Operations Research Letters 49 (2021) 338–344

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

A many-server functional strong law for a non-stationary loss model

Prakash Chakraborty [a,1], Harsha Honnappa [b,*]

[a] Department of Statistics, Purdue University, West Lafayette IN, USA
[b] School of Industrial Engineering, Purdue University, West Lafayette IN, USA

ARTICLE INFO

Article history:
Received 23 December 2019
Received in revised form 4 October 2020
Accepted 8 March 2021
Available online 20 March 2021

Keywords:
Fluid limit
Many-server
Loss model
Non-stationary

ABSTRACT

The purpose of this note is to show that it is possible to establish a many-server functional strong law of large numbers (FSLLN) for the fraction of occupied servers (i.e., the scaled number-in-system) without explicitly tracking through a measure valued process either the age or the residual service times of the jobs in a non-Markovian, non-stationary loss model. This considerable analytical simplification is achieved by exploiting a semimartingale representation. The fluid limit is shown to be the unique solution of a Volterra integral equation.

1. Introduction

This note establishes a functional strong law of large numbers (FSLLN) for a non-Markovian, non-stationary $M_t/G/n/n$ loss model in the many-server limit as $n \to \infty$. Stationary loss models have been studied extensively, with the Erlang-B formula being a cornerstone consequence of this literature. Non-stationary models, on the other hand, are much harder to analyze, and closed form expressions are almost impossible to derive. Stochastic process approximations are therefore crucial for performance analysis of loss models. There is a significant body of work focused on establishing fluid approximations in many-server settings, though the methods are non-trivial. In the formative paper [8], the elapsed waiting time (or 'age') of the jobs in the system are tracked, using which it is possible to obtain a martingale representation of the number-in-system process that yields the desired FSLLN for a $G/GI/n/\infty$ queue. In contrast, [22] develops a method where the residual service times of jobs in the $G/GI/n/\infty$ queue are tracked, in which case it is possible to establish the fluid limit without recourse to a martingale representation. Of course, while the latter approach in essence assumes that the service times are known at the time of arrival, as commented on in [22] the approach offers significant analytical simplification.

The purpose of this note is to show that it is possible to establish a many-server FSLLN for the fraction of occupied servers (i.e., the scaled number-in-system) in a $M_t/G/n/n$ loss model, without explicitly tracking through a measure-valued process either the age or the residual service times of the jobs. We capitalize on the fact that this process is the sum of a pure jump bounded semimartingale and a bounded finite variation process. Indeed, we show that in the many server limit the fraction of occupied servers converges to the solution of a non-linear Volterra integral equation by exploiting the fact that the semimartingale is uniformly zero and the bounded finite variation process converges to a deterministic function. This semimartingale representation is natural and allows us to avoid tracking the residual service times. Our proofs are also considerably simpler and easier to follow. Consequently, we anticipate that our analysis will be intuitive and useful for a broad range of applications. For instance, as a consequence of our main result, we present a fluid limit for the fraction of arrivals that are blocked. This result can be used as a proxy for the blocking probability in the many-server limit. A crucial motivation for this paper is the need to develop 'transitory fluid' traffic models; i.e., systems where a finite volume of jobs (in a continuum) enter a system over time. Queueing models fed by this type of traffic have been studied in [2,5,7] – however all of these consider discrete-event models of traffic. Transitory *fluid* traffic models have not been studied in the literature, and would be of considerable use in the modeling of capacitated energy storage systems and high-speed computer networks. As an auxiliary result, therefore, we also establish a FSLLN for the integrated fraction of occupied servers. This process is a non-decreasing stochastic fluid with a maximum rate of increase. This type of model can be used to model the energy production from a solar array, for instance.

While our proof of the main result is not complicated, some commentary is in order. We consider a sequence of $M_t/G/n/n$ models with nonstationary Poisson traffic with deterministic intensity $(\lambda^n(t) : t \geq 0)$ where $\lambda^n(\cdot) = n\lambda(\cdot)$ and stationary

* Corresponding author.
E-mail addresses: cprakash@umich.edu (P. Chakraborty), honnappa@purdue.edu (H. Honnappa).
[1] Present Affiliation: Department of Mathematics, University of Michigan.

https://doi.org/10.1016/j.orl.2021.03.004
0167-6377/

general service times with finite first moment. We assume that the traffic and service processes are statistically independent of each other for every $n \geq 1$. Thus, the number of servers (i.e., the system capacity) is in scale with the arrival intensity. Using the fact that the fraction of occupied servers in the nth system can be represented as a stochastic integral with respect to a random counting measure, we extract the desired semimartingale representation. In Theorem 3.4 we show that the fraction of occupied servers converges to a deterministic limit. Identifying the limit function itself turns out to be a little tricky, owing to the fact that the fraction of occupied servers is the solution of a discontinuous stochastic integral equation. In order to identify the limit, we smooth the representation of the process by using a mollifier of the discontinuity. This allows us to identify the limit function as the solution of a specific non-linear Volterra integral equation. Finally, to establish uniqueness of the limit, we exploit the fact that the first time that the limit function hits the level 1 (i.e., the system fluid level is full) is unique, from which it follows recursively that the first (and subsequent) times the limit leaves the fully occupied level and/or (re)enters state 1 are unique.

