\bar{X}(t) = {}& \bar{\mathcal{Z}}_0(C_t) + \bar{Q}_0 G^c(t) \\ & + \frac{\lambda}{\mu}\int_0^t F^c\left(F_d^{-1}\left(\frac{(\bar{X}(t-s)-1)^+}{\lambda}\right)\right) dG_e(s) \\ & + \int_0^t \left(\bar{X}(t-s)-1\right)^+ dG(s). \end{aligned}$$

Now, introduce

$$H(x) = \begin{cases} F^c(F_d^{-1}(\frac{x}{\lambda})) & \text{if } 0 \leq x < \lambda N_F, \\ 0 & \text{if } x \geq \lambda N_F, \end{cases} \tag{4.5}$$

and $\zeta_0(\cdot) = \bar{\mathcal{Z}}_0(C_0 + \cdot) + \bar{Q}_0 G^c(\cdot)$. It then follows that

$$\bar{X}(t) = \zeta_0(t) + \rho\int_0^t H\left(\left(\bar{X}(t-s)-1\right)^+\right) dG_e(s) + \int_0^t \left(\bar{X}(t-s)-1\right)^+ dG(s). \tag{4.6}$$

Note that $\zeta_0(\cdot)$ depends only on the initial condition and $H(\cdot)$ is a function defined by the arrival rate λ and the patience time distribution $F(\cdot)$. Equation (4.6) serves as a key to the analysis of the fluid model.

Proof of Theorem 3.1 We first prove the existence. Given a valid initial condition $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$ (i.e. an element in $\mathbf{M} \times \mathbf{M}_+$ that satisfies (3.5)–(3.8)), we now construct a solution $(\bar{\mathcal{R}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ to the fluid model (λ, F, G) with this initial condition.

It is clear that $\zeta_0(\cdot)$ satisfies condition (A.3) of Lemma A.2 since the initial condition is valid. By condition (3.12), $F(\cdot)$ is Lipschitz continuous. Let L_F denote the Lipschitz constant. First, consider the case where $N_F < \infty$. Fix any $\delta \in (0, \lambda N_F)$. By the definition of $H(\cdot)$ in (4.5), for any $x_1, x_2 \in [0, \lambda N_F - \delta]$,

$$\begin{aligned}\bigl|H(x_2) - H(x_1)\bigr| &\le L_F \left| F_d^{-1}\left(\frac{x_2}{\lambda}\right) - F_d^{-1}\left(\frac{x_1}{\lambda}\right)\right| \\ &\le L_F \sup_{x \in [0, y_\delta]} \frac{1}{1 - F(y)} \frac{1}{\lambda} |x_2 - x_1| \\ &\le L_F \frac{1}{1 - F(y_\delta)} \frac{1}{\lambda} |x_2 - x_1|,\end{aligned}$$

where $y_\delta = F_d^{-1}(N_F - \delta/\lambda)$. Since $F_d(y_\delta) < N_F$, $F(y_\delta) < 1$. So $H(\cdot)$ is Lipschitz continuous on $[0, \lambda N_F - \delta]$. Next, for the case where $N_F = \infty$, the above argument remains true if we replace $\lambda N_F - \delta$ by any $M > 0$. Thus, the function $H(\cdot)$ satisfies the condition (A.6) in Lemma A.2. It is also clear that $H(\cdot)$ satisfies the conditions (A.4)–(A.5). It follows from Lemma A.2 that equation (4.6) has a unique solution $\bar{X}(\cdot)$. Define $\bar{Q}(t) = (\bar{X}(t) - 1)^+$. We now claim that $\bar{Q}(t)/\lambda \le N_F$ for all $t \ge 0$. The claim is automatically true if $N_F = \infty$. Now, let us consider the case where $N_F < \infty$. Since $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$ is a valid initial condition, $\bar{Q}(0)/\lambda \le N_F$. Suppose there exists $t_1 > 0$ such that $\bar{Q}(t_1)/\lambda > N_F$. Let $t_0 = \sup\{s : \bar{Q}(s)/\lambda \le N_F, s \le t_1\}$. So we have $\lim_{t \to t_0} \bar{Q}(t)/\lambda \le N_F$, since $\bar{Q}(\cdot)$ has left limit. Let $\delta = (Q(t_1)/\lambda - N_F)/4$ and pick $t_\delta \in [t_0 - \delta, t_0]$ such that $\bar{Q}(t_\delta)/\lambda \le N_F + \delta$. By Lemma A.3,

$$\frac{\bar{Q}(t')}{\lambda} - \frac{\bar{Q}(t)}{\lambda} \le \int_t^{t'} F^c\left(F_d^{-1}\left(\frac{\bar{Q}(s)}{\lambda}\right)\right) ds \tag{4.7}$$

for any $t < t'$. This gives

$$\begin{aligned}\frac{\bar{Q}(t_1)}{\lambda} &\le \frac{\bar{Q}(t_\delta)}{\lambda} + \int_{t_\delta}^{t_1} \left[1 - F\left(F_d^{-1}\left(\frac{\bar{Q}(s)}{\lambda}\right)\right)\right] ds \\ &\le N_F + \delta + \int_{t_\delta}^{t_0} 1\, ds + \int_{t_0}^{t_1} 0\, ds \\ &\le N_F + 2\delta < \frac{\bar{Q}(t_1)}{\lambda},\end{aligned}$$

which is a contradiction. This proves the claim. Let

$$\begin{aligned}\bar{Z}(t) &= \min\bigl(\bar{X}(t), 1\bigr), \\ \bar{R}(t) &= \lambda F_d^{-1}\left(\frac{\bar{Q}(t)}{\lambda}\right), \\ \bar{B}(t) &= \lambda t - \bar{R}(t),\end{aligned}$$

for all $t \geq 0$. We now construct a fluid model solution by letting

$$\bar{\mathcal{R}}(t)(C_x) = \lambda \int_{t-\frac{\bar{R}(t)}{\lambda}}^{t} F^c(x+t-s)\,ds, \tag{4.8}$$

$$\bar{\mathcal{Z}}(t)(C_x) = \bar{\mathcal{Z}}_0(C_x+t) + \int_0^t F^c\left(\frac{\bar{R}(s)}{\lambda}\right) G^c(x+t-s)\,d\bar{B}(s), \tag{4.9}$$

for all $t \geq 0$. According to (4.2), the integral in the above involving $d\bar{B}(s)$ can be written as

$$\begin{aligned}
&\int_0^t F^c\left(\frac{\bar{R}(s)}{\lambda}\right) G^c(x+t-s)\,d\bar{B}(s) \\
&\quad = \int_0^t F^c\left(\frac{\bar{R}(s)}{\lambda}\right) G^c(x+t-s)\lambda\,ds - \int_0^t F^c\left(\frac{\bar{R}(s)}{\lambda}\right) G^c(x+t-s)\,d\bar{R}(s) \\
&\quad = \int_0^t F^c\left(\frac{\bar{R}(s)}{\lambda}\right) G^c(x+t-s)\lambda\,ds - \int_0^t G^c(x+t-s)\,d\bar{Q}(s). \qquad (4.10)
\end{aligned}$$

For any given partition $0 = t_0 < t_1 < \cdots < t_K = T$, let $I^+ = \{k : \bar{Q}(t_k) - \bar{Q}(t_{k-1}) \geq 0\}$ and $I^- = \{k : \bar{Q}(t_k) - \bar{Q}(t_{k-1}) < 0\}$. According to (4.7)

$$\begin{aligned}
\sum_{k=1}^{K} \left|\bar{Q}(t_k) - \bar{Q}(t_{k-1})\right| &= \sum_{k\in I^+} \left[\bar{Q}(t_k) - \bar{Q}(t_{k-1})\right] - \sum_{k\in I^-} \left[\bar{Q}(t_k) - \bar{Q}(t_{k-1})\right] \\
&= 2\sum_{k\in I^+} \left[\bar{Q}(t_k) - \bar{Q}(t_{k-1})\right] + \bar{Q}(0) - \bar{Q}(T) \\
&\leq 2\lambda T + \bar{Q}(0) - \bar{Q}(T).
\end{aligned}$$

This implies that $\bar{Q}(\cdot)$ has bounded total variation. So the integral in (4.10) is well defined. Thus, the integral in (4.9) is also well defined. It is clear that the above defined $(\bar{\mathcal{R}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ satisfies the fluid dynamic equations (3.1) and (3.2) and constraints (3.3) and (3.4). So we conclude that $(\bar{\mathcal{R}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ is a fluid model solution.

It now remains to show the uniqueness. Suppose there is another solution to the fluid model (λ, F, G) with initial condition $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$, denoted by $(\bar{\mathcal{R}}^\dagger(\cdot), \bar{\mathcal{Z}}^\dagger(\cdot))$. Similarly, denote

$$\bar{R}^\dagger(t) = \bar{\mathcal{R}}^\dagger(\mathbb{R}),$$

$$\bar{Z}^\dagger(t) = \bar{\mathcal{Z}}^\dagger\big((0,\infty)\big),$$

for all $t \geq 0$. It must satisfy the fluid dynamic equations (3.1) and (3.2) and constraints (3.3) and (3.4). For all $t \geq 0$, let

$$\bar{Q}^\dagger(t) = \lambda F_d\left(\frac{\bar{R}^\dagger(t)}{\lambda}\right).$$

According to the algebra at the beginning of Sect. 4.1, $\bar{X}^\dagger(\cdot)$ must also satisfy (4.6). By the uniqueness of the solution to (4.6) in Lemma A.1,

$$\bar{X}^\dagger(t) = \bar{X}(t) \quad \text{for all } t \geq 0.$$

This implies that $\bar{R}^\dagger(t) = \bar{R}(t)$. By the dynamic equations (3.1) and (3.2), we must have

$$\bigl(\bar{\mathcal{R}}^\dagger(t), \bar{\mathcal{Z}}^\dagger(t)\bigr) = \bigl(\bar{\mathcal{R}}(t), \bar{\mathcal{Z}}(t)\bigr) \quad \text{for all } t \geq 0.$$

This completes the proof. □

4.2 Equilibrium state of the fluid model solution

In this section we first intuitively explain what an equilibrium should be. Then we rigorously prove it in Theorem 3.2. To provide some intuition, note that in the equilibrium, by (3.1), one should have

$$\bar{\mathcal{R}}_\infty(C_x) = \lambda \int_0^{\bar{R}_\infty/\lambda} F^c(x+s)\,ds,$$

for the buffer. This immediately implies that

$$\bar{\mathcal{R}}_\infty(C_x) = \lambda\left[F_d\left(x + \frac{\bar{R}_\infty}{\lambda}\right) - F_d(x)\right].$$

So the rate at which customers leave the buffer due to abandonment is

$$\lim_{x\to 0} \frac{\bar{\mathcal{R}}_\infty(C_0) - \bar{\mathcal{R}}_\infty(C_x)}{x} = \lambda F\left(\frac{\bar{R}_\infty}{\lambda}\right).$$

In the equilibrium, intuitively, the number of customers in service should not change and the distribution for the remaining service time should be the equilibrium distribution $G_e(\cdot)$, i.e.

$$\bar{\mathcal{Z}}_\infty(C_x) = \bar{Z}_\infty\bigl[1 - G_e(x)\bigr].$$

The rate at which customers depart from the servers is

$$\lim_{x\to 0} \frac{\bar{\mathcal{Z}}_\infty(C_0) - \bar{\mathcal{Z}}_\infty(C_x)}{x} = \bar{Z}_\infty \mu.$$

The arrival rate must be equal to the summation of the departure rate from the servers (due to service completion) and the one from the buffer (due to abandonment), i.e.

$$\lambda = \lambda F\left(\frac{\bar{R}_\infty}{\lambda}\right) + \bar{Z}_\infty \mu. \tag{4.11}$$

It follows directly from (4.2) that

$$\bar{Q}_\infty = \lambda F_d\left(\frac{\bar{R}_\infty}{\lambda}\right). \tag{4.12}$$

If $\bar{R}_\infty > 0$, then according to (4.12) we have $\bar{Q}_\infty > 0$. Thus $\bar{Z}_\infty = 1$ according to non-idling constraints. By (4.11), $\rho > 1$ and $\frac{\bar{R}_\infty}{\lambda}$ is a solution to the equation $F(w) = \frac{\rho-1}{\rho}$. If $\bar{R}_\infty = 0$, then according to (4.11) we have $\rho = \bar{Z}_\infty \le 1$. In summary, we have

$$\bar{Q}_\infty = \lambda F_d(w),$$
$$\bar{Z}_\infty = \min(\rho, 1),$$

where w is a solution to the equation $F(w) = \max(\frac{\rho-1}{\rho}, 0)$. This is consistent with the one in [31], which is derived from a conjecture of a fluid model. Now, we rigorously prove this result.

Proof of Theorem 3.2 If $(\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$ is an equilibrium state, then according to (3.1) and (3.2) and Definition 3.1, it must satisfy

$$\bar{\mathcal{R}}_\infty(C_x) = \lambda \int_{t-\frac{\bar{R}_\infty}{\lambda}}^{t} F^c(x+t-s)\,ds, \quad t \ge 0, \tag{4.13}$$

$$\bar{\mathcal{Z}}_\infty(C_x) = \bar{\mathcal{Z}}_\infty(C_x+t) + \int_0^t F^c\left(\frac{\bar{R}_\infty}{\lambda}\right) G^c(x+t-s)\,d\lambda s, \quad t \ge 0. \tag{4.14}$$

It follows from (4.14) that

$$\begin{aligned}\bar{\mathcal{Z}}_\infty(C_x) - \bar{\mathcal{Z}}_\infty(C_x+t) &= \rho F^c\left(\frac{\bar{R}_\infty}{\lambda}\right) \mu \int_0^t G^c(x+t-s)\,ds \\ &= \rho F^c\left(\frac{\bar{R}_\infty}{\lambda}\right)\big[G_e(x+t) - G_e(x)\big], \quad t \ge 0.\end{aligned}$$

Taking $t \to \infty$, one has

$$\bar{\mathcal{Z}}_\infty(C_x) = \rho F^c\left(\frac{\bar{R}_\infty}{\lambda}\right) G_e^c(x). \tag{4.15}$$

Thus $\bar{Z}_\infty = \rho F^c(\frac{\bar{R}_\infty}{\lambda})$. According to (4.2), we have

$$\bar{Q}_\infty = \lambda F_d\left(\frac{\bar{R}_\infty}{\lambda}\right).$$

First assume that $\bar{R}_\infty > 0$. Then $\bar{Q}_\infty > 0$, and thus $\bar{Z}_\infty = 1$ by the non-idling constraints (3.3) and (3.4). Therefore, $\rho F^c(\frac{\bar{R}_\infty}{\lambda}) = 1$, which implies that $F(\frac{\bar{R}_\infty}{\lambda}) = \frac{\rho-1}{\rho}$ and $\rho > 1$. Now assume that $\bar{R}_\infty = 0$. Then $\bar{Z}_\infty = \rho$, which must be less than or equal to 1 by the non-idling constraints. Summarizing the cases where $\rho > 1$ and $\rho \le 1$, we find that the equilibrium state must satisfy (3.13)–(3.15).

If a state $(\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$ satisfies (3.13)–(3.15), then let

$$\big(\bar{\mathcal{R}}(t), \bar{\mathcal{Z}}(t)\big) = (\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty),$$

for all $t \geq 0$. If $\rho \leq 1$, then $\bar{\mathcal{R}}(\cdot) \equiv \mathbf{0}$ and $\bar{Z}(\cdot) \equiv \rho$; if $\rho > 1$, then $\bar{R}(\cdot) \equiv \lambda w$ and $\bar{Z}(\cdot) \equiv 1$, where w is a solution to (3.15). It is easy to check that $(\bar{\mathcal{R}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ is a fluid model solution in both cases. So by definition, the state $(\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$ is a equilibrium state. □

5 Fluid approximation of the stochastic models

Similar to (2.3), let

$$B^n(t) = E^n(t) - R^n(t). \tag{5.1}$$

As explained in Sect. 2 that $B^n(t)$ is the index (by the order of arrival) of the head-of-the-line customer in the virtual buffer. The only way customers can leave the virtual buffer is when there is an available server, and they leave according to the order of the arrival. So the index $B^n(t)$ can only go up or stay as time t increases. We define its fluid scaling as $\bar{B}^n(t) = \frac{1}{n} B^n(t)$. For the convenience of notation we will need in the future, denote

$$\begin{aligned} \bar{E}^n(s,t) &= \bar{E}^n(t) - \bar{E}^n(s) \\ \bar{B}^n(s,t) &= \bar{B}^n(t) - \bar{B}^n(s) \end{aligned}$$

for any $0 \leq s \leq t$.

It follows from (2.4) and (2.5) that the dynamics for the fluid scaled processes can be written as

$$\bar{\mathcal{R}}^n(t)(C) = \frac{1}{n} \sum_{i=B^n(t)+1}^{E^n(t)} \delta_{u_i^n}\big(C + t - a_i^n\big), \quad \text{for all } C \in \mathscr{B}(\mathbb{R}), \tag{5.2}$$

$$\begin{aligned} \bar{\mathcal{Z}}^n(t)(C) &= \bar{\mathcal{Z}}^n(s)(C + t - s) \\ &\quad + \frac{1}{n} \sum_{i=B^n(s)+1}^{B^n(t)} \delta_{(u_i^n, v_i^n)}\big(C_0 + \tau_i^n - a_i^n\big) \times \big(C + t - \tau_i^n\big), \\ &\text{for all } C \in \mathscr{B}\big((0,\infty)\big), \end{aligned} \tag{5.3}$$

for all $0 \leq s \leq t$.

5.1 Precompactness

We first establish the following precompactness for the sequence of fluid scaled stochastic processes $\{(\bar{\mathcal{R}}^n(\cdot), \bar{\mathcal{Z}}^n(\cdot))\}$.

Theorem 5.1 *Assume* (3.17)–(3.20). *The sequence of the fluid scaled stochastic processes* $\{(\bar{\mathcal{R}}^n(\cdot), \bar{\mathcal{Z}}^n(\cdot))\}_{n\in\mathbb{N}}$ *is precompact in the space* $\mathbf{D}([0,\infty), \mathbf{M} \times \mathbf{M}_+)$; *namely,* *for each subsequence* $\{(\bar{\mathcal{R}}^{n_k}(\cdot), \bar{\mathcal{Z}}^{n_k}(\cdot))\}_{n_k}$ *with* $n_k \to \infty$, *there exists a further subsequence* $\{(\bar{\mathcal{R}}^{n_{k_j}}(\cdot), \bar{\mathcal{Z}}^{n_{k_j}}(\cdot))\}_{n_{k_j}}$ *such that*

$$\left(\bar{\mathcal{R}}^{n_{k_j}}(\cdot), \bar{\mathcal{Z}}^{n_{k_j}}(\cdot)\right) \quad \Rightarrow \quad \left(\tilde{\mathcal{R}}(\cdot), \tilde{\mathcal{Z}}(\cdot)\right) \quad \textit{as } j \to \infty,$$

for some $(\tilde{\mathcal{R}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \in \mathbf{D}([0,\infty), \mathbf{M} \times \mathbf{M}_+)$.

The remaining of this section is devoted to proving the above theorem. According to [10] that both $\mathbf{M}$ and $\mathbf{M}_+$ are separable and complete (thus so is the product space $\mathbf{M} \times \mathbf{M}_+$ as defined in Sect. 1.1) with the Prohorov metric $\mathbf{d}$ defined in Sect. 1.1. It follows from Theorem 3.5.6 in [7] that the space $\mathbf{D}([0,\infty), \mathbf{M} \times \mathbf{M}_+)$ is separable and complete since so is the space $\mathbf{M} \times \mathbf{M}_+$. By Theorem 3.7.2 in [7], it suffices to verify (a) the compact containment property, Lemma 5.1 and (b) the oscillation bound, Lemma 5.4 below.

5.1.1 Compact containment

A set $\mathbf{K} \subset \mathbf{M}$ is relatively compact if $\sup_{\xi \in \mathbf{K}} \xi(\mathbb{R}) < \infty$, and there exists a sequence of nested compact sets $A_j \subset \mathbb{R}$ such that $\bigcup A_j = \mathbb{R}$ and

$$\lim_{j\to\infty} \sup_{\xi \in \mathbf{K}} \xi\left(A_j^c\right) = 0,$$

where A_j^c denotes the complement of A_j; see [16], Theorem A7.5. The first major step to prove Theorem 5.1 is to establish the following *compact containment* property.

Lemma 5.1 *Assume* (3.17)–(3.20). *Fix* $T > 0$. *For each* $\eta > 0$ *there exists a compact set* $\mathbf{K} \subset \mathbf{M}$ *such that*

$$\liminf_{n\to\infty} \mathbb{P}^n\left(\left(\bar{\mathcal{R}}^n(t), \bar{\mathcal{Z}}^n(t)\right) \in \mathbf{K} \times \mathbf{K} \textit{ for all } t \in [0,T]\right) \ge 1 - \eta.$$

To prove this result, we first need to establish some bound estimations. It follows immediately from condition (3.17) that for each $\epsilon > 0$ there exists an n_0 such that when $n > n_0$,

$$\mathbb{P}^n\left(\sup_{0\le s<t\le T} \left|\bar{E}^n(s,t) - \lambda(t-s)\right| < \epsilon\right) \ge 1 - \epsilon. \tag{5.4}$$

To facilitate some arguments later on, we derive the following result from inequality (5.4) in the above.

Lemma 5.2 *Fix* $T > 0$. *There exists a function* $\epsilon_E(\cdot)$, *with* $\lim_{n\to\infty} \epsilon_E(n) = 0$ *such that*

$$\mathbb{P}^n\left(\sup_{0\le s<t\le T} \left|\bar{E}^n(s,t) - \lambda(t-s)\right| < \epsilon_E(n)\right) \ge 1 - \epsilon_E(n),$$

for each $n \ge 0$.

The derivation of the above lemma from (5.4) follows the same as the proof of Lemma 5.1 in [33]. We omit the proof for brevity. Based on the above lemma, we construct the following event:

$$\Omega_E^n = \Big\{ \sup_{t\in[0,T]} \big|\bar{E}^n(s,t) - \lambda(t-s)\big| < \epsilon_E(n) \Big\}. \tag{5.5}$$

We see that on this event, the arrival process is regular, i.e. $\bar{E}^n(s,t)$ is "close" to $\lambda(t-s)$. This event has "large" probability, i.e.

$$\lim_{n\to\infty} \mathbb{P}^n\big(\Omega_E^n\big) = 1. \tag{5.6}$$

Proof of Lemma 5.1 By the convergence of the initial condition (3.19), for any $\epsilon > 0$, there exists a relatively compact set $\mathbf{K}_0 \subset \mathbf{M}$ such that

$$\liminf_{n\to\infty} \mathbb{P}^n\big(\bar{\mathcal{R}}^n(0) \in \mathbf{K}_0 \text{ and } \bar{\mathcal{Z}}^n(0) \in \mathbf{K}_0\big) > 1 - \epsilon. \tag{5.7}$$

Denote the event in the above probability by Ω_0^n. On this event, by the definition of relatively compact set in the space $\mathbf{M}$, there exists a function $\kappa_0(\cdot)$ with $\lim_{x\to\infty} \kappa_0(x) = 0$ such that

$$\bar{\mathcal{R}}^n(0)(C_x) \le \kappa_0(x), \qquad \bar{\mathcal{Z}}^n(0)(C_x) \le \kappa_0(x), \tag{5.8}$$

and

$$\bar{\mathcal{R}}^n(0)\big(C_x^-\big) \le \kappa_0(x), \tag{5.9}$$

for all $x \ge 0$, where $C_x^- = (-\infty, -x)$ for any $y \in \mathbb{R}$. (Remember that $\bar{\mathcal{Z}}^n(0)$ is a measure on $(0,\infty)$, so we do not need to consider its measure of C_x^-.) It is clear that on the event $\Omega_E^n \cap \Omega_0^n$, for any $t \le T$ and all large n,

$$\bar{\mathcal{R}}^n(t)(\mathbb{R}) \le \sup_n \bar{\mathcal{R}}^n(0)(\mathbb{R}) + 2\lambda T,$$

$$\bar{\mathcal{Z}}^n(t)\big((0,\infty)\big) \le 1,$$

where the last inequality is due to the fact that $Z^n(\cdot) \le n$. Again, by the definition of relative compact set in $\mathbf{M}$, we have $\sup_n \bar{\mathcal{R}}^n(0)(\mathbb{R}) = M_0 < \infty$. It follows from the dynamic equations (5.2) and (5.3) that for all $x > 0$,

$$\bar{\mathcal{R}}^n(t)(C_x) \le \bar{\mathcal{R}}^n(0)(C_x) + \frac{1}{n}\sum_{i=1}^{E^n(t)} \delta_{u_i^n}(C_x),$$

$$\bar{\mathcal{Z}}^n(t)(C_x) \le \bar{\mathcal{Z}}^n(0)(C_x) + \frac{1}{n}\sum_{i=B^n(0)+1}^{E^n(t)} \delta_{v_i^n}(C_x).$$

Denote $\bar{\mathcal{L}}_1^n(t) = \frac{1}{n}\sum_{i=1}^{E^n(t)} \delta_{u_i^n}$ and $\bar{\mathcal{L}}_2^n(t) = \frac{1}{n}\sum_{i=B^n(0)+1}^{E^n(t)} \delta_{v_i^n}$. Let us first study these two terms. Since the probability measure ν_F^n of u_i^n's converges to ν_F, and the probability measure ν_G^n of v_i^n's converges to ν_G (by condition (3.18)), we can define $\bar{f}$ and

$\bar{f}_2$ by (B.5) and (B.6). By the definition of Ω_0^n there exists a constant $M_{R,0}$ such that $|B^n(0)/n| \le M_{R,0}$ on Ω_0^n. Recall the definition of the event $\Omega_{\mathrm{GC}}^n(M,L)$ in (B.10). For the application here, it is enough to set $M = M_{R,0}$ and $L = 2\lambda T + M$. On the event $\Omega_E^n \cap \Omega_{\mathrm{GC}}^n(M,L)$, we have

$$\langle \bar{f}, \bar{\mathcal{L}}_1^n(t)\rangle \le \left\langle \bar{f}, \frac{1}{n}\sum_{i=-\lfloor Mn\rfloor}^{\lfloor 2\lambda Tn\rfloor} \delta_{u_i^n}\right\rangle \le (2\lambda T + M)\langle \bar{f}, \nu_F\rangle + 1,$$

for all large enough n. Similarly, on the same event we have

$$\langle \bar{f}, \bar{\mathcal{L}}_2^n(t)\rangle \le \left\langle \bar{f}, \frac{1}{n}\sum_{i=-\lfloor Mn\rfloor}^{\lfloor 2\lambda Tn\rfloor} \delta_{v_i^n}\right\rangle \le (2\lambda T + M)\langle \bar{f}, \nu_G\rangle + 1,$$

for all large enough n. Denote $M_b = (2\lambda T + M)\max(\langle \bar{f}, \nu_F\rangle, \langle \bar{f}, \nu_G\rangle) + 1$. By Markov's inequality, for all $x > 0$ (again, on the same event and for all large n)

$$\bar{\mathcal{L}}_1^n(t)(C_x) < M_b/\bar{f}(x), \qquad \bar{\mathcal{L}}_2^n(t)(C_x) < M_b/\bar{f}(x),$$

where the upper bound vanishes as $x \to \infty$ by (B.7). Unlike the measure $\mathcal{Z}(t) \in \mathbf{M}_+$, the measure $\mathcal{R}(t) \in \mathbf{M}$. So we need to consider all the test set $C_x^- = (-\infty, -x)$ for $x \ge 0$. The following inequality again follows from (5.2):

$$\bar{\mathcal{R}}^n(t)\big(C_x^-\big) \le \bar{\mathcal{R}}^n(0)\big(C_x^- + t\big) + \frac{1}{n}\sum_{i=1}^{E^n(t)} \delta_{u_i^n}\big(C_x^- + t\big).$$

Note that if we take $x > T$, then $\delta_{u_i^n}(C_x^- + t) = 0$. So we have

$$\bar{\mathcal{R}}^n(t)\big(C_x^-\big) \le \bar{\mathcal{R}}^n(0)\big(C_x^- + T\big) = \bar{\mathcal{R}}^n(0)\big(C_{x-T}^-\big), \quad \text{for all } t \le T. \tag{5.10}$$

Now, define the set $\mathbf{K} \subset \mathbf{M}$ by

$$\begin{aligned}\mathbf{K} = \big\{\xi \in \mathbf{M} : {}& \xi(\mathbb{R}) < 1 + M_0 + 2\lambda T,\\ & \xi(C_x) < \kappa_0(x) + M_b/\bar{f}(x) \text{ for all } x > 0,\\ & \xi\big(C_x^-\big) \le \kappa_0(x - T) \text{ for all } x \ge T\big\}.\end{aligned}$$

It is clear that $\mathbf{K}$ is relatively compact and on the event $\Omega_E^n \cap \Omega_{\mathrm{GC}}^n(M,L) \cap \Omega_0^n$,

$$\big(\bar{\mathcal{R}}^n(t), \bar{\mathcal{Z}}^n(t)\big) \in \mathbf{K} \times \mathbf{K} \quad \text{for all } t \in [0,T].$$

The result of this lemma then follows immediately from (5.6), (5.7), and (B.11). □

5.1.2 Oscillation bound

The second major step to prove precompactness is to obtain the oscillation bound in Lemma 5.4 below. The oscillation of a *càdlàg* function $\zeta(\cdot)$ (taking values in a metric space $(\mathbf{E},\pi)$) on a fixed interval $[0,T]$ is defined as

$$\mathbf{w}_T\bigl(\zeta(\cdot),\delta\bigr)=\sup_{s,t\in[0,T],|s-t|<\delta}\pi\bigl[\zeta(s),\zeta(t)\bigr].$$

If the metric space is $\mathbb{R}$, we just use the Euclidean metric; if the space is $\mathbf{M}$ or $\mathbf{M}_+$, we use the Prohorov metric $\mathbf{d}$ defined in Sect. 1.1. For the measure-valued processes in our model, oscillations mainly result from sudden departures of a large number of customers. To control the departure process, we show that $\bar{\mathcal{Z}}^n(\cdot)$ and $\bar{\mathcal{R}}^n(\cdot)$ assign arbitrarily small mass to small intervals.

Lemma 5.3 *Assume* (3.10), (3.17)–(3.20). *Fix* $T>0$. *For each* $\epsilon,\eta>0$ *there exists a* $\kappa>0$ (*depending on* ϵ *and* η) *such that*

$$\liminf_{n\to\infty}\mathbb{P}^n\Bigl(\sup_{t\in[0,T]}\sup_{x\in\mathbb{R}_+}\bar{\mathcal{Z}}^n(t)\bigl([x,x+\kappa]\bigr)\le\epsilon\Bigr)\ge 1-\eta. \tag{5.11}$$

Proof First, We see that for any $\epsilon,\eta>0$, there exists a κ such that

$$\liminf_{n\to\infty}\mathbb{P}^n\Bigl(\sup_{x\in\mathbb{R}_+}\bar{\mathcal{Z}}^n(0)\bigl([x,x+\kappa]\bigr)\le\epsilon/2\Bigr)\ge 1-\eta. \tag{5.12}$$

This inequality is derived from the initial condition. The derivation is exactly the same as in the proof of (5.14) in [33], so we omit it here for brevity.