Related literature. There is a significant body of work establishing many-server fluid limits for stationary and non-stationary models, both with and without abandonment, starting with the seminal work in [6]; see [20] for a recent survey. Our work is related to the development of proof techniques for many-server limits, and to work on approximations to nonstationary loss models. As noted before, Kaspi and Ramanan established a fluid limit for the number-in-system process in the formative paper [8], using a martingale representation of extracted using the elapsed waiting time or age of the jobs in the system. [17] on the other hand established a fluid (and diffusion) limit for the number-in-system process of a stationary $G/GI/n$ queue by using a representation of the number in system process that is similar to the system equations of a $G/GI/\infty$ queue. By establishing a link between the equations, [17] was able to prove both a FSLLN and a functional central limit theorem (FCLT). Our approach is similar, in the sense that we exploit a random measure representation of the number-in-system process akin to system state representations in infinite server queues. Note that since we focus on nonstationary loss models, our representation is different from that of [17]. While the analysis of models without abandonment are most relevant to our setting, [22] analyzed the number in system process of a $G/GI/n + GI$ queue by tracking the residual service times. In the nonstationary setting, in a series of papers [10–12] Liu and Whitt proved a fluid limit for a $G_t/GI/n + GI$ queue that experiences alternating periods of overload and underload, by tracking the age of the jobs in the system *a la* [8]. More broadly, there has been a growing body of work on nonstationary loss models and various approximations, particularly for computing blocking probabilities [13–15,20,21]. Our results complement these works by providing fluid limits that characterize the fraction of arrivals that encounter a blocked system.

2. Preliminaries

In this section we present some preliminary results that will be useful later on.

2.1. Right continuous functions

Let $D = D[0,T]$ denote the space of right continuous functions on $[0,T]$ that have left limits. For a function $f \in D$ and $T_0 \subset [0,T]$, let $w_f(T_0) = \sup\{|f(t) - f(s)| : s,t \in T_0\}$. and for $\delta \in (0,T)$, let $w'_f(\delta) = \inf_{\mathcal{P}:\|\mathcal{P}\|\leq\delta} \max_{0<i\leq|\mathcal{P}|} w_f([t_{i-1},t_i))$, where $\mathcal{P}$ runs over the set of all partitions of $[0,T]$, in the sense that a generic $\mathcal{P}$ is defined as $\mathcal{P} = \{0 = t_0, \ldots, t_{|\mathcal{P}|} = T\}$, and $\|\mathcal{P}\|$ denotes the mesh or norm of the partition $\mathcal{P}$: $\|\mathcal{P}\| = \max_{1\leq i<|\mathcal{P}|} |t_i - t_{i-1}|$. It can be shown that a function f lies in D if and only if $\lim_{\delta\downarrow 0} w'_f(\delta) = 0$. The proof of this result and related discussion can be found in [3, Ch. 14]. The Skorohod distance between two functions f and g in D is defined by

$$d_S(f,g) = \inf\Big\{ \varepsilon > 0 : \exists \text{ strictly increasing function } \lambda : [0,T] \mapsto [0,T], \text{ and } \sup_{t\in[0,T]} |\lambda(t) - t| \leq \varepsilon, \sup_{t\in[0,T]} |f(\lambda(t)) - g(t)| \leq \varepsilon \Big\}.$$

The topology induced on D by the Skorohod distance is the Skorohod topology.

Theorem 2.1. *A set $A \subset D[0,T]$ has compact closure in the Skorohod topology if and only if $\sup_{f\in A} \sup_{t\in[0,T]} |f(t)| < \infty$, and $\lim_{\delta\downarrow 0} \sup_{f\in A} w'_f(\delta) = 0$.*

Remark 2.2. It can be shown that D is not a complete space with respect to the Skorohod distance d_S but there exists a topologically equivalent metric d_0 with respect to which D is complete.

2.2. Counting measure

Let $(\Omega, \mathcal{F}, \boldsymbol{\mathcal{F}} = (\mathcal{F}_t)_{t\geq 0}, P)$ be a filtered probability space. Let $(N_t)_{t\geq 0}$ be a point process given by a sequence $(T_n)_{n\geq 0}$ of jump times, that is $N_t := \sum_{i=1} \mathbf{1}_{\{T_i\leq t\}}$. Suppose in addition the n^{th} jump time or arrival T_n has a corresponding random variable Z_n taking values in some measurable space $(E,\mathcal{E})$. Then $(T_n,Z_n)_{n\geq 1}$ is called an E-marked point process. For each $A \in \mathcal{E}$, let the counting process $N_t(A)$ be given by $N_t(A) := \sum_{i=1}^{\infty} \mathbf{1}_{\{Z_n\in A\}}\mathbf{1}_{\{T_n\leq t\}}$, and the corresponding counting measure $p(dt\times dz)$ by $p(\omega,(0,t]\times A) = N_t(\omega,A)$. This means that for functions $H : \Omega\times[0,\infty)\times\mathbb{R} \mapsto \mathbb{R}$

$$\int_0^t \int_{\mathbb{R}} H(\omega,u,x)p(\omega,du\times dx) = \sum_{i=1}^{\infty} H(\omega,T_i(\omega),Z_i(\omega))\mathbf{1}_{\{T_i(\omega)\leq t\}}. \tag{1}$$