Now we need to extend this result to the interval $[0,T]$. Denote the event in (5.12) by Ω^n_{0s}, and the event in Lemma 5.1 by $\Omega^n_C(\mathbf{K})$. Fix $M=1$ and $L=2\lambda T$. Let

$$\Omega^n_1(M,L)=\Omega^n_{0s}\cap\Omega^n_C(\mathbf{K})\cap\Omega^n_E\cap\Omega^n_{\mathrm{GC}}(M,L). \tag{5.13}$$

By (5.12), Lemma 5.1, (5.6), and (B.11), for any fixed $M,L>0$,

$$\liminf_{n\to\infty}\mathbb{P}^n\bigl(\Omega^n_1(M,L)\bigr)\ge 1-\eta.$$

In the remainder of the proof, all random objects are evaluated at a fixed sample path in $\Omega^n_1(M,L)$.

It follows from the fluid scaled stochastic dynamic equation (5.3) that

$$\bar{\mathcal{Z}}^n(t)\bigl([x,x+\kappa]\bigr)\le\bar{\mathcal{Z}}^n(0)\bigl([x,x+\kappa]+t\bigr)+\frac{1}{n}\sum_{i=B^n(0)+1}^{B^n(t)}\delta_{v_i^n}\bigl([x,x+\kappa]+t-\tau_i^n\bigr),$$

for each $x,\kappa\in\mathbb{R}_+$. By (5.12), the first term on the right hand side of the above equation is always upper bounded by $\epsilon/2$. Let S denote the second term on the right hand side of the preceding equation. Now it only remains to show that $S<\epsilon/2$.

Let $0=t_0<t_1<\cdots<t_J=t$ be a partition of the interval $[0,t]$ such that $|t_{j+1}-t_j|<\delta$ for all $j=0,\ldots,J-1$, where δ and J are to be chosen below. Write S

as the summation

$$S = \sum_{j=0}^{J-1} \frac{1}{n} \sum_{i=B^n(t_j)+1}^{B^n(t_{j+1})} \delta_{v_i^n}\bigl([x, x+\kappa] + t - \tau_i^n\bigr).$$

Recall that τ_i^n is the time that the ith job starts service, so on each sub-interval $[t_j, t_{j+1}]$ those i's to be summed must satisfy $t_j \le \tau_i^n \le t_{j+1}$. This implies that

$$t - t_{j+1} \le t - \tau_i^n \le t - t_j.$$

Then

$$S \le \sum_{j=0}^{J-1} \frac{1}{n} \sum_{i=B^n(t_j)+1}^{B^n(t_{j+1})} \delta_{v_i^n}\bigl([x + t - t_{j+1}, x + t - t_j + \kappa]\bigr).$$

By (5.1), we have for all $j = 0, \ldots, J$

$$-\bar{R}^n(0) \le \bar{B}^n(t_j) \le \bar{E}^n(T).$$

Thus,

$$0 \le \bar{B}^n(t_J) - \bar{B}^n(t_0) \le \bar{E}^n(T) + \bar{R}^n(0).$$

By Lemmas 5.1 and 5.2, $\bar{R}^n(0) < M_0$ and $\bar{E}^n(T) \le 2\lambda T$ on $\Omega_C^n(\mathbf{K}) \cap \Omega_E^n$ for some constant M_0. Take $M = \max(M_0, 2\lambda T)$ and $L = M_0 + 2\lambda T$, it follows from the Glivenko–Cantelli estimate (B.10) that

$$\begin{aligned}
&\frac{1}{n} \sum_{i=B^n(t_j)+1}^{B^n(t_{j+1})} \delta_{v_i^n}\bigl([x + t - t_{j+1}, x + t - t_j + \kappa]\bigr) \\
&\quad \le \bigl(\bar{B}^n(t_{j+1}) - \bar{B}^n(t_j)\bigr) \nu_G^n\bigl([x + t - t_{j+1}, x + t - t_j + \kappa]\bigr) + \frac{\epsilon}{4J},
\end{aligned} \tag{5.14}$$

for each $j < J$. By condition (3.18), for any $\epsilon_2 > 0$,

$$\mathbf{d}\bigl[\nu_G^n, \nu_G\bigr] < \epsilon_2,$$

for all large n. By the definition of Prohorov metric, we have

$$\nu_G^n\bigl([x+t-t_{j+1}, x+t-t_j+\kappa]\bigr) \le \nu_G\bigl([x+t-t_{j+1}-\epsilon_2, x+t-t_j+\kappa+\epsilon_2]\bigr) + \epsilon_2,$$

for all large n. Since $[x + t - t_{j+1} - \epsilon_2, x + t - t_j + \kappa + \epsilon_2]$ is a close interval with length less than $\kappa + \delta + 2\epsilon_2$, by condition (3.10) and the fact that ν_G is a probability measure, we can choose $\kappa, \delta, \epsilon_2$ small enough such that

$$\nu_G\bigl([x + t - t_{j+1} - \epsilon_2, x + t - t_j + \kappa + \epsilon_2]\bigr) + \epsilon_2 \le \frac{\epsilon}{4M}. \tag{5.15}$$

Note that by making δ small, we need to choose $J \geq \lceil t/\delta \rceil$. It then follows from (5.14) that

$$\begin{aligned} S &\leq \frac{\epsilon}{4M} \sum_{j=0}^{J-1} \big[\bar{B}^n(t_{j+1}) - \bar{B}^n(t_j) \big] + \frac{\epsilon}{4} \\ &\leq \frac{\epsilon}{4M} \big[\bar{B}^n(t_J) - \bar{B}^n(t_0) \big] + \frac{\epsilon}{4} \\ &\leq \epsilon/2. \end{aligned}$$

This completes the proof. □

Lemma 5.4 *Assume* (3.10), (3.17)–(3.20). *Fix* $T > 0$. *For each* $\epsilon, \eta > 0$ *there exists a* $\delta > 0$ (*depending on* ϵ *and* η) *such that*

$$\liminf_{n\to\infty} \mathbb{P}^n \big(\mathbf{w}_T \big((\bar{\mathcal{R}}^n, \bar{\mathcal{Z}}^n)(\cdot), \delta \big) \leq 3\epsilon \big) \geq 1 - \eta. \tag{5.16}$$

Proof Define

$$\Omega^n_{\text{Reg}}(\epsilon, \kappa) = \Big\{ \sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} \bar{\mathcal{Z}}^n(t)\big([x, x+\kappa]\big) \leq \epsilon \Big\}.$$

By (5.6) and Lemma 5.3, for each $\epsilon, \eta > 0$ there exists a $\kappa > 0$ such that

$$\liminf_{n\to\infty} \mathbb{P}^n \big(\Omega^n_E \cap \Omega^n_{\text{Reg}}(\epsilon, \kappa) \big) > 1 - \eta. \tag{5.17}$$

On the event $\Omega^n_E \cap \Omega^n_{\text{Reg}}(\epsilon, \kappa)$, we have some control over the dynamics of the system. First, when $t - s \leq \min(\frac{\epsilon}{2\lambda}, \kappa)$, by the definition of Ω^n_E and $\Omega^n_{\text{Reg}}(\epsilon, \kappa)$, we have

$$\bar{E}^n(s, t) \leq \epsilon \tag{5.18}$$

Second, by the dynamic equation (5.2), for any $s < t$ and any set $C \in \mathscr{B}(\mathbb{R})$,

$$\begin{aligned} &\bar{\mathcal{R}}^n(t)(C) - \bar{\mathcal{R}}^n(s)\big(C^{3\epsilon}\big) \\ &\quad \leq -\frac{1}{n} \sum_{i=B^n(s)+1}^{B^n(t)} \delta_{u_i^n}\big(C^{3\epsilon} + sa_i^n\big) + \bar{E}^n(s,t) \\ &\qquad + \frac{1}{n} \sum_{i=B^n(t)+1}^{E^n(s)} \big[\delta_{u_i^n}\big(C + t - a_i^n\big) - \delta_{u_i^n}\big(C^{3\epsilon} + s - a_i^n\big) \big], \end{aligned} \tag{5.19}$$

where C^a is the a-enlargement of the set C as defined in Sect. 1.1. The first term on the right hand side is clearly non-positive. Note that when $t - s \leq \epsilon$, $C + t - a_i^n \subseteq C^\epsilon + s - a_i^n$ for all $i \in \mathbb{Z}$, which implies that the third term in the above inequality is less than zero. It follows from (5.18) that

$$\bar{\mathcal{R}}^n(t)(C) - \bar{\mathcal{R}}^n(s)\big(C^\epsilon\big) \leq \epsilon.$$

By Property (ii) on page 72 in [2], we have

$$\mathbf{d}\big[\bar{\mathcal{R}}^n(t), \bar{\mathcal{R}}^n(s)\big] \le \epsilon. \tag{5.20}$$

Finally, by the dynamic equation (5.3),

$$\bar{\mathcal{Z}}^n(t)(C) \le \bar{\mathcal{Z}}^n(s)(C + t - s) + \bar{B}^n(s,t).$$

Plug in $C = \mathbb{R}$ into (5.19), we have

$$\bar{B}^n(s,t) \le \big|\bar{\mathcal{R}}^n(t)(\mathbb{R}) - \bar{\mathcal{R}}^n(s)(\mathbb{R})\big| + \bar{E}^n(s,t) \le 2\epsilon,$$

by (5.18) and (5.20). Note that when $t - s \le 2\epsilon$, $C + t - s \subseteq C^{2\epsilon}$, where C^a is the a-enlargement of the set C as defined in Sect. 1.1. Thus, we have

$$\bar{\mathcal{Z}}^n(t)(C) \le \bar{\mathcal{Z}}^n(s)\big(C^{2\epsilon}\big) + 2\epsilon.$$

By Property (ii) on page 72 in [2], we have

$$\mathbf{d}\big[\bar{\mathcal{Z}}^n(s), \bar{\mathcal{Z}}^n(t)\big] \le 2\epsilon. \tag{5.21}$$

The result of this lemma follows immediately from (5.17), (5.20), and (5.21). □

5.2 Convergence to the fluid model solution

We have established the precompactness in Theorem 5.1. So every subsequence of the fluid scaled processes has a further subsequence which converges to some limit. For simplicity of notations, we index the convergent subsequence again by n. So we have

$$\big(\bar{\mathcal{R}}^n(\cdot), \bar{\mathcal{Z}}^n(\cdot)\big) \quad \Rightarrow \quad \big(\tilde{\mathcal{R}}(\cdot), \tilde{\mathcal{Z}}(\cdot)\big) \quad \text{as } n \to \infty. \tag{5.22}$$

By the oscillation bound in Lemma 5.4, the limit $(\tilde{\mathcal{R}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ is almost surely continuous. We have the following result, which further characterizes the above limit.

Lemma 5.5 *Assume* (3.10)–(3.12) *and* (3.17)–(3.20). *The limit* $(\tilde{\mathcal{R}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ *in* (5.22) *is almost surely the solution to the fluid model* (λ, F, G) *with initial condition* $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$.

The rest of this section is devoted to characterizing the limits. To better structure the proof, we first provide some preliminary estimates based on the dynamic equations (5.2) and (5.3).

Lemma 5.6 *Let* $\{t_j\}_{j=0}^J$ *be a partition of the interval* $[s,t]$ *such that* $s = t_0 < t_1 < \cdots < t_J = t$. *We have for any* $x \in \mathbb{R}$,

$$\bar{\mathcal{R}}^n(t)(C_x) \le \sum_{j=0}^{J-1} \frac{1}{n} \sum_{i=1+E^n(t_j)}^{E^n(t_{j+1})} \delta_{u_i^n}(C_x + t - t_{j+1}) + \big|\bar{E}^n(s) - \bar{B}^n(t)\big|, \tag{5.23}$$

$$\bar{\mathcal{R}}^n(t)(C_x) \geq \sum_{j=0}^{J-1} \frac{1}{n} \sum_{i=1+E^n(t_j)}^{E^n(t_{j+1})} \delta_{u_i^n}(C_x + t - t_j) - \left|\bar{E}^n(s) - \bar{B}^n(t)\right|. \tag{5.24}$$

On the event where $\sup_{\tau\in[s,t]} |\bar{E}^n(\tau) - \lambda\tau| < \epsilon$, *then for any* $x > 0$,

$$\begin{aligned} \bar{\mathcal{Z}}^n(t)(C_x) &\leq \bar{\mathcal{Z}}^n(s)(C_x + t - s) \\ &\quad + \sum_{j=0}^{J-1} \frac{1}{n} \sum_{i=1+B^n(t_j)}^{B^n(t_{j+1})} \delta_{u_i^n}\left(C_0 + \frac{\bar{R}^n_{L,j} - 2\epsilon}{\lambda}\right) \delta_{v_i^n}(C_x + t - t_{j+1}), \end{aligned} \tag{5.25}$$

$$\begin{aligned} \bar{\mathcal{Z}}^n(t)(C_x) &\geq \bar{\mathcal{Z}}^n(s)(C_x + t - s) \\ &\quad + \sum_{j=0}^{J-1} \frac{1}{n} \sum_{i=1+B^n(t_j)}^{B^n(t_{j+1})} \delta_{u_i^n}\left(C_0 + \frac{\bar{R}^n_{U,j} + 2\epsilon}{\lambda}\right) \delta_{v_i^n}(C_x + t - t_j), \end{aligned} \tag{5.26}$$

where $\bar{R}^n_{L,j} = \inf_{t\in[t_j,t_{j+1}]} \bar{R}^n(t)$ *and* $\bar{R}^n_{U,j} = \sup_{t\in[t_j,t_{j+1}]} \bar{R}^n(t)$.

Proof Note that $0 \leq \delta_{u_i^n}(C) \leq 1$ for any Borel set C and any random variable u_i^n. So by the dynamic equation (5.2), we have

$$\left|\bar{\mathcal{R}}^n(t)(C) - \frac{1}{n} \sum_{i=E^n(s)+1}^{E^n(t)} \delta_{u_i^n}\left(C + t - a_i^n\right)\right| \leq \left|\bar{E}^n(s) - \bar{B}^n(t)\right|.$$

For those i's such that $E^n(t_j) < i \leq E^n(t_{j+1})$, we have

$$t_j < a_i^n \leq t_{j+1}. \tag{5.27}$$

This implies that $C_x + t - a_i \subseteq C_x + t - t_{j+1}$. So we have

$$\sum_{i=1+E^n(t_j)}^{E^n(t_{j+1})} \delta_{u_i^n}(C_x + t - a_i) \leq \sum_{i=1+E^n(t_j)}^{E^n(t_{j+1})} \delta_{u_i^n}(C_x + t - t_{j+1}).$$

This establishes (5.23). Also, (5.27) implies $C_x + t - t_j \subseteq C_x + t - a_i$. So (5.24) follows in the same way.

For those i's such that $B^n(t_j) < i \leq B^n(t_{j+1})$, we have

$$t_j < \tau_i^n \leq t_{j+1}.$$

Note that $\bar{R}^n(\tau_i^n) = \bar{E}^n(\tau_i^n) - \bar{E}^n(a_i^n)$ for each i. So, by the closeness between $\bar{E}^n(\cdot)$ and $\lambda\cdot$, we have

$$\begin{aligned} &\left|\bar{R}^n\left(\tau_i^n\right) - \lambda\left(\tau_i^n - a_i^n\right)\right| \\ &\quad \leq \left|\bar{R}^n\left(\tau_i^n\right) - \bar{E}^n\left(\tau_i^n\right) + \bar{E}^n\left(a_i^n\right)\right| + \left|\bar{E}^n\left(\tau_i^n\right) - \bar{E}^n\left(a_i^n\right) - \lambda\left(\tau_i^n - a_i^n\right)\right| \\ &\quad \leq 2\epsilon. \end{aligned}$$

So

$$\bar{R}^n_{L,j} - 2\epsilon \le \lambda\big(\tau^n_i - a^n_i\big) \le \bar{R}^n_{U,j} + 2\epsilon,$$

for all i's such that $B^n(t_j) < i \le B^n(t_{j+1})$. Thus,

$$\begin{aligned}
&\sum_{i=1+B^n(t_j)}^{B^n(t_{j+1})} \delta_{u^n_i}\big(C_0 + \tau^n_i - a^n_i\big)\delta_{v^n_i}\big(C_x + t - \tau^n_i\big) \\
&\quad \le \sum_{i=1+B^n(t_j)}^{B^n(t_{j+1})} \delta_{u^n_i}\left(C_0 + \frac{\bar{R}^n_{L,j} - 2\epsilon}{\lambda}\right)\delta_{v^n_i}(C_x + t - t_{j+1}).
\end{aligned}$$

This implies (5.25). Also (5.26) can be proved in the same way. □

Recall the notations $\bar{\mathcal{L}}^n(m,l)$, $\bar{\mathcal{L}}^n_F(m,l)$ and $\bar{\mathcal{L}}^n_G(m,l)$ are defined in (B.1)–(B.3) in the appendix. Using these notations, Lemma 5.6 can be written as the follows.

Lemma 5.7 *Let $\{t_j\}_{j=0}^J$ be a partition of the interval $[s,t]$ such that $s = t_0 < t_1 < \cdots < t_J = t$. We have for any $x \in \mathbb{R}$,*

$$\bar{\mathcal{R}}^n(t)(C_x) \le \sum_{j=0}^{J-1}\big\langle 1_{(C_x+t-t_{j+1})}, \bar{\mathcal{L}}^n_F\big(E^n(t_j), \bar{E}^n(t_j,t_{j+1})\big)\big\rangle + \big|\bar{E}^n(s) - \bar{B}^n(t)\big|, \tag{5.28}$$

$$\bar{\mathcal{R}}^n(t)(C_x) \ge \sum_{j=0}^{J-1}\big\langle 1_{(C_x+t-t_j)}, \bar{\mathcal{L}}^n_F\big(E^n(t_j), \bar{E}^n(t_j,t_{j+1})\big)\big\rangle - \big|\bar{E}^n(s) - \bar{B}^n(t)\big|. \tag{5.29}$$

If in addition $\sup_{\tau\in[s,t]}|\bar{E}^n(\tau) - \lambda\tau| < \epsilon$, *then for any* $x > 0$,

$$\begin{aligned}
\bar{\mathcal{Z}}^n(t)(C_x) &\le \bar{\mathcal{Z}}^n(s)(C_x + t - s) \\
&\quad + \sum_{j=0}^{J-1}\big\langle 1_{(C_0+\frac{\bar{R}^n_{L,j}-2\epsilon}{\lambda})\times(C_x+t-t_{j+1})}, \bar{\mathcal{L}}^n\big(B^n(t_j), \bar{B}^n(t_j,t_{j+1})\big)\big\rangle,
\end{aligned} \tag{5.30}$$

$$\begin{aligned}
\bar{\mathcal{Z}}^n(t)(C_x) &\ge \bar{\mathcal{Z}}^n(s)(C_x + t - s) \\
&\quad + \sum_{j=0}^{J-1}\big\langle 1_{(C_0+\frac{\bar{R}^n_{U,j}+2\epsilon}{\lambda})\times(C_x+t-t_j)}, \bar{\mathcal{L}}^n\big(B^n(t_j), \bar{B}^n(t_j,t_{j+1})\big)\big\rangle.
\end{aligned} \tag{5.31}$$

Fix a constant $T > 0$ and let $M = 1$ and $L = 2\lambda T$. Denote the random variable

$$\begin{aligned}
\bar{V}^n_{M,L} = \max_{-nM<m<nM}\sup_{l\in[0,L]}\sup_{x,y\in\mathbb{R}}\Big\{&\big|\bar{\mathcal{L}}^n(m,l)(C_x \times C_y) - l\nu^n_F(C_x)\nu^n_G(C_y)\big| \\
&+ \big|\bar{\mathcal{L}}^n_F(m,l)(C_x) - l\nu^n_F(C_x)\big| + \big|\bar{\mathcal{L}}^n_G(m,l)(C_x) - l\nu^n_G(C_x)\big|\Big\}.
\end{aligned} \tag{5.32}$$

By Lemma B.1, for any fixed constants $M, L > 0$,

$$\bar{V}^n_{M,L} \quad \Rightarrow \quad 0 \quad \text{as } n \to \infty.$$

By the assumption (3.17), we have

$$\bar{E}^n(\cdot) \quad \Rightarrow \quad \lambda\,\cdot \quad \text{as } n \to \infty.$$

Since both the above two limits are deterministic, those convergences are joint with the convergence of $(\bar{\mathcal{R}}^n(\cdot), \bar{\mathcal{Z}}^n(\cdot))$. Now, for each $n \geq 1$, we can view $(\bar{E}^n(\cdot), \bar{\mathcal{R}}^n(\cdot), \bar{\mathcal{Z}}^n(\cdot), V_{M,L})$ as a random variable in the space $\mathbf{E}_1$, which is the product space of three $\mathbf{D}([0,\infty), \mathbb{R})$ spaces and the space $\mathbb{R}$. $(\bar{\mathcal{L}}^n(m,\cdot), \bar{\mathcal{L}}^n_F(m,\cdot), \bar{\mathcal{L}}^n_G(m,\cdot) : m \in \mathbb{Z})$ in the product space $\mathbf{E}_2$ of countably many $\mathbf{D}([0,\infty), \mathbf{M})$ spaces. It is clear that both $\mathbf{E}_1$ and $\mathbf{E}_2$ are complete and separable metric spaces. Using the extension of the Skorohod representation Theorem, Lemma C.1, we assume without loss of generality that $\bar{E}^n(\cdot), \bar{\mathcal{R}}^n(\cdot), \bar{\mathcal{Z}}^n(\cdot), \bar{V}^n_{M,L}, \bar{\mathcal{L}}^n(m,\cdot), \bar{\mathcal{L}}^n_F(m,\cdot), \bar{\mathcal{L}}^n_G(m,\cdot), m \in \mathbb{Z}$, and $(\tilde{\mathcal{R}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ are defined on a common probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ such that, almost surely,

$$\bigl(\bigl(\bar{\mathcal{R}}^n(\cdot), \bar{\mathcal{Z}}^n(\cdot)\bigr), \bar{V}^n_{M,L}, \bar{E}^n(\cdot)\bigr) \to \bigl(\bigl(\tilde{\mathcal{R}}(\cdot), \tilde{\mathcal{Z}}(\cdot)\bigr), 0, \lambda\cdot\bigr) \quad \text{as } n \to \infty, \tag{5.33}$$

and inequalities (5.28)–(5.31) and (5.32) also hold almost surely. Note that the convergence of each function component in the above is in the Skorohod J_1 topology. Since the limit is continuous, the convergence is equivalent to the convergence in the uniform norm on compact intervals. Thus as $n \to \infty$,

$$\sup_{t\in[0,T]} \mathbf{d}\bigl[\bar{\mathcal{R}}^n(t), \tilde{\mathcal{R}}(t)\bigr] \to 0, \tag{5.34}$$

$$\sup_{t\in[0,T]} \mathbf{d}\bigl[\bar{\mathcal{Z}}^n(t), \tilde{\mathcal{Z}}(t)\bigr] \to 0, \tag{5.35}$$

$$\sup_{t\in[0,T]} \bigl|\bar{E}^n(t) - \lambda t\bigr| \to 0, \tag{5.36}$$

where $\mathbf{d}$ is the Prohorov metric defined in Sect. 1.1. In the same way as on the original probability space, let

$$\bar{R}^n(\cdot) = \bigl\langle 1, \bar{\mathcal{R}}^n(\cdot)\bigr\rangle, \qquad \bar{Q}^n(\cdot) = \bigl\langle 1_{(0,\infty)}, \bar{\mathcal{R}}^n(\cdot)\bigr\rangle,$$
$$\bar{Z}^n(\cdot) = \bigl\langle 1, \bar{\mathcal{Z}}^n(\cdot)\bigr\rangle, \qquad \bar{X}^n(\cdot) = \bar{Q}^n(\cdot) + \bar{Z}^n(\cdot),$$

and

$$\bar{B}^n(\cdot) = \bar{E}^n(\cdot) - \bar{R}^n(\cdot).$$

According to (5.34) and (5.36), we have

$$\sup_{t\in[0,T]} \bigl|\bar{B}^n(t) - \tilde{B}(t)\bigr| \to 0. \tag{5.37}$$

For each n, let $\tilde{\Omega}_{n,2}$ be an event of probability one on which the stochastic dynamic equations (5.2) and (5.3) and the non-idling constraints (2.7) and (2.8) hold. Define $\tilde{\Omega}_0 = \tilde{\Omega}_1 \cap (\bigcap_{n=0}^{\infty} \tilde{\Omega}^n_{n,2})$, where $\tilde{\Omega}_1$ is the event of probability one on which (5.33) holds. Then $\tilde{\Omega}_0$ also has probability one. Based on Lemma 5.6 and the above argument using Skorohod Representation theorem, we can now prove Lemma 5.5.

Proof of Lemma 5.5 For any $t \geq 0$, fix a constant $T > t$. Let us now study $(\tilde{\mathcal{R}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ on the time interval $[0, T]$. It is enough to show that on the event $\tilde{\Omega}_0$, $(\tilde{\mathcal{R}}(t), \tilde{\mathcal{Z}}(t))$ satisfies the fluid model equation (3.1)–(3.2) and the constraints (3.3)–(3.4). Assume for the remainder of this proof that all random objects are evaluated at a sample path in the event $\tilde{\Omega}_0$.

We first verify (3.1). For any $\epsilon > 0$, consider the difference

$$\begin{aligned}\tilde{\mathcal{R}}(t)(C_x) &- \int_{t-\frac{\tilde{R}(t)}{\lambda}}^{t} F^c(x+t-s)\,d\lambda s\\ &= \tilde{\mathcal{R}}(t)(C_x) - \bar{\mathcal{R}}^n(t)\big(C^{\epsilon}_x\big) + \bar{\mathcal{R}}^n(t)\big(C^{\epsilon}_x\big) - \int_{t-\frac{\tilde{R}(t)}{\lambda}}^{t} F^c(x+t-s)\,d\lambda s,\end{aligned}$$

where C^{ϵ}_x is the ϵ-enlargement of the set C_x as defined in Sect. 1.1, which is essentially $C_{x-\epsilon}$. Let $t_0 = t - \tilde{R}(t)/\lambda$. According to (5.28), we have

$$\begin{aligned}\tilde{\mathcal{R}}(t)(C_x) &- \int_{t-\frac{\tilde{R}(t)}{\lambda}}^{t} F^c(x+t-s)\,d\lambda s\\ &\leq \tilde{\mathcal{R}}(t)(C_x) - \bar{\mathcal{R}}^n(t)\big(C^{\epsilon}_x\big) + \big|\bar{E}^n(t_0) - \bar{B}^n(t)\big|\\ &\quad \times \sum_{j=0}^{J-1}\big\langle 1_{(C^{\epsilon}_x+t-t_{j+1})}, \bar{\mathcal{L}}^n_F\big(E^n(t_j), \bar{E}^n(t_j, t_{j+1})\big)\big\rangle\\ &\quad - \int_{t_0}^{t} F^c(x+t-s)\,d\lambda s,\end{aligned} \tag{5.38}$$

where $\{t_j\}_{j=0}^J$ is a partition of the interval $[t_0, t]$ such that $t_0 < t_1 < \cdots < t_J = t$ and $\max_j(t_{j+1} - t_j) < \delta$ for some $\delta > 0$. By the definition of Prohorov metric and the convergence in (5.34), the first term on the right hand side of (5.38) is bounded by ϵ for all large n. By (5.34) and (5.36)

$$\begin{aligned}\big|\bar{B}^n(t) - \bar{E}^n(t_0)\big| &= \big|\bar{E}^n(t) - \bar{R}^n(t) - \bar{E}^n(t_0)\big|\\ &\leq \big|\bar{E}^n(t) - \lambda t\big| + \big|\bar{R}^n(t) - \tilde{R}(t)\big| + \big|\bar{E}^n(t_0) - \lambda t_0\big| < 3\epsilon,\end{aligned}$$

for all large n. So

$$\begin{aligned}\tilde{\mathcal{R}}(t)(C_x) &- \int_{t-\frac{\tilde{R}(t)}{\lambda}}^{t} F^c(x+t-s)\,d\lambda s\\ &\leq 4\epsilon + \sum_{j=0}^{J-1}\big\langle 1_{(C^{\epsilon}_x+t-t_{j+1})}, \bar{\mathcal{L}}^n_F\big(E^n(t_j), \bar{E}^n(t_j, t_{j+1})\big)\big\rangle - \int_{t_0}^{t} F^c(x+t-s)\,d\lambda s,\end{aligned} \tag{5.39}$$

for all large n. Similarly, according to (5.29), we have

$$\begin{aligned}
&\tilde{\mathcal{R}}(t)(C_x) - \int_{t-\frac{\tilde{R}(t)}{\lambda}}^{t} F^c(x+t-s)\,d\lambda s \\
&\quad \geq -4\epsilon + \sum_{j=0}^{J-1} \bigl\langle 1_{(C_x^{\epsilon}+t-t_j)}, \bar{\mathcal{L}}_F^n\bigl(E^n(t_j), \bar{E}^n(t_j,t_{j+1})\bigr)\bigr\rangle - \int_{t_0}^{t} F^c(x+t-s)\,d\lambda s,
\end{aligned} \tag{5.40}$$

for all large n. Note that for each j, we have

$$\begin{aligned}
&\bigl\langle 1_{(C_x+t-t_{j+1})}, \bar{\mathcal{L}}_F^n\bigl(E^n(t_j), \bar{E}^n(t_j,t_{j+1})\bigr)\bigr\rangle \\
&\quad \leq \bigl\langle 1_{(C_x+t-t_{j+1})}, \bar{\mathcal{L}}_F^n\bigl(E^n(t_j), \lambda(t_{j+1}-t_j)+2\epsilon\bigr)\bigr\rangle \\
&\quad \leq \bigl[\lambda(t_{j+1}-t_j)+2\epsilon\bigr]\nu_F^n\bigl(C_x^{\epsilon}+t-t_{j+1}\bigr)+\epsilon \\
&\quad \leq \bigl[\lambda(t_{j+1}-t_j)+2\epsilon\bigr]\bigl[\nu_F(C_x+t-t_{j+1})+\epsilon\bigr]+\epsilon \\
&\quad \leq \lambda(t_{j+1}-t_j)\nu_F(C_x+t-t_{j+1})+(3+\lambda\delta)\epsilon
\end{aligned}$$

for all large n, where the first inequality is due to (5.36), the second one is due to (5.33) (the component of $\bar{V}_{M,L}^n$), the third one is due to (3.18) and the definition of the Prohorov metric, and the last one is due to algebra. Similarly, we can show that

$$\bigl\langle 1_{(C_x+t-t_j)}, \bar{\mathcal{L}}_F^n(E^n(t_j), \bar{E}^n(t_j,t_{j+1})\bigr\rangle \geq \lambda(t_{j+1}-t_j)\nu_F(C_x+t-t_j)-(3+\lambda\delta)\epsilon$$

for all large n. Note that $\sum_{j=0}^{J-1}\lambda(t_{j+1}-t_j)F^c(x+t-t_{j+1})$ and $\sum_{j=0}^{J-1}\lambda(t_{j+1}-t_j)F^c(x+t-t_j)$ serve as the upper and lower Reimann sum of the integral $\int_{t_0}^{t}F^c(x+t-s)\,d\lambda s$, which converge to the integration as $n\to\infty$. So by (5.39) and (5.40), we have, for all large n,

$$\left|\tilde{\mathcal{R}}(t)(C_x) - \int_{t-\frac{\tilde{R}(t)}{\lambda}}^{t} F^c(x+t-s)\,d\lambda s\right| \leq (3+\lambda\delta)J\epsilon+5\epsilon.$$

We conclude that $\tilde{\mathcal{R}}(t)(C_x) - \int_{t-\frac{\tilde{R}(t)}{\lambda}}^{t} F^c(x+t-s)\,d\lambda s = 0$ since ϵ in the above can be arbitrary. This verifies (3.1).