For a point process $(N_t)_{t\geq 0}$, its intensity with respect to a given filtration $(\mathcal{F}_t)_{t\geq 0}$ is given by $\lambda_t = \lim_{\delta\downarrow 0} P(N(t+\delta t) - N(t)|\mathcal{F}_t)$, $t > 0$. If $(Z_n)_{n\geq 1}$ and $(T_n)_{n\geq 1}$ are independent, and $(Z_n)_{n\geq 1}$ are independent and identically distributed (iid) from a distribution with density ν, then it is easy to see that the intensity of the marked point process $N_t(A)$ for some $A \in \mathcal{E}$ is given by $\lambda_t(A) = \lambda_t\nu(A)$. We now say that $p(dt\times dz)$ admits the intensity kernel $\lambda_t\nu(dz)$. Let $\mathcal{P}(\boldsymbol{\mathcal{F}})$ denote the predictable σ-field on $\Omega\times(0,\infty)$. Then for any mapping $H : \Omega\times(0,\infty)\times E \mapsto \mathbb{R}$, measurable with respect to $\mathcal{P}(\boldsymbol{\mathcal{F}})\otimes\mathcal{E}$ satisfies the following projection result (cf. [4, T3 Theorem, pp 235])

$$E\left[\int_0^{\infty}\int_E H(s,z)p(ds\times dz)\right] = E\left[\int_0^{\infty}\int_E H(s,z)\lambda_s\nu(dz)ds\right]. \tag{2}$$

Thus defining the compensated measure $q(ds\times dz) = p(ds\times dz) - \lambda_s\nu(dz)ds$, we have for every H as in (2) that $\int_0^t\int_E H(s,z)q(ds\times dz)$ is a $(P,\mathcal{F}_t)$ local martingale.

A final note: we denote convergence in probability by $\xrightarrow{p}$, convergence uniformly on compact intervals and in probability by $\xrightarrow{ucp}$.

3. Model and results

3.1. Description of model

We now introduce our model along with a useful representation of our main quantity of interest.

Assumption 3.1. Consider a $M_t/G/n/n$ loss model; namely, a queueing model with

i. a non-homogeneous Poisson arrival process A^n with rate $n\lambda$, where λ is locally integrable;
ii. general service times sampled iid from a distribution F with density ν; and,
iii. n servers and zero buffer.

Let $\mathbf{R} = (R_i)_{i\geq 1}$ be the marked process where $R_i = (T_i, S_i)$, T_i are the arrival time epochs corresponding to the arrival process A^n, and S_i denotes the corresponding service time sampled iid from F. We assume that relevant random variables for every n sit in a common filtered probability space $(\Omega, \mathcal{F}, \boldsymbol{\mathcal{F}} = (\mathcal{F}_t)_{t>0}, \mathbf{P})$. Let $p^R(du, dx)$ denote the counting measure associated with the process R. Recall from (1), $p^R(du, dx)$ is a random measure on $[0,\infty)\times\mathbb{R}^+$ such that for functions $W : \Omega \times [0,\infty)\times\mathbb{R} \mapsto \mathbb{R}$ we have:

$$\int_0^t \int_{\mathbb{R}} W(\omega; u, x)p^R(du, dx) = \sum_{i=1}^{\infty} W(\omega; T_i(\omega), S_i(\omega))\mathbf{1}_{\{T_i(\omega)\leq t\}}. \quad (3)$$

Moreover since $\lambda_t(A) = \lambda_t\nu(A)$, p^R is a random measure with intensity $p^R_c(du, dx) = n\lambda_u\nu(x)dudx$. Denote p^R_* to be the compensated random measure:

$$p^R_* = p^R - p^R_c. \quad (4)$$

Remark 3.2. We explain the need for Poisson arrival processes in our considerations. In the sequel we would need finiteness of the second moment of stochastic integrals of bounded predictable processes with respect to the compensated measure, that is $\mathbf{E}\left(\int_0^t \int_{\mathbb{R}} W(u, x)p^*_R(du, dx)\right)^2 < \infty$ for a bounded predictable process W. This is well established when p^R_* results from Poisson arrivals. However, we note that this is the only crucial requirement and all the results stated in this article hold true for any arrival process satisfying this second moment condition.

3.2. Fraction of occupied servers

Let ρ^n_t denote the fraction of occupied servers at time t. Observe that the number of busy servers at time t is the cumulative sum of arrivals at times u, $u \in [0, t]$ satisfying:

(i) the number of occupied servers at time u is less than n.
(ii) the corresponding service requirement exceeds $t - u$.

Consequently we have:

$$\rho^n_t = \frac{1}{n}\sum_{i=1}^{\infty} \mathbf{1}_{\{\rho^n_{T_i-}<1\}}\mathbf{1}_{\{S_i>t-T_i\}}\mathbf{1}_{\{T_i\leq t\}} \quad (5)$$

Using (3), the right hand side of (5) can be expressed as a stochastic integral with respect to the counting measure p^R:

$$\rho^n_t = \int_0^t \int_{\mathbb{R}} W_n(t, u, x)p^R(du, dx), \quad (6)$$

where $W_n(t, u, x) = \frac{1}{n}\mathbf{1}_{\{\rho^n_{u-}<1\}}\mathbf{1}_{\{u<t\}}\mathbf{1}_{\{x>t-u\}}$, is a predictable process.

Remark 3.3. Note that ρ^n has paths of finite variation on compacts. Indeed, the process ρ^n is piecewise constant with jumps corresponding to arrivals according to A^n only if the current state ρ^n is less than one. This means that the total variation of ρ^n is bounded by that of A^n which is a non-homogeneous Poisson process and hence is of finite variation. In addition, ρ^n is adapted and càdlàg, and consequently by [16, Theorem 26] ρ^n is a pure jump quadratic semimartingale.

We now state a functional fluid limit for ρ^n as n tends to infinity.

Theorem 3.4. *Let the conditions in Assumption 3.1 hold. Then for any $T > 0$ we have:*

$$\lim_{n\to\infty} \sup_{t\in[0,T]} \left|\rho^n_t - \rho_t\right| = 0, \text{ almost surely,} \quad (7)$$

where ρ is the solution to a non-linear Volterra integral equation:

$$\rho_t = \int_0^t \mathbf{1}_{\{\rho_{u-}<1\}}\bar{F}(t-u)\lambda_u du \quad \text{for } t > 0 \quad \text{and } \rho_0 = 0. \quad (8)$$

Proof. Recall that the counting measure p^R possesses a compensator p^R_c. Now, observe that from (5) and (4) the fraction of occupied servers ρ^n has the following decomposition:

$$\begin{aligned}\rho^n_t &= \int_0^t \int_{\mathbb{R}} W_n(t, u, x)p^R_*(du, dx) + \int_0^t \int_{\mathbb{R}} W_n(t, u, x)p^R_c(du, dx)\\ &= \int_0^t \int_{\mathbb{R}} W_n(t, u, x)p^R_*(du, dx) + \int_0^t \mathbf{1}_{\{\rho^n_{u-}<1\}}\bar{F}(t-u)\lambda_u du\\ &=: X^n_t + \gamma^n_t,\end{aligned}$$

where p^R_* is given by (4) and X^n is another bounded cadlag semimartingale. Indeed, X^n is the difference of a bounded cadlag semimartingale ρ^n and the bounded finite variation process γ^n. In fact, we have $\sup_n \sup_t |X^n_t| \leq 1 + \int_0^T \lambda_u du$, where the second quantity is finite because of our assumption that λ is locally integrable according to Assumption 3.1. In addition we have almost surely $\lim_{\delta\downarrow 0}\sup_n w'_{X^n}(\delta) = 0$, where $w'_x(\delta) = \inf_{t_i} w_x[t_i, t_{i+1})$. This follows from the facts that $w'_{\rho^n}(\delta) = 0$ (almost surely ρ^n has finitely many jumps in $[0, T]$ and is constant in between), and the fact that for all n one must have $\lim_{\delta\downarrow 0}\sup_n w_{\gamma^n}(\delta) = 0$. In order to obtain this last assertion observe that $|\gamma^n_t - \gamma^n_s| \leq 2\int_s^t \lambda_u du \leq 2w_\Lambda(|s-t|)$ where $\Lambda(t) = \int_0^t \lambda_u du$ is continuous on $[0, T]$ and hence also uniformly continuous.

We thus have that $\{X^n\}_{n\geq 1}$ has compact closure in the Skorohod topology. In other words we have obtained tightness.

Next, fix t and obtain (cf. [1, Theorem 2.3.7]):

$$\begin{aligned}\mathbf{E}\left(X^n_t\right)^2 &= \mathbb{E}\left[\int_0^t \int_{\mathbb{R}} W^2_n(t, u, x)p^R_*(du, dx)\right] \leq \frac{1}{n^2}\int_0^t \int_{\mathbb{R}} p^R_c(du, dx)\\ &= \frac{1}{n}\int_0^t \lambda_u du \leq \frac{1}{n}\Lambda \longrightarrow 0.\end{aligned}$$

Consequently for each $t \in [0, T]$, $X^n_t \xrightarrow{p} 0$. Thus for any $(t_1, t_2, \ldots, t_d) \in [0, T]^d$, the finite dimensional vectors $(X^n_{t_1}, \ldots, X^n_{t_n}) \xrightarrow{p} (0, \ldots, 0)$ as a consequence of the Cramer–Wold device [3, Theorem 7.7]. Recalling the tightness condition we have thus obtained that X^n converges in distribution to the constant zero function and hence also in probability. Since the limiting function is non-random, the convergence is also in probability under the uniform topology. Thus we have: $X^n \xrightarrow{ucp} 0$.

Observe that we have

$$\rho^n_t = X^n_t + \gamma^n_t = X^n_t + \int_0^t \mathbf{1}_{\{\rho^n_{u-}<1\}}\bar{F}(t-u)\lambda_u du. \quad (9)$$

For every $\omega \in \Omega$, ρ^n by definition belongs to the Skorohod space $D[0,T]$. Furthermore there exists $\Omega_1 \subset \Omega$ such that $P(\Omega_1) = 1$ and for every $\omega \in \Omega_1$ the sequence $\{\rho^n(\omega)\}_{n\geq 1}$ has compact closure in the Skorohod topology because (i) $\sup_n \sup_t \left|\rho_t^n\right| \leq 1$ and (ii) $\lim_{\delta\downarrow 0} \sup_n w'_{\rho^n}(\delta) = 0$, where $w'_x(\delta) = \inf_{t_i} w_x[t_i, t_{i+1})$, and w_x is the modulus of continuity of x in $[t_i, t_{i+1})$. Note that item (ii) is true almost surely because almost surely ρ will have finitely many jumps in the time horizon $[0,T]$.