Next, we verify (3.2). For any $\epsilon>0$, consider the difference

$$\begin{aligned}
&\left|\tilde{\mathcal{Z}}(t)(C_x) - \bar{\mathcal{Z}}_0(C_x+t) - \int_0^t F^c\left(\frac{\tilde{R}(s)}{\lambda}\right)G^c(x+t-s)\,d\bigl[\lambda s-\tilde{R}(s)\bigr]\right| \\
&\quad \leq \bigl|\tilde{\mathcal{Z}}(t)(C_x) - \bar{\mathcal{Z}}^n(t)\bigl(C_x^{\epsilon}\bigr)\bigr| + \bigl|\tilde{\mathcal{Z}}_0(C_x+t) - \bar{\mathcal{Z}}^n(0)\bigl(C_x^{\epsilon}+t\bigr)\bigr| \\
&\qquad + \left|\bar{\mathcal{Z}}^n(t)\bigl(C_x^{\epsilon}\bigr) - \bar{\mathcal{Z}}^n(0)\bigl(C_x^{\epsilon}+t\bigr)\right. \\
&\qquad \left. - \int_0^t F^c\left(\frac{\tilde{R}(s)}{\lambda}\right)G^c(x+t-s)\,d\bigl[\lambda s-\tilde{R}(s)\bigr]\right|,
\end{aligned} \tag{5.41}$$

where the above inequality is due to the fluid scaled stochastic dynamic equation (5.3). Again, by the definition of Prohorov metric and the convergence in (5.35), each of the first two terms on the right hand side in the above inequality is less than ϵ for all large n. Let $\{t_j\}_{j=0}^{J}$ be a partition of the interval $[0,t]$ such that $0=t_0<t_1<\cdots<t_J=t$ and $\max_j(t_{j+1}-t_j)<\delta$ for some $\delta>0$. Let

$$\tilde{R}_{U,j}=\sup_{t\in[t_j,t_{j+1}]}\tilde{R}(t),\qquad \tilde{R}_{L,j}=\inf_{t\in[t_j,t_{j+1}]}\tilde{R}(t).$$

According to the definition of supremum, there exists $t\in[t_j,t_{j+1}]$ such that $\tilde{R}_{U,j}\le\tilde{R}(t)+\epsilon/2$. By (5.34), we have $\tilde{R}(t)\le\bar{R}^n(t)+\epsilon/2$ for all large n. This implies that $\tilde{R}_{U,j}\le\bar{R}^n_{U,j}+\epsilon$. Similarly, we can prove that $\bar{R}^n_{U,j}\le\tilde{R}_{U,j}+\epsilon$. The same approach can be applied to $\tilde{R}_{L,j}$ and $\bar{R}^n_{L,j}$. Thus we have

$$\bigl|\bar{R}^n_{U,j}-\tilde{R}_{U,j}\bigr|\le\epsilon,\qquad \bigl|\bar{R}^n_{L,j}-\tilde{R}_{L,j}\bigr|\le\epsilon,$$

for all large n. So for each j, we have

$$\begin{aligned}
&\bigl\langle 1_{(C_0+\frac{\bar{R}^n_{L,j}-2\epsilon}{\lambda})\times(C^{\epsilon}_x+t-t_{j+1})},\bar{\mathcal{L}}^n\bigl(B^n(t_j),\bar{B}^n(t_j,t_{j+1})\bigr)\bigr\rangle\\
&\quad\le\bigl\langle 1_{(C_0+\frac{\tilde{R}_{L,j}-3\epsilon}{\lambda})\times(C^{\epsilon}_x+t-t_{j+1})},\bar{\mathcal{L}}^n\bigl(B^n(t_j),\tilde{B}(t_{j+1})-\tilde{B}(t_j)+2\epsilon\bigr)\bigr\rangle\\
&\quad\le\bigl[\tilde{B}(t_{j+1})-\tilde{B}(t_j)+2\epsilon\bigr]\nu^n_F\biggl(C_0+\frac{\tilde{R}_{L,j}-3\epsilon}{\lambda}\biggr)\nu^n_G\bigl(C^{\epsilon}_x+t-t_{j+1}\bigr)+\epsilon\\
&\quad\le\bigl[\tilde{B}(t_{j+1})-\tilde{B}(t_j)+2\epsilon\bigr]\biggl[\nu_F\biggl(C_0+\frac{\tilde{R}_{L,j}}{\lambda}\biggr)+\frac{3\epsilon}{\lambda}\biggr]\bigl[\nu_G(C_x+t-t_{j+1})+\epsilon\bigr]+\epsilon
\end{aligned}$$

for all large n, where the first inequality is due to (5.37), the second one is due to (5.33) (the component of $\bar{V}^n_{M,L}$), the third one is due to (3.18). Let M_B be a finite upper bound of $\tilde{B}(t_J)-\tilde{B}(t_0)$; the above inequality can be further bounded by

$$\bigl[\tilde{B}(t_{j+1})-\tilde{B}(t_j)\bigr]\nu_F\biggl(C_0+\frac{\tilde{R}_{L,j}}{\lambda}\biggr)\nu_G(C_x+t-t_{j+1})+\biggl(\frac{3}{\lambda}+2\biggr)M_B\epsilon+3\epsilon.$$

Similarly, we can show that

$$\begin{aligned}
&\bigl\langle 1_{(C_0+\frac{\bar{R}^n_{U,j}+2\epsilon}{\lambda})\times(C_x+t-t_j)},\bar{\mathcal{L}}^n\bigl(B^n(t_j),\bar{B}^n(t_j,t_{j+1})\bigr)\bigr\rangle\\
&\quad\ge\bigl[\tilde{B}(t_{j+1})-\tilde{B}(t_j)\bigr]\nu_F\biggl(C_0+\frac{\tilde{R}_{L,j}}{\lambda}\biggr)\nu_G(C_x+t-t_j)-\biggl(\frac{3}{\lambda}+2\biggr)M_B\epsilon-3\epsilon.
\end{aligned}$$

Note that $\sum_{j=0}^{J-1}[\tilde{B}(t_{j+1})-\tilde{B}(t_j)]F^c(\frac{\tilde{R}_{U,j}}{\lambda})G^c(x+t-t_{j+1})$ and $\sum_{j=0}^{J-1}[\tilde{B}(t_{j+1})-\tilde{B}(t_j)]F^c(\frac{\tilde{R}_{L,j}}{\lambda})G^c(x+t-t_j)$ serve as the upper and lower Reimann sum of the integral $\int_{t_0}^{t}F^c(\frac{\tilde{R}(s)}{\lambda})G^c(x+t-s)\,d\tilde{B}(s)$, which converge to the integration as $n\to\infty$.

So, by (5.30) and (5.31), we have, for all large n,

$$\begin{aligned}&\left|\bar{\mathcal{Z}}^n(t)\big(C_x^\epsilon\big) - \bar{\mathcal{Z}}^n(0)\big(C_x^\epsilon + t\big) - \int_{t_0}^t F^c\left(\frac{\tilde{R}(s)}{\lambda}\right) G^c(x+t-s)\, d\tilde{B}(s)\right|\\ &\quad\le \left(\frac{3}{\lambda}+2\right) M_B\epsilon + 3\epsilon + \epsilon.\end{aligned}$$

In summary, the right hand side of (5.41) can be bounded by a finite multiple of ϵ. We conclude that the left hand side of (5.41) must be 0 since it does not depend on ϵ, which can be arbitrary. This verifies (3.2).

The verification of fluid constrains (3.3) and (3.4) is quite straightforward. Basically, it is just passing the fluid scaled stochastic constraints

$$\bar{Q}^n(t) = \big(\bar{X}^n(t) - 1\big)^+,$$
$$\bar{Z}^n(t) = \big(\bar{X}^n(t) \wedge 1\big),$$

to $n \to \infty$. We omit it for brevity. □

6 The special case with exponential distribution

In this section we verify that the fluid model developed in this paper for the general patience and service time distributions is consistent with the one in [30], that was obtained in the special case where both distributions are assumed to be exponential.

Our fluid model equations implies the key relationship (4.6). Now, we specialize in the case with exponential distribution, i.e.

$$F(t) = F_e(t) = 1 - e^{-\alpha t}, \qquad G(t) = G_e(t) = 1 - e^{-\mu t}, \quad \text{for all } t \ge 0.$$

Now (4.6) becomes

$$\begin{aligned}\bar{X}(t) &= \zeta_0(t) + \rho \int_0^t \left[1 - \frac{\alpha}{\lambda}\big(\big(\bar{X}(t-s) - 1\big)^+\big)\right] \mu e^{-\mu s}\, ds\\ &\quad + \int_0^t \big(\bar{X}(t-s) - 1\big)^+ \mu e^{-\mu s}\, ds.\end{aligned}$$

In the case of exponential service time distribution, the remaining service time of those initially in service and the service times of those initially waiting in queue are also assumed to be exponentially distributed. So we have

$$\zeta_0(t) = \bar{\mathcal{Z}}_0(C_0 + t) + \bar{Q}_0 e^{-\mu t} = \bar{X}_0 e^{-\mu t},$$

where $\bar{X}_0 = \bar{Z}_0 + \bar{Q}_0$ is the initial number of customers in the system. By some algebra, the above two equations can be simplified as the follows:

$$\bar{X}(t) = \bar{X}_0 e^{-\mu t} + \rho\big[1 - e^{-\mu t}\big] + (\mu - \alpha) \int_0^t \big(\bar{X}(t-s) - 1\big)^+ e^{-\mu s}\, ds. \tag{6.1}$$

By the change of variable $t-s \to s$, the above integration can be written as

$$\int_0^t \big(\bar{X}(t-s)-1\big)^+ e^{-\mu s}\,ds = e^{-\mu t}\int_0^t \big(\bar{X}(s)-1\big)^+ e^{\mu s}\,ds.$$

Taking the derivative on both sides of (6.1) yields

$$\begin{aligned}
\bar{X}'(t) &= -\mu X_0 e^{-\mu t} + \mu\rho e^{\mu t} \\
&\quad + (\mu-\alpha)\left[-\mu e^{-\mu t}\int_0^t \big(\bar{X}(s)-1\big)^+ e^{\mu s}\,ds + e^{-\mu t}\big(\bar{X}(t)-1\big)^+ e^{\mu t}\right] \\
&= -\mu X_0 e^{-\mu t} - \mu\rho\big[1-e^{\mu t}\big] + \mu\rho \\
&\quad - \mu(\mu-\alpha)e^{-\mu t}\int_0^t \big(\bar{X}(s)-1\big)^+ e^{\mu s}\,ds + (\mu-\alpha)\big(\bar{X}(t)-1\big)^+ \\
&= -\mu\bar{X}(t) + \mu\rho + (\mu-\alpha)\big(\bar{X}(t)-1\big)^+.
\end{aligned}$$

Using the notation in [30], $a^- = -\min(0,a)$ for any $a \in \mathbb{R}$. Note that $a = \min(a,1) + (a-1)^+ = 1-(a-1)^- + (a-1)^+$. So the above equation further implies

$$\bar{X}'(t) = \mu(\rho-1) - \alpha\big(\bar{X}(t)-1\big)^+ + \mu\big(\bar{X}(t)-1\big)^-, \quad \text{for all } t \geq 0.$$

This equation is consistent with Theorem 2.2 in [30] (μ is assumed to be 1 in that paper).

Acknowledgements The author would like to thank Jim Dai and Bert Zwart for many useful discussions. The author is grateful to Christian Gromoll for suggesting a nice method on using the Skorohod Representation theorem to make the presentation in Sect. 5.2 rigorous. This research is supported in part by Hong Kong GRF grant 622110 and National Science Foundation grants CMMI-0727400 and CNS-0718701.

Appendix A: A convolution equation

The main purpose for this appendix is to study the convolution equation

$$x(t) = \zeta(t) + \rho\int_0^t H\big(\big(x(t-s)-1\big)^+\big)\,dG_e(s) + \int_0^t \big(x(t-s)-1\big)^+\,dG(s), \quad \text{(A.1)}$$

where G_e is the equilibrium distribution of distribution function G as defined in Sect. 3.1. We first show that (A.1) has a unique solution under the assumption that the function $H(\cdot)$ is Lipschitz continuous. The proof follows the application of the classical contraction mapping theorem.

Lemma A.1 *Assume that $G(\cdot)$ is a distribution function with $G(0) < 1$, $\zeta(\cdot) \in$ $\mathbf{D}([0,T],\mathbb{R})$, $H(\cdot)$ is a Lipschitz continuous function, and $\rho \in \mathbb{R}$. There exists a unique solution $x^*(\cdot) \in \mathbf{D}([0,T],\mathbb{R})$ to (A.1).*

Proof Suppose $H(\cdot)$ is Lipschitz continuous with constant L. The equilibrium distribution has density $\mu[1-G(\cdot)]$, so $|G_e(t)-G_e(s)| \le \mu|t-s|$ for any $s,t \in \mathbb{R}$. Since $G(0)<1$, there exists $b>0$ such that

$$\kappa := \rho L\big[G_e(b)-G_e(0)\big] + \big[G(b)-G(0)\big] < 1.$$

According to Theorem 3.5.6 in [7], the space $\mathbf{D}([0,b],\mathbb{R})$ (all real valued *càdlàg* functions on $[0,b]$, cf. Sect. 1.1) is complete since $\mathbb{R}$ is complete. Now consider the space $\mathbf{D}([0,b],\mathbb{R})$ (each function in it is bounded and measurable by Corollaries 12.2.3 and 12.2.4 in [29], respectively) is a subset of the Banach space of bounded, measurable functions on $[0,b]$, equipped with the sup norm. One can check that this subset is closed in the Banach space. Thus, the space $\mathbf{D}([0,b],\mathbb{R})$ itself, equipped with the uniform metric υ_T (defined in Sect. 1.1), is complete.

For any $y \in \mathbf{D}([0,b],\mathbb{R})$, define $\Psi(y)$ by

$$\Psi(y)(t) = \zeta(t) + \rho \int_0^t H\big(\big(y(t-s)-1\big)^+\big)\,dG_e(s) + \int_0^t \big(y(t-s)-1\big)^+\,dG(s),$$

for any $t \in [0,b]$. By convention, the integration $\int_0^t y(t-s)\,dF(s)$ is interpreted to be $\int_{(0,t]} y(t-s)\,dF(s)$ (cf. p. 43 in [4]). We prove the existence and uniqueness of the solution to (A.1) by showing that Ψ is a contraction mapping on $\mathbf{D}([0,b],\mathbb{R})$. According to the proof of Lemma A.1 in [33], the convolution of a *càdlàg* function with a distribution function is still a *càdlàg* function. So Ψ is a mapping from $\mathbf{D}([0,b],\mathbb{R})$ to $\mathbf{D}([0,b],\mathbb{R})$. Next, we show that the mapping Ψ is a contraction. For any $y, y' \in \mathbf{D}([0,b],\mathbb{R})$, we have

$$\begin{aligned}\upsilon_b\big[\Psi(y),\Psi\big(y'\big)\big] &\le \sup_{t\in[0,b]} \rho \int_0^t L\big|\big(y(t-s)-1\big)^+ - \big(y'(u-v)-1\big)^+\big|\,dG_e(s) \\ &\quad + \sup_{t\in[0,b]} \int_0^t \big|\big(y(t-s)-1\big)^+ - \big(y'(t-s)-1\big)^+\big|\,dG(s) \\ &\le \rho L \int_0^b \upsilon_b\big[y,y'\big]\,dG_e(s) + \int_0^b \upsilon_b\big[y,y'\big]\,dG(s) \\ &\le \kappa\upsilon_b\big[y,y'\big].\end{aligned}$$

Since $\kappa < 1$, the mapping Ψ is a contraction. By the contraction mapping theorem (cf. Theorem 3.2 in [14]), Ψ has a unique fixed point x, i.e. $x = \psi(x)$. This implies that $x \in \mathbf{D}([0,b],\mathbb{R})$ is the unique solution to (A.1) on $[0,b]$.