Consider any $\omega \in \Omega_1$. The above considerations thus show that for every subsequence n_k of the naturals, ρ^{n_k} has a convergent subsequence which converges to an element of D. That is, there exists a subsequence $\{m_k\} \subset \{n_k\}$ such that

$$\rho^{m_k} \to \rho, \text{ in the Skorohod topology.} \tag{10}$$

Henceforth, we try to identify ρ. To that attempt, we give a slightly different representation of ρ^n. Observe that the set $\{\rho^n_{u-} < 1\}$ is identical to the set $\{\rho^n_{u-} \leq 1 - \frac{1}{n}\}$. This is because ρ^n only takes values in $\{\frac{i}{n} : i = 1, \ldots, n\}$. This gives us the opportunity to replace the indicator $\mathbf{1}_{\{\rho^n_{u-}<1\}}$ in (9) by a smooth approximation. In particular consider a sequence of smooth functions $\mathbf{1}^d : \mathbb{R} \to [0,1]$ for $d \in (0,1)$ such that

$$\mathbf{1}^d(x) = \begin{cases} 1, & \text{for } x \leq 1 - \frac{2d}{3}, \\ 0 & \text{for } x \geq 1 - \frac{d}{3}. \end{cases}$$

In addition let $\underline{\mathbf{1}}^d : \mathbb{R} \mapsto [0,1]$ for $d \in (0,1)$ be defined as:

$$\underline{\mathbf{1}}^d(x) = \begin{cases} 1 & \text{for } x \leq 1 - d \\ 0 & \text{for } x > 1 - d. \end{cases}$$

Using this notation we can replace $\mathbf{1}_{\{\rho^n_{u-}<1\}}$ in (9) by $\mathbf{1}^{\frac{1}{n}}(\rho^n_{u-})$ as both the quantities are the same. Thus our alternate representation of ρ^n is given by: $\rho^n_t = X^n_t + \int_0^t \mathbf{1}^{\frac{1}{n}}(\rho^n_{u-})\bar{F}(t-u)\lambda_u du$. Fix any arbitrary $t \in [0,T]$. Since $L^1[0,t]$ is a separable Banach space with dual $L^\infty[0,t]$ and $\mathbf{1}^{\frac{1}{m_k}}(\rho^{m_k})$ is bounded in $L^\infty[0,t]$, there is a subsequence $\{l_k\} \subset \{m_k\}$, where m_k is as in (10) such that

$$\lim_{k\to\infty} \int_0^t h(u)\mathbf{1}^{\frac{1}{l_k}}(\rho^{l_k}_{u-})du = \int_0^t h(u)w(u)du. \tag{11}$$

for every $h \in L^1[0,t]$. In particular, let us consider the function h given by: $h(u) = \bar{F}(t-u)\lambda_u$. Observe that our assumption on λ ensures that this specific h lies in $L^1[0,t]$. We thus have:

$$\lim_{k\to\infty} \int_0^t \mathbf{1}^{\frac{1}{l_k}}(\rho^{l_k}_{u-})\bar{F}(t-u)\lambda_u = \int_0^t w(u)\bar{F}(t-u)\lambda_u du,$$

Recall that $X^n \xrightarrow{ucp} 0$. Consequently there exists a subsequence $\{r_k\} \subset \{l_k\}$, (which we conveniently choose to be a subset of $\{l_k\}$) such that $\|X^{r_k}\|_T \longrightarrow 0$, for every $\omega \in \Omega_2$, where $\Omega_2 \subset \Omega_1$ satisfies $P(\Omega_2) = 1$. For this sequence $\{r_k\}$ we thus obtain:

$$\begin{aligned}\lim_{k\to\infty} \rho^{r_k}_t &= \lim_{k\to\infty} \left(X^{r_k}_t + \int_0^t \mathbf{1}^{\frac{1}{r_k}}(\rho^{r_k}_{u-})\bar{F}(t-u)\lambda_u\right) \\ &= \int_0^t w(u)\bar{F}(t-u)\lambda_u du.\end{aligned}$$

Due to (10) we must then have: $\rho_t = \int_0^t w(u)\bar{F}(t-u)\lambda_u du$. Observe that the above representation guarantees that ρ is continuous and the convergence stated in (10) is in the uniform topology for each $\omega \in \Omega_2$. Consequently fix $\varepsilon > 0$ and choose N large enough such that for all $k > N$ we have $r_k > \frac{3}{\varepsilon}$ and $\|\rho^{r_k} - \rho\|_T < \frac{\varepsilon}{3}$. Then it is readily checked that $\underline{\mathbf{1}}^\varepsilon(\rho_{u-}) \leq \mathbf{1}^{\frac{1}{r_k}}(\rho^{r_k}_{u-}) \leq \mathbf{1}_{\{\rho_{u-}<1\}}$. Consider any $h \geq 0$ such that $h \in L^1[0,T]$. Let us multiply each side of the above equation by h and integrate. We thus obtain:

$$\int_0^t h(u)\underline{\mathbf{1}}^\varepsilon(\rho_{u-})du \leq \int_0^t h(u)\mathbf{1}^{\frac{1}{r_k}}(\rho^{r_k}_{u-})du \leq \int_0^t h(u)\mathbf{1}_{\{\rho_{u-}<1\}}du.$$

Note that $\lim_{d\downarrow 0} \underline{\mathbf{1}}^d(x) = \lim_{d\downarrow 0} \mathbf{1}^d(x) = \mathbf{1}_{\{x<1\}}$. Consequently taking $k \to \infty$ (this is okay as we need $k > N = N(\varepsilon)$) and then $\varepsilon \downarrow 0$ we have by the dominated convergence theorem and (11) that:

$$\int_0^t h(u)\mathbf{1}_{\{\rho_{u-}<1\}}du \leq \int_0^t w(u)h(u)du \leq \int_0^t h(u)\mathbf{1}_{\{\rho_{u-}<1\}}du.$$