It now remains to extend the existence and uniqueness result from $[0,b]$ to $[0,T]$. Denote $x_b(t) = x(b+t)$, $\zeta_b(t) = \zeta(b+t) + \rho\int_t^{b+t} H((x(b+t-s)-1)^+)\,dG_e(s) + \int_t^{b+t}(x(b+t-s)-1)^+\,dG(s)$, then we have for $t \in [0,T-b]$,

$$x_b(t) = \zeta_b(t) + \rho\int_0^t H\big(\big(x_b(t-s)-1\big)^+\big)\,dG_e(s) + \int_0^t \big(x_b(t-s)-1\big)^+\,dG(s). \tag{A.2}$$

It follows from the previous argument that there is unique solution $x_b(\cdot)$ to the above equation. Thus, we obtain a unique extension of the solution to (A.1) on the interval $[0,2b]$. Repeating this approach for N time with $N \geq \lceil T/b \rceil$ gives a unique solution on the interval $[0,T]$. □

For the application of this paper, $H(\cdot)$ sometimes cannot be guaranteed to be Lipschitz continuous. Recall the definition of $H(\cdot)$ in (4.5), one can easily see that when the patience time distribution $F(\cdot)$ has finite support, $H(\cdot)$ will not necessarily be Lipschitz continuous. A simple example is that when $F(x) = 1 - x$, the distribution function for a random variable uniformly distributed on the interval [0, 1]. However, the function $H(\cdot)$ given by (4.5) is non-increasing and is Lipschitz continuous on any sub-interval within its support (cf. Proof of Theorem 3.1 for detailed discussion). In the following, we prove the existence and uniqueness of the solution to (A.1) by leveraging these properties of $H(\cdot)$.

Lemma A.2 *Assume that $G(\cdot)$ is a distribution function with mean $\mu \in (0,\infty)$ and $G(0) < 1$, $\rho = \lambda/\mu$ with $\lambda \in (0,\infty)$, $\zeta(\cdot) \in \mathbf{D}([0,T],\mathbb{R})$ that satisfies the following condition*:

$$\zeta(t) = g_0(t) + \bigl(\zeta(0) - 1\bigr)^+\bigl[1 - G(t)\bigr], \tag{A.3}$$

where $g_0(\cdot)$ is a non-increasing function, and $H(\cdot)$ is a function that satisfies the following conditions:

$$H(x) \geq 0 \quad \textit{for all } x \geq 0, \tag{A.4}$$

$$H(\cdot) \textit{ is non-increasing}, \tag{A.5}$$

$$\begin{aligned} &H(\cdot) \textit{ is Lipschitz continuous on } [0, S_H - \delta] \textit{ for any } \delta > 0 \textit{ if } S_H < \infty;\\ &\qquad\qquad\qquad\qquad \textit{on } [0, M] \textit{ for any } M > 0 \textit{ if } S_H = \infty, \end{aligned} \tag{A.6}$$

where $S_H = \inf\{x \geq 0 : H(x) = 0\}$. There exists a unique solution $x^*(\cdot) \in \mathbf{D}([0,T],\mathbb{R})$ to (A.1).

This proof applies part of the argument in Lemma A.1 and some transformation of (A.1) based on the condition (A.3). To better structure the proof, we first show the following auxiliary result.

Lemma A.3 *Assume that $G(\cdot)$ is a distribution function with mean $\mu \in (0,\infty)$ and $G(0) < 1$, $\rho = \lambda/\mu$ with $\lambda \in (0,\infty)$, $\zeta(\cdot) \in \mathbf{D}([0,T],\mathbb{R})$ satisfies* (A.3) *and $H(\cdot)$ is continuous and satisfies* (A.4). *Suppose $x(\cdot) \in \mathbf{D}([0,T],\mathbb{R})$ is a function* (*if any*) *that satisfies* (A.1). *Then the function*

$$\bigl(x(t) - 1\bigr)^+ - \lambda \int_0^t H\bigl(\bigl(x(s) - 1\bigr)^+\bigr)\,ds$$

is non-increasing.

Proof To simplify the notation, let $Q(t) = (x(t) - 1)^+$ and

$$D(t) = Q(t) - \lambda \int_0^t H\big(Q(s)\big)\,ds \tag{A.7}$$

for all $t \in [0, T]$. Since $G_e(\cdot)$ is the equilibrium distribution, we have

$$\begin{aligned} x(t) &= \zeta(t) + \rho \int_0^t H\big(Q(t-s)\big)\mu\big[1 - G(s)\big]\,ds + \int_0^t Q(t-s)\,dG(s) \\ &= \zeta(t) + \lambda \int_0^t H\big(Q(s)\big)\,ds - \lambda \int_0^t H\big(Q(s)\big)G(t-s)\,ds + \int_0^t Q(t-s)\,dG(s). \end{aligned}$$

Applying Fubini's Theorem (cf. Theorem 8.4 in [20]) to the second to the last integral in the above, we have

$$\begin{aligned} \int_0^t H\big(Q(s)\big)G(t-s)\,ds &= \int_0^t \int_0^{t-s} H\big(Q(s)\big)\,dG(\tau)\,ds \\ &= \int_0^t \int_0^{t-\tau} H\big(Q(s)\big)\,ds\,dG(\tau). \end{aligned}$$

So we obtain

$$x(t) - \lambda \int_0^t H\big(Q(s)\big)\,ds = \zeta(t) + \int_0^t \left[Q(t-s) - \lambda \int_0^{t-s} H\big(Q(\tau)\big)\,d\tau\right] dG(s).$$

According to the above definition of $D(\cdot)$, we have

$$\big(x(t) \wedge 1\big) + D(t) = \zeta(t) + \int_0^t D(t-s)\,dG(s). \tag{A.8}$$

It now remains to use (A.8) to show that $D(\cdot)$ is non-increasing, i.e. for any $t, t' \in [0, T]$ with $t \le t'$, we have $D(t) \ge D(t')$. Since $G(0) < 1$, there exists $a > 0$ such that $G(a) < 1$. We first show that $D(\cdot)$ is non-increasing on the interval $[0, a]$. Let

$$D^* = \sup_{\{(t,t') \in [0,a] \times [0,a] : t \le t'\}} D\big(t'\big) - D(t).$$

Since $D(\cdot)$ is *càdlàg*, according to Theorem 6.2.2 in the supplement of [29], it is bounded on the interval $[0, a]$. Thus, D^* is finite. We will prove by contradiction that $D^* \le 0$, which shows that $D(\cdot)$ is non-increasing on $[0, a]$. Assume on the contrary that $D^* > 0$. Applying (A.8), we have

$$\begin{aligned} D\big(t'\big) - D(t) = {}& \big(x(t) \wedge 1\big) - \big(x\big(t'\big) \wedge 1\big) + \zeta\big(t'\big) - \zeta(t) \\ & + \int_0^{t'} D\big(t' - s\big)\,dG(s) - \int_0^t D(t-s)\,dG(s) \end{aligned}$$

$$= \bigl(x(t) \wedge 1\bigr) - \bigl(x(t') \wedge 1\bigr) + \zeta(t') - \zeta(t)$$
$$+ \int_t^{t'} D(t'-s)\, dG(s) + \int_0^t \bigl[D(t'-s) - D(t-s)\bigr]\, dG(s).$$

It follows from (A.1) and (A.7) that $D(0) = (\zeta(0) - 1)^+$. This together with condition (A.3) implies that

$$\zeta(t') - \zeta(t) = g_0(t') - g_0(t) + D(0)\bigl[G(t) - G(t')\bigr]. \tag{A.9}$$

So

$$\begin{aligned} D(t') - D(t) = \bigl(x(t) \wedge 1\bigr) - \bigl(x(t') \wedge 1\bigr) + g_0(t') - g_0(t) \\ + \int_t^{t'} \bigl[D(t'-s) - D(0)\bigr]\, dG(s) \\ + \int_0^t \bigl[D(t'-s) - D(t-s)\bigr]\, dG(s). \end{aligned} \tag{A.10}$$

If $x(t') < 1$, by (A.7),

$$D(t') - D(t) = -\lambda \int_t^{t'} H\bigl(Q(s)\bigr)\, ds - Q(t),$$

which is always non-positive; if $x(t') \geq 1$, then $(x(t) \wedge 1) - (x(t') \wedge 1) \leq 0$. So it follows from (A.10) and $g_0(\cdot)$ being non-increasing that

$$\begin{aligned} D(t') - D(t) &\leq \int_t^{t'} \bigl[D(t'-s) - D(0)\bigr]\, dG(s) + \int_0^t \bigl[D(t'-s) - D(t-s)\bigr]\, dG(s) \\ &\leq \int_0^{t'} D^*\, dG(s) = D^* G(t') \leq D^* G(a), \end{aligned}$$

where the last inequality follows from the assumption that D^* is non-negative. Summarizing both cases of $x(t')$, we have

$$D(t') - D(t) \leq \max\bigl(0, D^* G(a)\bigr)$$

for all $t, t' \in [0, a] > 0$ with $t \leq t'$. Taking the supremum on both sides over the set $\{(t, t') \in [0, a] \times [0, a] : t \leq t'\}$ gives $D^* \geq F(a) D^*$. This implies that $[1 - G(a)] D^* \leq 0$. Since $G(a) < 1$, it contradicts the assumption that $D^* > 0$. So we must have $D^* \leq 0$, this implies that $D(\cdot)$ is non-increasing on $[0, a]$. We next extend this property to the interval $[0, T]$ using induction. Suppose we can show that $D(\cdot)$ is non-decreasing on the interval $[0, na]$ for some $n \in \mathbb{N}$. Introduce $D_{na}(t) = D(na + t)$, $x_{na}(t) = x(na + t)$ and

$$\zeta_{na}(t) = \zeta(na + t) + \int_0^{na} D(na - s)\, dG(t + s). \tag{A.11}$$

It is clear that the shifted functions satisfy

$$\bigl(x_{na}(t)\wedge 1\bigr)+D_{na}(t)=\zeta_{na}(t)+\int_0^t D_{na}(t-s)\,dG(s). \tag{A.12}$$

To show that $D(\cdot)$ is non-increasing on $[na,(n+1)a]$ is the same as to show that $D_{na}(\cdot)$ is non-increasing on $[0,a]$. For this purpose, it is enough to verify that $\zeta_{na}(\cdot)$ satisfy the condition (A.9). Performing integration by parts on (A.11) gives

$$\begin{aligned}\zeta_{na}(t)&=\zeta(na+t)+\bigl(\zeta(0)-1\bigr)^+\bigl[1-G(na+t)\bigr]+\int_0^{na} D(na-s)\,dG(t+s)\\ &=\zeta(na+t)+\bigl(\zeta(0)-1\bigr)^+\bigl[1-G(na+t)\bigr]\\ &\quad+D(0)G(na+t)-D(na)G(t)-\int_0^{na} G(t+s)\,dD(na-s).\end{aligned}$$

It follows from (A.1) and (A.7) that $D(0)=(\zeta(0)-1)^+$, so we can write $\zeta_{na}(\cdot)$ as

$$\zeta_{na}(t)=h_{na}(t)+D_{na}(0)\bigl[1-G(t)\bigr],$$

where $g_{na}(t)=\zeta(na+t)+(\zeta(0)-1)^+-D_{na}(0)-\int_0^{na}G(t+s)\,dD(na-s)$. Since $G(\cdot)$ is non-decreasing and $D(\cdot)$ is non-increasing, the integral $-\int_0^{na}G(t+s)\,dD(na-s)$ is non-increasing as a function of t. So we can conclude that $g_{na}(\cdot)$ is non-increasing, i.e. $\zeta_{na}(\cdot)$ satisfies condition (A.9). Thus, we extend the non-increasing interval to $[0,(n+1)a]$. By induction, the function $D(\cdot)$ is non-increasing on the interval $[0,T]$. □

Proof of Lemma A.2 If $H(0)=0$, then by conditions (A.4) and (A.5), $H\equiv 0$. In this case, $H(\cdot)$ is Lipschitz, thus the result follows from Lemma A.1. For the rest of the proof, assume that $H(0)>0$, thus $S_H>0$. According to Lemma A.3, any function $x(\cdot)$ that satisfies (A.1) must also satisfy

$$\bigl(x(t')-1\bigr)^+-\bigl(x(t)-1\bigr)^+\le\lambda\int_t^{t'}H\bigl(\bigl(x(s)-1\bigr)^+\bigr)\,ds\le\lambda H(0)\bigl(t'-t\bigr) \tag{A.13}$$

for any $t<t'$. So we can prove an upper abound for $x(\cdot)$ on the interval $[0,T]$,

$$\sup_{t\in[0,T]}x(t)\le 1+H(0)T.$$

If $S_H=\infty$, set $M=1+H(0)T$. Recall the constant b as defined in the proof of Lemma A.1. We can apply the same argument by restricting the map Ψ, also defined in the proof of Lemma A.1, on $\mathbf{D}([0,b],(-\infty,M])$ instead of $\mathbf{D}([0,b],\mathbb{R})$. By condition (A.6), H is Lipschitz continuous on $[0,M]$. The result is proved by using the same application of contraction mapping theory as in Lemma A.1. Now we focus on the case where $S_H<\infty$. Fix a $\delta\in(0,S_H/2)$, and consider the following two cases.

Case 1, $\zeta(0)<1+S_H-2\delta$. Let $b'=\delta/(\lambda H(0))$. Since $x(0)=\zeta(0)$, according to (A.13), we have an upper bound,

$$\sup_{t\in[0,b']}x(t)\le 1+S_H-\delta.$$

By condition (A.6), H is Lipschitz continuous on $[0, S_H - \delta]$. Again, we can apply the contraction mapping theorem as in Lemma A.1 with the following adjustment. Let $b_1 = \min[b, b']$ and restrict the mapping Ψ on the space $\mathbf{D}([0, b_1], [0, 1 + S_H - \delta])$ instead of $\mathbf{D}([0, b_1], \mathbb{R})$. Thus, we have the existence and uniqueness of the solution on a small interval $[0, b_1]$. To extend the solution to $[kb_1, (k+1)b_1]$, $k = 1, 2, \dots$ till we cover the interval $[0, T]$, we also apply the same approach as in proving Lemma A.1. However, we have to stop at k whenever $x(kb_1)$ goes above $1 + S_H - 2\delta$. We now turn to the analysis of the second case.

Case 2, $\zeta(0) \geq 1 + S_H - 2\delta$ (which is strictly larger than 1 according to the definition of δ). By right-continuity, we know that there exists $b_2 > 0$ such that

$$x(t) \geq 1 \quad \text{for all } t \in [0, b_2].$$

Same as in the proof of Lemma A.3, denote $Q(t) = (x(t) - 1)^+$ and

$$D(t) = Q(t) - \lambda \int_0^t H\bigl(Q(s)\bigr)\, ds \tag{A.14}$$

to simplify the notation. For $t \in [0, b_2]$, (A.8) obtained in the proof of Lemma A.3 becomes

$$D(t) = \zeta(t) - 1 + \int_0^t D(t-s)\, dG(s).$$

According to Theorem 2.4 in Chap. V of [1], the above renewal equation has a unique solution $D(\cdot)$. Since $\zeta(\cdot) \in \mathbf{D}([0, b_2], \mathbb{R})$, $D(\cdot) \in \mathbf{D}([0, b_2], \mathbb{R})$. It now remains to prove that (A.14) with a known $D(\cdot)$ has a unique solution. Now let $-Q_0(\cdot) \equiv 0$ and define

$$-Q_{k+1}(t) = D(t) - 2Q_k(t) + \lambda \int_0^t H\bigl(Q_k(s)\bigr)\, ds, \quad k = 0, 1, 2, \dots,$$

Specialize $k = 0$ in the above and plug in (A.14), we have

$$\begin{aligned} Q_0(t) - Q_1(t) &= D(t) - 0 + \lambda \int_0^t H(0)\, ds \\ &= Q(t) - \lambda \int_0^t H\bigl(Q(s)\bigr)\, ds + \lambda \int_0^t H(0)\, ds \\ &\geq 0 + \lambda \int_0^t \bigl[H(0) - H\bigl(Q(s)\bigr)\bigr]\, ds \geq 0, \end{aligned}$$

where the last inequality is due to that (A.5) and $Q(t) \geq 0$. So we have $-Q_0(t) \leq -Q_1(t)$ for all $t \geq 0$. Note that for all $k \geq 1$,

$$\begin{aligned} -Q_{k+1}(t) - \bigl(-Q_k(t)\bigr) = {} & 2\bigl[-Q_k(t) - \bigl(-Q_{k-1}(t)\bigr)\bigr] \\ & + \lambda \int_0^t \bigl[H\bigl(Q_k(s)\bigr) - H\bigl(Q_{k-1}(s)\bigr)\bigr]\, ds. \end{aligned}$$

Note that $H(x) = H(-(-x))$ and by condition (A.5), $H(-\cdot)$ is non-decreasing. Thus, we can prove by induction that $-Q_k(t) \leq -Q_{k+1}(t)$ for all $k = 0, 1, 2, \dots$.

We have to use the trick $-Q_k$ just because the function H is non-increasing. Define

$$-Q(t) = \lim_{k\to\infty} -Q_k(t), \quad \text{for all } t \geq 0.$$

It now remains to verify that $Q(t)$ satisfies equation (A.14). This follows immediately from the monotone convergence theorem (Theorem 4.3.2 in [6]):

$$\int_0^t H\bigl(Q_k(s)\bigr)\,ds \to \int_0^t H\bigl(Q(s)\bigr)\,ds \quad \text{as } k \to \infty.$$

So we have resolved case 2. With the help of case 2, we can further extend the solution to a point where the solution $x(\cdot)$ reaches 1. Starting from there, we can apply case 1 to extend the solution to an extra small interval with length b_1. With the process continuing, we can cover the interval $[0, T]$. □

Appendix B: Glivenko–Cantelli estimates

An important preliminary result is the following estimate due to Glivenko–Cantelli. It is used in Sect. 5. It is convenient to state it as a general result, since the Glivenko–Cantelli estimate requires weaker conditions and gives stronger results than those in this paper.

For each n, let $\{u_i^n\}_{i\in\mathbb{Z}}$ be a sequence of i.i.d. random variables with probability measure $\nu_F^n(\cdot)$, let $\{u_i^n\}_{i\in\mathbb{Z}}$ be a sequence of i.i.d. random variables with probability measure $\nu_G^n(\cdot)$. For any $n \in \mathbb{N}$, $m \in \mathbb{Z}$ and $l \in \mathbb{R}_+$, define

$$\bar{\mathcal{L}}_F^n(m,l) = \frac{1}{n}\sum_{i=m+1}^{m+\lfloor nl\rfloor} \delta_{u_i^n}, \tag{B.1}$$

$$\bar{\mathcal{L}}_G^n(m,l) = \frac{1}{n}\sum_{i=m+1}^{m+\lfloor nl\rfloor} \delta_{v_i^n}, \tag{B.2}$$

$$\bar{\mathcal{L}}^n(m,l) = \frac{1}{n}\sum_{i=m+1}^{m+\lfloor nl\rfloor} \delta_{(u_i^n,v_i^n)}, \tag{B.3}$$

where δ_x denotes the Dirac measure of point x on $\mathbb{R}$ and $\delta_{(x,y)}$ denotes the Dirac measure of point (x,y) on $\mathbb{R}\times\mathbb{R}$. So $\bar{\mathcal{L}}_F^n(m,l)$ and $\bar{\mathcal{L}}_G^n(m,l)$ are measures on $\mathbb{R}$ and $\bar{\mathcal{L}}^n(m,l)$ is a measure on $\mathbb{R}\times\mathbb{R}$.

Denote $C_x = (x,\infty)$, for all $x \in \mathbb{R}$. We define two classes of testing functions by

$$\mathscr{V} = \bigl\{1_{C_x}(\cdot) : x \in \mathbb{R}\bigr\},$$
$$\mathscr{V}_2 = \bigl\{1_{C_x\times C_y}(\cdot,\cdot) : x, y \in \mathbb{R}\bigr\}.$$

It is clear that $\mathscr{V}$ is a set of functions on $\mathbb{R}$ and $\mathscr{V}_2$ is a set of functions on $\mathbb{R}\times\mathbb{R}$. Define an envelop function for $\mathscr{V}$ as follows. Since $\nu_F^n \to \nu_F$, by Skorohod representation theorem, there exist random variables X^n (with law ν_F^n) and X (with law ν_F), such that $X^n \to X$ almost surely as $r \to \infty$. Thus there exists a random variable X^* such that almost surely,

$$X^* = \sup_n X^n.$$

Let ν_F^* be the law of X^*. Since $L_2(\nu_F^*)$ (the space of square integrable functions with respect to the measure ν_F^*) contains continuous unbounded functions, there exists a continuous unbounded function $f_{\nu_F} : \mathbb{R}_+ \to \mathbb{R}$ that is increasing, satisfies $f_{\nu_F} \geq 1$ and $\langle f_{\nu_F}^2, \nu_F^* \rangle < \infty$. This implies that

$$\langle f_{\nu_F}^2, \nu_F \rangle = \mathbb{E}\bigl[f_{\nu_F}^2(X)\bigr] \leq \mathbb{E}\bigl[f_{\nu_F}^2\bigl(X^*\bigr)\bigr] = \langle f_{\nu_F}^2, \nu_F^* \rangle < \infty. \tag{B.4}$$

Similarly, based on the weak convergence $\nu_G^n \to \nu_G$, we can construct a function f_{ν_G} that is increasing, satisfies $f_{\nu_G} \geq 1$ and $\langle f_{\nu_G}^2, \nu_G \rangle < \infty$. Now, define function $\bar{f} : \mathbb{R}_+ \to \mathbb{R}$ by

$$\bar{f}(x) = \min\bigl(f_{\nu_F}(x), f_{\nu_G}(x)\bigr) \tag{B.5}$$

and function $\bar{f}_2 : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ by

$$\bar{f}_2(x, y) = \min\bigl(f_{\nu_F}(x), f_{\nu_G}(y)\bigr) \tag{B.6}$$

for all $x, y \in \mathbb{R}_+$. Note that we have to following properties:

$$\bar{f} \text{ is increasing and unbounded,} \tag{B.7}$$

$$f \leq \bar{f} \quad \text{for all } f \in \mathcal{V}, \tag{B.8}$$

$$f \leq \bar{f}_2 \quad \text{for all } f \in \mathcal{V}_2. \tag{B.9}$$

So we call $\bar{f}$ and $\bar{f}_2$ the envelop function for $\mathcal{V}$ and $\mathcal{V}_2$, respectively. Finally, let $\bar{\mathcal{V}} = \{\bar{f}\} \cup \mathcal{V}$ and $\bar{\mathcal{V}}_2 = \{\bar{f}_2\} \cup \mathcal{V}_2$.

Lemma B.1 *Assume that*

$$\nu_F^n \to \nu_F, \qquad \nu_G^n \to \nu_G \quad \textit{as } n \to \infty.$$

Fix constants $M, L > 0$. *For all* $\epsilon, \eta > 0$,

$$\limsup_{n\to\infty} \mathbb{P}^n\Bigl(\max_{-nM<m<nM} \sup_{l\in[0,L]} \sup_{f\in\bar{\mathcal{V}}} \bigl|\langle f, \bar{\mathcal{L}}_F^n(m,l)\rangle - l\langle f, \nu_F^n\rangle\bigr| > \epsilon\Bigr) < \eta,$$

$$\limsup_{n\to\infty} \mathbb{P}^n\Bigl(\max_{-nM<m<nM} \sup_{l\in[0,L]} \sup_{f\in\bar{\mathcal{V}}} \bigl|\langle f, \bar{\mathcal{L}}_G^n(m,l)\rangle - l\langle f, \nu_G^n\rangle\bigr| > \epsilon\Bigr) < \eta,$$

$$\limsup_{n\to\infty} \mathbb{P}^n\Bigl(\max_{-nM<m<nM} \sup_{l\in[0,L]} \sup_{f\in\bar{\mathcal{V}}_2} \bigl|\langle f, \bar{\mathcal{L}}^n(m,l)\rangle - l\langle f, \bigl(\nu_F^n, \nu_G^n\bigr)\rangle\bigr| > \epsilon\Bigr) < \eta.$$

These kinds of result have been widely used in the study of measure valued processes, see [10, 12, 33]. The proof of the first two inequalities in the above lemma follows exactly the same way as the one for Lemma B.1 in [33], and the proof of the third inequality in the above lemma follows exactly the same as the one for Lemma 5.1 in [12]. We omit the proof for brevity. By the same reasoning as for Lemma 5.2, there exists a function $\epsilon_{\mathrm{GC}}(\cdot)$, which vanishes at infinity such that the ϵ and η in the above lemma can be replaced by the function $\epsilon_{\mathrm{GC}}(n)$ for each index n. Based on this, we construct the following event:

$$\begin{aligned}
&\Omega^n_{\mathrm{GC}}(M,L)\\
&\quad=\Bigl\{\max_{-nM<m<nM}\sup_{l\in[0,L]}\sup_{f\in\bar{\mathcal{V}}}\bigl|\bigl\langle f,\bar{\mathcal{L}}^n_F(m,l)\bigr\rangle-l\bigl\langle f,\nu^n_F\bigr\rangle\bigr|\le\epsilon_{\mathrm{GC}}(n)\Bigr\}\\
&\qquad\cap\Bigl\{\max_{-nM<m<nM}\sup_{l\in[0,L]}\sup_{f\in\bar{\mathcal{V}}}\bigl|\bigl\langle f,\bar{\mathcal{L}}^n_G(m,l)\bigr\rangle-l\bigl\langle f,\nu^n_G\bigr\rangle\bigr|\le\epsilon_{\mathrm{GC}}(n)\Bigr\}\\
&\qquad\cap\Bigl\{\max_{-nM<m<nM}\sup_{l\in[0,L]}\sup_{f\in\bar{\mathcal{V}}_2}\bigl|\bigl\langle f,\bar{\mathcal{L}}^n(m,l)\bigr\rangle-l\bigl\langle f,\bigl(\nu^n_F,\nu^n_G\bigr)\bigr\rangle\bigr|\le\epsilon_{\mathrm{GC}}(n)\Bigr\}. \qquad \text{(B.10)}
\end{aligned}$$

It is clear that for any fixed $M,L>0$,

$$\lim_{n\to\infty}\mathbb{P}^n\bigl(\Omega^n_{\mathrm{GC}}(M,L)\bigr)=1. \qquad \text{(B.11)}$$

Intuitively, on the event $\Omega^n_{\mathrm{GC}}(M,L)$ (whose probability goes to 1 as $n\to\infty$ for any fixed constants M,L), the measures $\bar{\mathcal{L}}^n_F(m,l)$, $\bar{\mathcal{L}}^n_G(m,l)$ and $\bar{\mathcal{L}}^n(m,l)$ are very "close" to $l\nu^n_F$, $l\nu^n_G$ and $l(\nu^n_F,\nu^n_G)$, respectively.

Appendix C: An extension of Skorohod representation theorem

In this section we present a slight extension, Lemma C.1 below, of the Skorohod Representation Theorem (cf. Theorem 3.2.2 in [29]). The proof of Lemma C.1 is built on the proof of Theorem 3.2.2 provided in the supplement of [29], with slight extension to deal with the product of two metric spaces.

Let $(\mathbf{E}_1,\pi_1)$ and $(\mathbf{E}_2,\pi_2)$ be two complete and separable metric spaces. Let $(\mathbf{E}_1\times\mathbf{E}_2,\pi)$ denote the product space of them, with the product metric π obtained by the maximum metric.

Lemma C.1 *Consider a sequence of random variables* $\{(X_n,Y_n),n\ge 1\}$ *in the product space* $\mathbf{E}_1\times\mathbf{E}_2$. *If* $X_n\Rightarrow X$, *then there exist other random elements of* $\mathbf{E}_1\times\mathbf{E}_2$, $\{(\tilde{X}_n,\tilde{Y}_n),n\ge 1\}$, *and* $\tilde{X}$, *defined on a common underlying probability space*, *such that*

$$(\tilde{X}_n,\tilde{Y}_n)\stackrel{d}{=}(X_n,Y_n),\quad n\ge 1,\qquad \tilde{X}\stackrel{d}{=}X$$

and almost surely,

$$\tilde{X}_n\to\tilde{X}\quad \textit{as } n\to\infty.$$

Proof In order to present the proof, we first need some preliminaries. A nested family of countably partitions of a set A is a collection of subsets $A_{i_1,\dots,i_k}$ indexed by k-tuples of positive integers such that $\{A_i:i\ge 1\}$ is a partition of A and $\{A_{i_1,\dots,i_{k+1}}:$

$i_{k+1} \geq 1\}$ is a partition of $A_{i_1,\ldots,i_k}$ for all $k \geq 1$ and $(i_1, \ldots, i_k) \in \mathbb{N}_+^k$. Let $\mathbb{P}_1$ denote the probability measure on the space where X lives on. Since the space $(\mathbf{E}_1, \pi_1)$ is separable, according to Lemma 1.9 in the supplement of [29], there exists a nested family of countably partitions $\{E^1_{i_1,\ldots,i_k}\}$ of $(\mathbf{E}_1, \pi_1)$ that satisfies

$$\operatorname{rad}\bigl(E^1_{i_1,\ldots,i_k}\bigr) < 2^{-k}, \tag{C.1}$$

$$\mathbb{P}_1\bigl(\partial E^1_{i_1,\ldots,i_k}\bigr) = 0, \tag{C.2}$$

where $\operatorname{rad}(A)$ denotes the radius of the set A in a metric space, and $\partial(A)$ denote the boundary of the set A. Since the space $(\mathbf{E}_2, \pi_2)$ is separable, by the same lemma, there exists a nested sequence of countably partitions $\{E^2_{i'_1,\ldots,i'_{k'}}\}$ of $(\mathbf{E}_2, \pi_2)$ that satisfies

$$\operatorname{rad}\bigl(E^2_{i'_1,\ldots,i'_{k'}}\bigr) < 2^{-k'}. \tag{C.3}$$

Note that for space $(\mathbf{E}_2, \pi_2)$, we only need a weaker version of Lemma 1.9 in the supplement of [29].