We now take $h = \bar{F}(t-\cdot)\lambda$ and thus we conclude that $\rho_t = \int_0^t \mathbf{1}_{\{\rho_{u-}<1\}}\bar{F}(t-u)\lambda_u du$. Observe that our considerations above hold for any $t \in [0,T]$ and hence (8) holds true for any $t \in [0,T]$. □

Remark 3.5. Our considerations so far as stated in Assumption 3.1 are constrained on systems which start empty, that is, $\rho^n_0 = 0$, for all n. However this is easily relaxed as stated in the following corollary which holds under the following assumption

Assumption 3.6. Let the conditions under Assumption 3.1 hold. In addition let the initial fraction of occupied servers ρ^n_0 satisfy the following asymptotic result:

$$\lim_{n\to\infty} \left|\rho^n_0 - r_0\right| = 0, \quad \text{almost surely,}$$

where $r_0 \in (0,1]$. Moreover, assume that the remaining service times for each of these occupied servers are iid drawn from a distribution G.

Remark 3.7. Assumption 3.6 can be extended to include more general initial distributions. In particular, one may consider the empirical distribution G^n of the remaining service times of the initial customers ρ^n_0 along with an assumption on its almost sure uniform convergence:

$$\lim_{n\to\infty} \sup_t \left|G^n(t) - G(t)\right| = 0, \quad \text{almost surely.}$$

Note that under this setup the remaining service times need not be independent or identically distributed. There are a number of classic results in probability theory establishing Glivenko–Cantelli type strong law results that imply this condition under different hypotheses. For instance, in the case where the remaining service times of the initial jobs are independent, but non-identically distributed [18] immediately implies the desired result. On the other hand, if the remaining service times form an ordered statistic (thereby introducing a weak form of dependence between them) then [19] provides a Glivenko–Cantelli theorem corresponding to this case.

Corollary 3.8. *Let the conditions in Assumption 3.6 hold. Then for any $T > 0$ we have:*

$$\lim_{n\to\infty} \sup_{t\in[0,T]} \left|\rho^n_t - \rho_t\right| = 0, \quad \text{almost surely,} \tag{12}$$

where ρ is the solution to a non-linear Volterra integral equation:

$$\rho_t = \rho_0\bar{G}(t) + \int_0^t \mathbf{1}_{\{\rho_{u-}<1\}}\bar{F}(t-u)\lambda_u du \quad \text{for } t > 0 \quad \text{and } \rho_0 = r_0. \tag{13}$$

Proof. Let the initial number of occupied servers be N^n_0, so that $\rho^n_0 = \frac{N^n_0}{n}$. Let the remaining service times for these N^n_0 many jobs be $(S^0_i)_{1\leq i\leq N^n_0}$. Then ρ^n_t can be represented as:

$$\rho^n_t = \frac{1}{n}\sum_{i=0}^{N^n_0} \mathbf{1}_{\{S^0_i > t\}} + \int_0^t \int_{\mathbb{R}} W_n(t,u,x)p^R(du,dx).$$

Observing that N_0^n goes to infinity since $r_0 > 0$, we can represent this as follows:

$$\rho_t^n = \sum_{i=0}^{N_0^n} \mathbf{1}_{\{S_i^0 > t\}} \frac{\rho_0^n}{N_0^n} + \int_0^t \int_{\mathbb{R}} W_n(t,u,x) p^R(du,dx).$$

We have already analyzed the second quantity on the right hand side in Theorem 3.4. Now using Assumption 3.6 we have that the first quantity converges almost surely. In particular, we have:

$$\lim_{n\to\infty} \left| \frac{\sum_{i=0}^{N_0^n} \mathbf{1}_{\{S_i^0 > t\}}}{N_0^n} \rho_0^n - \rho_0 \bar{G}(t) \right| = 0, \quad \text{almost surely.}$$

This completes the proof. □

Remark 3.9. Under the conditions of Remark 3.7 the fluid limit stays the same as in (12) but now G is the almost sure uniform limit of G^n.

3.3. Existence and uniqueness of fluid limit

In this subsection we prove the existence and uniqueness of the fluid limits ρ and Θ.

Theorem 3.10. *For all $r_0 \in [0,1]$ there exists a unique solution to the non-linear Volterra equation given by (13).*

Proof. Existence of a solution is well known and its proof is very similar to what we have presented in the proof to Theorem 3.4. Namely, we mollify the discontinuous coefficient $\mathbf{1}_{\{\cdot<1\}}$ by a smooth version, use existence results for smooth coefficients and then show that the limit satisfies (13). See [9] for the existence result in a more general setup, and for a more general definition of solution to nonlinear Volterra equations with discontinuous coefficient. Now we will show that for all $T > 0$, (13) has a unique solution for $t \in [0,T]$. We first show that ρ given by:

$$\rho_t = \rho_0 \bar{G}(t) + \int_0^t \mathbf{1}_{\{\rho_{u-}<1\}} \bar{F}(t-u)\lambda_u du,$$

takes values in $[0,1]$. The fact that ρ is positive is immediate by the positivity of ρ_0, $\bar{F}$, $\bar{G}$ and λ. Suppose there exists a $t_0 \in [0,T]$ such that $\rho_{t_0} > 1$. Since $\bar{G}$ is non-increasing and $\int_0^t \mathbf{1}_{\{\rho_{u-}<1\}} \bar{F}(t-u)\lambda_u du$ is continuous as a function of t, the jumps of ρ if any are negative. Thus there must exist an $s_0 < t_0$ such that $\rho_{s_0} = 1$ and $\rho_s \geq 1$ for $s \in (s_0, t_0]$. Consequently we must have:

$$\begin{aligned}\rho_{t_0} &= \rho_0 \bar{G}(t_0) + \int_0^{t_0} \mathbf{1}_{\{\rho_{u-}<1\}} \bar{F}(t_0-u)\lambda_u du \\ &= \rho_0 \bar{G}(t_0) + \int_0^{s_0} \mathbf{1}_{\{\rho_{u-}<1\}} \bar{F}(t_0-u)\lambda_u du.\end{aligned}$$