The first step is to use this nested sequence of countably partitions to construct random variables $\{(\tilde{X}_n, \tilde{Y}_n), n \geq 1\}$ with the same distribution for each n. For $n \geq 1$, we first construct sub-intervals $I^n_{i_1,\ldots,i_k} \subseteq [0, 1)$ corresponding to the marginal probability of X_n. Let $I^n_1 = [0, \mathbb{P}^n(E^1_1 \times \mathbf{E}_2))$ and

$$I^n_i = \left[\sum_{j=1}^{i-1} \mathbb{P}^n\bigl(E^1_j \times \mathbf{E}_2\bigr), \sum_{j=1}^{i} \mathbb{P}^n\bigl(E^1_j \times \mathbf{E}_2\bigr)\right), \quad i > 1,$$

where $\mathbb{P}^n$ is the probability measure on the space where (X_n, Y_n) lives. Let $\{I^n_{i_1,\ldots,i_{k+1}} : i_{k+1} \geq 1\}$ be a countable partition of sub-intervals of $I^n_{i_1,\ldots,i_k}$. If $I^n_{i_1,\ldots,i_k} = [a_n, b_n)$, then

$$I^n_{i_1,\ldots,i_{k+1}} = \left[a_n + \sum_{j=1}^{i_{k+1}-1} \mathbb{P}^n\bigl(E^1_{i_1,\ldots,i_k,j} \times \mathbf{E}_2\bigr), a_n + \sum_{j=1}^{i_{k+1}} \mathbb{P}^n\bigl(E^1_{i_1,\ldots,i_k,j} \times \mathbf{E}_2\bigr)\right).$$

The length of each sub-interval $I^n_{i_1,\ldots,i_k}$ is the probability $\mathbb{P}^n(E^1_{i_1,\ldots,i_k} \times \mathbf{E}_2)$. We then construct further sub-intervals $I^n_{i_1,\ldots,i_k;i'_1,\ldots,i'_{k'}} \subseteq I^n_{i_1,\ldots,i_k}$ corresponding to (X_n, Y_n). If $I^n_{i_1,\ldots,i_k} = [a_n, b_n)$, then let $I^n_{i_1,\ldots,i_k;1} = [a_n, a_n + \mathbb{P}^n(E^1_{i_1,\ldots,i_k} \times E^2_1))$ and

$$I^n_{i_1,\ldots,i_k;i'} = \left[a_n + \sum_{j'=1}^{i'-1} \mathbb{P}^n\bigl(E^1_{i_1,\ldots,i_k} \times E^2_{j'}\bigr), a_n + \sum_{j'=1}^{i'} \mathbb{P}^n\bigl(E^1_{i_1,\ldots,i_k} \times E^2_{j'}\bigr)\right), \quad i' > 1.$$

Let $\{I^n_{i_1,\ldots,i_k;i'_1,\ldots,i'_{k'+1}} : i'_{k'+1} \geq 1\}$ be countable partition of $I^n_{i_1,\ldots,i_k;i'_1,\ldots,i'_{k'}}$. If $I^n_{i_1,\ldots,i_k;i'_1,\ldots,i'_{k'}} = [a_n, b_n)$, then

$$I^n_{i_1,\dots,i_k;i'_1,\dots,i'_{k'+1}}$$
$$= \left[a_n + \sum_{j'=1}^{i'_{k'+1}-1} \mathbb{P}^n\bigl(E^1_{i_1,\dots,i_k} \times E^2_{i'_1,\dots,i'_k,j'}\bigr), a_n + \sum_{j'=1}^{i'_{k'+1}} \mathbb{P}^n\bigl(E^1_{i_1,\dots,i_k} \times E^2_{i'_1,\dots,i'_k,j'}\bigr)\right).$$

The length of each sub-interval $I^n_{i_1,\dots,i_k;i'_1,\dots,i'_{k'}}$ is the probability $\mathbb{P}^n(E^1_{i_1,\dots,i_k} \times E^2_{i'_1,\dots,i'_{k'}})$. Now from each non-empty subset $E^1_{i_1,\dots,i_k} \times E^2_{i'_1,\dots,i'_k}$ we choose one point $(x_{i_1,\dots,i_k}, y_{i'_1,\dots,i'_k})$. For each $n \geq 1$ and $k \geq 1$, we define functions $(x^k_n, y^k_n) : [0, 1) \to \mathbf{E}_1 \times \mathbf{E}_2$ by letting $x^k_n(w) = x_{i_1,\dots,i_k}$ and $y^k_n(w) = y_{i'_1,\dots,i'_k}$ for $\omega \in I^n_{i_1,\dots,i_k;i'_1,\dots,i'_k}$. By the nested partition property and inequalities C.1 and C.3,

$$\pi\bigl(\bigl(x^k_n(\omega), x^k_n(\omega)\bigr), \bigl(x^{k+j}_n(\omega), x^{k+j}_n(\omega)\bigr)\bigr) < 2^{-k} \quad \text{for all } j, k, n$$

and $\omega \in [0, 1)$. Since $(\mathbf{E}_1 \times \mathbf{E}_2, \pi)$ is a complete metric space, the above implies that there is $(x_n(\omega), y_n(\omega)) \in \mathbf{E}_1 \times \mathbf{E}_2$ such that

$$\pi\bigl(\bigl(x^k_n(\omega), x^k_n(\omega)\bigr), \bigl(x_n(\omega), x_n(\omega)\bigr)\bigr) \to 0 \quad \text{as } k \to \infty.$$

We let $(\tilde{X}_n, \tilde{Y}_n) = (x_n, y_n)$ on $[0, 1)$ for $n \geq 0$.

The next step is to construct $\tilde{X}$ and show that $\tilde{X}_n \to \tilde{X}$ almost surely. For each $n \geq 1$, let $\mathbb{P}^n_1$ denote the marginal probability of X^n. It is clear that $I^n_{i_1,\dots,i_k}$ is the probability $\mathbb{P}^n_1(E^1_{i_1,\dots,i_k})$. By (C.2), we have $\mathbb{P}^n_1(E^1_{i_1,\dots,i_k}) \to \mathbb{P}_1(E^1_{i_1,\dots,i_k})$, as $n \to \infty$. Consequently, the length of the interval $I^n_{i_1,\dots,i_k}$ converges to the length of the interval $I_{i_1,\dots,i_k}$, which is defined in a similar way as for $I^n_{i_1,\dots,i_k}$ by letting

$$I_{i_1,\dots,i_{k+1}} = \left[a_n + \sum_{j=1}^{i_{k+1}-1} \mathbb{P}_1(E_{i_1,\dots,i_k,j}), a_n + \sum_{j=1}^{i_{k+1}} \mathbb{P}_1(E_{i_1,\dots,i_k,j})\right),$$

if $I_{i_1,\dots,i_k} = [a_n, b_n)$. Now from each non-empty subset $E_{i_1,\dots,i_k}$ we choose one point $x_{i_1,\dots,i_k}$. For each $k \geq 1$, we define functions $x^k : [0, 1) \to \mathbf{E}_1$ by letting $x^k(\omega) = x_{i_1,\dots,i_k}$ for $\omega \in I^n_{i_1,\dots,i_k}$. By the nested partition property and inequalities C.1,

$$\pi_1\bigl(x^k(\omega), x^{k+j}(\omega)\bigr) < 2^{-k} \quad \text{for all } j, k$$

and $\omega \in [0, 1)$. Since $(\mathbf{E}_1, \pi_1)$ is a complete metric space, the above implies that there is $x(\omega) \in \mathbf{E}_1$ such that

$$\pi_1\bigl(x^k(\omega), x(\omega)\bigr) \to 0 \quad \text{as } k \to \infty.$$

We let $\tilde{X} = x$ on $[0, 1)$. Since

$$\begin{aligned}\pi_1\bigl(\tilde{X}_n(\omega), \tilde{X}(\omega)\bigr) &\leq \pi_1\bigl(\tilde{X}_n(\omega), \tilde{X}^k_n(\omega)\bigr) + \pi_1\bigl(\tilde{X}^k_n(\omega), \tilde{X}^k(\omega)\bigr) + \pi_1\bigl(\tilde{X}^k(\omega), \tilde{X}(\omega)\bigr)\\ &\leq 3 \times 2^{-k},\end{aligned}$$

for all ω in the interior of $I_{i_1,\dots,i_k}$,

$$\lim_{n\to\infty} \pi_1\big(\tilde{X}_n(\omega), \tilde{X}(\omega)\big) \le 3 \times 2^{-k}.$$

Since k is arbitrary, we must have $\tilde{X}_n(\omega) \to \tilde{X}(\omega)$ as $n \to \infty$ for all but at most countably many $\omega \in [0, 1)$.

It remains to show that $(\tilde{X}_n, \tilde{Y}_n)$ has the probability laws $\mathbb{P}^n$. Let $\tilde{\mathbb{P}}$ denote the Lebesgue measure on $[0, 1)$. It suffices to show that $\tilde{\mathbb{P}}((\tilde{X}_n, \tilde{Y}_n) \in A) = \mathbb{P}^n(A)$ for each A such that $\mathbb{P}^n(\partial A) = 0$. Let A be such a set. Let A^k be the union of the sets $E^1_{i_1,\dots,i_k} \times E^2_{i'_1,\dots,i'_k}$ such that $E^1_{i_1,\dots,i_k} \times E^2_{i'_1,\dots,i'_k} \subseteq A$ and let A'^k be the union of the sets $E^1_{i_1,\dots,i_k} \times E^2_{i'_1,\dots,i'_k}$ such that $E^1_{i_1,\dots,i_k} \times E^2_{i'_1,\dots,i'_k} \cap A \neq \emptyset$. Then $A^k \subseteq A \subseteq A'^k$ and, by the construction above,

$$\tilde{\mathbb{P}}\big((\tilde{X}_n, \tilde{Y}_n) \in A^k\big) = \mathbb{P}^n\big(A^k\big) \quad \text{and} \quad \tilde{\mathbb{P}}\big((\tilde{X}_n, \tilde{Y}_n) \in A'^k\big) = \mathbb{P}^n\big(A'^k\big).$$

Now let $C^k = \{s \in \mathbf{E}_1 \times \mathbf{E}_2 : \pi(s, \partial A) \le 2^{-k}\}$. Then $A'^k - A^k \downarrow \partial A$ as $k \to \infty$. Since $\mathbb{P}^n(\partial A) = 0$ by assumption, $\mathbb{P}^n(C^k) \downarrow 0$ as $k \to \infty$. Hence

$$\tilde{\mathbb{P}}\big((\tilde{X}_n, \tilde{Y}_n) \in A\big) = \lim_{k\to\infty} \tilde{\mathbb{P}}\big((\tilde{X}_n, \tilde{Y}_n) \in A^k\big) = \lim_{k\to\infty} \mathbb{P}^n\big(A^k\big) = \mathbb{P}^n(A).$$

In the same way, we can show that $\tilde{X}$ has probability law $\mathbb{P}_1$. □

References

1. Asmussen, S.: Applied Probability and Queues, 2nd edn. Applications of Mathematics (New York), vol. 51. Springer, New York (2003)
2. Billingsley, P.: Convergence of Probability Measures, 2nd edn. Wiley Series in Probability and Statistics: Probability and Statistics. Wiley, New York (1999)
3. Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., Zhao, L.: Statistical analysis of a telephone call center: a queueing-science perspective. J. Am. Stat. Assoc. **100**(469), 36–50 (2005)
4. Chung, K.L.: A Course in Probability Theory, 3rd edn. Academic Press, San Diego (2001)
5. Dai, J.G., He, S., Tezcan, T.: Many-server diffusion limits for $G/Ph/n + GI$ queues. Ann. Appl. Probab. **20**(5), 1854–1890 (2010)
6. Dudley, R.M.: Real Analysis and Probability. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge (2002)
7. Ethier, S.N., Kurtz, T.G.: Markov Processes. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. Wiley, New York (1986)
8. Gamarnik, D., Momčilović, P.: Steady-State Analysis of a Multi-Server Queue in the Halfin–Whitt Regime (2007)
9. Garnett, O., Mandelbaum, A., Reiman, M.: Designing a call center with impatient customers. Manuf. Serv. Oper. Manag. **4**(3), 208–227 (2002)
10. Gromoll, H.C., Kruk, Ł.: Heavy traffic limit for a processor sharing queue with soft deadlines. Ann. Appl. Probab. **17**(3), 1049–1101 (2007)
11. Gromoll, H.C., Puha, A.L., Williams, R.J.: The fluid limit of a heavily loaded processor sharing queue. Ann. Appl. Probab. **12**(3), 797–859 (2002)
12. Gromoll, H.C., Robert, P., Zwart, B.: Fluid limits for processor sharing queues with impatience. Math. Oper. Res. **33**(2), 375–402 (2008)
13. Halfin, S., Whitt, W.: Heavy-traffic limits for queues with many exponential servers. Oper. Res. **29**(3), 567–588 (1981)
14. Hunter, J.K., Nachtergaele, B.: Applied Analysis. World Scientific, River Edge (2001)

15. Jelenković, P., Mandelbaum, A., Momčilović, P.: Heavy traffic limits for queues with many deterministic servers. Queueing Syst. **47**(1–2), 53–69 (2004)
16. Kallenberg, O.: Random Measures, 4th edn. Akademie-Verlag, Berlin (1986)
17. Kang, W., Ramanan, K.: Fluid limits of many-server queues with reneging. Ann. Probab. **20**(6), 2204–2260 (2010)
18. Kaspi, H., Ramanan, K.: Law of large numbers limits for many-server queues. Ann. Appl. Probab. **21**(1), 33–114 (2011)
19. Kaspi, H., Ramanan, K.: (2011). SPDE limits of many-server queues. Tech. Rep., Technion and Carnegie Mellon University
20. Lang, S.: Real Analysis, 2nd edn. Addison-Wesley Publishing Company Advanced Book Program, Reading (1983)
21. Mandelbaum, A., Momčilović, P.: Queues with many servers: the virtual waiting-time process in the QED regime. Math. Oper. Res. **33**(3), 561–586 (2008)
22. Mandelbaum, A., Momčilović, P.: Queues with many servers and impatient customers. Math. Oper. Res. **37**(1), 41–65 (2012)
23. Mandelbaum, A., Shimkin, N.: A model for rational abandonments from invisible queues. Queueing Syst. **36**(1–3), 141–173 (2000)
24. Pang, G., Whitt, W.: Service interruptions in large-scale service systems. Manag. Sci. **55**(9), 1499–1512 (2009)
25. Puhalskii, A.: The $M_t/M_t/K_t + M_t$ Queue in Heavy Traffic (2008)
26. Puhalskii, A.A., Reed, J.E.: On many-server queues in heavy traffic. Ann. Appl. Probab. **20**(1), 129–195 (2010)
27. Puhalskii, A.A., Reiman, M.I.: The multiclass $GI/PH/N$ queue in the Halfin–Whitt regime. Adv. Appl. Probab. **32**(2), 564–595 (2000)
28. Reed, J.E.: The $G/GI/N$ queue in the Halfin–Whitt regime. Ann. Appl. Probab. **19**(6), 2211–2269 (2009)
29. Whitt, W.: Stochastic-Process Limits. Springer Series in Operations Research. Springer, New York (2002)
30. Whitt, W.: Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. Manag. Sci. **50**(10), 1449–1461 (2004)
31. Whitt, W.: Fluid models for multiserver queues with abandonments. Oper. Res. **54**(1), 37–54 (2006)
32. Zeltyn, S., Mandelbaum, A.: Call centers with impatient customers: many-server asymptotics of the $M/M/n + G$ queue. Queueing Syst. **51**(3–4), 361–402 (2005)
33. Zhang, J., Dai, J.G., Zwart, B.: Law of large number limits of limited processor-sharing queues. Math. Oper. Res. **34**(4), 937–970 (2009)
34. Zhang, J., Dai, J.G., Zwart, B.: Diffusion limits of limited processor-sharing queues. Ann. Appl. Probab. **21**(2), 745–799 (2011)