However, since $\bar{G}$ and $\bar{F}$ are non-increasing, we have:

$$\begin{aligned}&\rho_0 \bar{G}(t_0) + \int_0^{s_0} \mathbf{1}_{\{\rho_{u-}<1\}} \bar{F}(t_0-u)\lambda_u du \leq \rho_0 \bar{G}(s_0) \\ &+ \int_0^{s_0} \mathbf{1}_{\{\rho_{u-}<1\}} \bar{F}(s_0-u)\lambda_u du = \rho_{s_0} = 1\end{aligned}$$

Thus we obtain the contradiction that $\rho_{t_0} \leq 1$. Having obtained that ρ takes values in $[0,1]$ it is easy to obtain the following hitting times to 1 and exit times from 1 for a solution ρ. We denote:

$$\text{if } \rho_0 < 1 \text{ then } \sigma_0 = 0, \text{ else } \sigma_0 = \inf\left\{t > 0 : \rho_0 \bar{G}(t) < 1\right\}$$

$$\tau_1 = \inf\left\{t > 0 : \rho_0 \bar{G}(t) + \int_0^t \bar{F}(t-u)\lambda_u du = 1\right\},$$

$$\sigma_1 = \inf\left\{t > \tau_1 : \rho_0 \bar{G}(t) + \int_0^{\tau_1} \bar{F}(t-u)\lambda_u du < 1\right\}.$$

Here τ_1 denotes the first hitting time from below for a solution ρ and σ_1 (σ_0) denotes the first exit time from 1 when the initial condition $\rho_0 < 1$ ($\rho_0 = 1$). The next set of hitting and exit times are defined similarly. For $k \geq 2$ denote:

$$\tau_k = \inf\left\{t > \sigma_{k-1} : \rho_0 \bar{G}(t) + \int_{I_{k,t}} \bar{F}(t-u)\lambda_u du = 1\right\},$$
$$\text{where } I_{k,t} = \cup_{i=1}^{k-1} [\sigma_{i-1}, \tau_i) \cup [\sigma_{k-1}, t).$$

and

$$\sigma_k = \inf\left\{t > \tau_k : \rho_0 \bar{G}(t) + \int_{J_k} \bar{F}(t-u)\lambda_u du < 1\right\}, \tag{14}$$
$$\text{where } J_k = \cup_{i=1}^{k} [\sigma_{i-1}, \tau_i).$$

In addition the specific solution ρ can now be actually represented as

$$\rho_t = \rho_0 \bar{G}(t) + \int_{J_t} \bar{F}(t-u)\lambda_u du, \tag{15}$$

where $J_t = \cup_{i=1}^{\infty} [\sigma_{i-1}, \tau_i) \cap [0,t]$. Let us justify the above representation rigorously. Consider first the case when G is continuous. This yields that ρ must also be continuous. Now observe that the interval $[0,1)$ is open in $[0,1]$ and as such the pre-image of $[0,1)$ with respect to ρ: $\rho^{-1}[0,1) = \{t \in [0,\infty) : \rho_t \in [0,1)\}$, is an open set of $[0,\infty)$. Consequently this pre-image can be represented as countable union of open intervals in $[0,\infty)$:

$$\rho^{-1}[0,1) = \cup_{k=1}^{\infty} L_k, \tag{16}$$

where L_k's are open intervals in $[0,\infty)$. If G however is not continuous, the fact that it is right continuous and non-decreasing guarantees, as we have already mentioned before, that $\bar{G}$ has at most countably many jumps of negative size. The addition of this complexity does not complicate the pre-image too much. We just have at most countably many L_k's in (16) replaced by left-closed right-open intervals, that is: $\rho^{-1}[0,1) = \cup_{k=1}^{\infty} \tilde{L}_k$, where each $\tilde{L}_k$ is either an open or a left-closed right-open interval of $[0,\infty)$. In our considerations above we have denoted J_t to be: $J_t = \left(\cup_{i=1}^{\infty} \tilde{L}_k\right) \cap [0,t]$.

Note that for the purposes of obtaining the solution from J_t using Eq. (15) we may replace the open intervals in $\{\tilde{L}_k\}_{k\geq 1}$ by left-closed right-open intervals without affecting the solution because the solution would be continuous at the left limit point of the said interval. We have thus obtained a one–one correspondence between solutions of (13) and the corresponding intervals $\tilde{L}_k = [\sigma_{k-1}, \tau_k)$ through relation (15). Now suppose (13) admits two solutions ρ^1 and ρ^2. By the one–one correspondence established these two solutions will differ only if they admit two different countable collection of intervals $\{\tilde{L}_k^1\}_{k\geq 1}$ and $\{\tilde{L}_k^2\}_{k\geq 1}$. Let $l = \min\{k : \tilde{L}_k^1 \neq \tilde{L}_k^2\}$ If $l = 1$, deriving a contradiction is straightforward. Indeed, if $\rho_0 < 1$, then it is immediate that the solution would be unique until the first time it hits 1, and consequently $\sigma_0 = 0$ and τ_1 are unique. Similarly, if $\rho_0 = 1$, $\sigma_0 = \inf\{t > 0 : \rho_0 \bar{G}(t) < 1\}$, which is unique and by translation one can obtain a new equation on $[\sigma_0, \infty)$ as follows: $\gamma_s = \rho_0 \bar{G}(s+\sigma_0) + \int_{\sigma_0}^{s} \bar{F}(s-u)\lambda_u du$, which also admits a unique solution until it hits 1. Thus in both cases, the first interval $\tilde{L}_1$ is determined by F, G and λ, and hence unique. The argument for the latter case can be modified and applied to derive a contradiction when $l > 1$. In this case σ_{l-1} is given by (14) with $J_{l-1} = \cup_{i=1}^{l-1} \tilde{L}_i$, and is hence unique. We therefore translate our equation to $[\sigma_{l-1}, \infty)$ to obtain like before: $\gamma_s = \rho_0 \bar{G}(s+\sigma_{l-1}) + H(s) + \int_{\sigma_{l-1}}^{s} \bar{F}(s-u)\lambda_u du$, where $H(s) = \int_{J_{l-1}} \bar{F}(s-u)\lambda_u du$. Again, this admits a unique solution until

the first time it hits 1, and hence $\tilde{L}_1$ is also unique. This provides our required contradiction. □

3.4. Related processes

3.4.1. Integrated fraction of occupied servers

Let Θ^n denote the integrated fraction of occupied servers, that is, for $t > 0$: $\Theta^n_t = \int_0^t \rho^n_u du$. Observe that $t - \Theta^n_t = \int_0^t (1 - \rho^n_u) du$ is the cumulative idleness (in the sense that this counts the time instants when any server is idle) since $\rho^n_u = 1$ only if all the servers are occupied. Indeed, the fluid limit in (18) below provides a relation between the mean service times, arrival intensity and the integrated process.

It is readily seen that Θ^n has an integral representation with respect to the counting measure p^R:

$$\Theta^n_t = \frac{1}{n}\sum_{i=1}^{\infty} \mathbf{1}_{\{\rho^n_{T_i-}<1\}}\mathbf{1}_{\{T_i<t\}} S_i \wedge (t-T_i) = \int_0^t \int_{\mathbb{R}} V_n(t,u,x) p^R(du,dx), \tag{17}$$

where $V_n(t,u,x) = \frac{1}{n}\mathbf{1}_{\{\rho^n_{u-}<1\}}\mathbf{1}_{\{u<t\}}(x \wedge (t-u))$. Similar to Theorem 3.4 we now have the following fluid limit for the integrated process Θ^n.

Theorem 3.11. *Let the conditions in Assumption 3.6 hold. Then for any $T > 0$ we have:*

$$\lim_{n\to\infty} \sup_{t\in[0,T]} \left|\Theta^n_t - \Theta_t\right| = 0, \quad \text{almost surely,}$$

where Θ for $t > 0$ is given by $\Theta_t = \int_0^t \rho_u du$, and where ρ is given by (13).

Remark 3.12. Observe that (13) implies that Θ has the alternate more explicit expression:

$$\Theta_t = \rho_0 \int_0^t \mathbf{E}[S^0 \wedge t] du + \int_0^t \mathbf{1}_{\{\rho_{u-}<1\}} \mathbf{E}[S \wedge (t-u)] \lambda_u du, \tag{18}$$

where S^0 is distributed as G, while S is distributed as F. This is readily obtained by integrating the right hand side of (13).

Proof. By Corollary 3.8 we have that $\lim_{n\to\infty} \sup_{t\in[0,T]} |\rho^n_t(\omega) - \rho_t| = 0$, for all $\omega \in \Omega_1$ such that $\Omega_1 \subset \Omega$ and $P(\Omega) = 1$. Consequently fix $\omega \in \Omega_1$ and $\varepsilon > 0$ to obtain $N(\omega)$ large enough such that $\left|\rho^n_t(\omega) - \rho_t\right| < \frac{\varepsilon}{T}$, for all $n > N(\omega)$. Thus we have $\left|\int_0^t \rho^n_u du - \int_0^t \rho_u du\right| \le \int_0^t \left|\rho^n_u - \rho_u\right| du \le \varepsilon$, for all $n > N(\omega)$. This completes the proof.

Alternate explicit proof. From (17) we have the following decomposition

$$\begin{aligned}
\Theta^n_t &= \int_0^t \int_{\mathbb{R}} V_n(t,u,x) p^R_*(du,dx) + \int_0^t \int_{\mathbb{R}} V_n(t,u,x) p^R_c(du,dx) \\
&= \int_0^t \int_{\mathbb{R}} V_n(t,u,x) p^R_*(du,dx) \\
&\quad + \int_0^t \int_{\mathbb{R}} \mathbf{1}_{\{\rho^n_{u-}<1\}}\mathbf{1}_{\{u<t\}}(x \wedge (t-u)) \nu(x) \lambda_u dx du \\
&= \int_0^t \int_{\mathbb{R}} V_n(t,u,x) p^R_*(du,dx) \\
&\quad + \int_0^t \mathbf{1}_{\{\rho^n_{u-}<1\}} \mathbf{E}[X \wedge (t-u)] \lambda_u du \\
&= X^{1,n}_t + \gamma^{1,n}_t,
\end{aligned}$$

where $X^{1,n}$ is a semimartingale bounded above by Θ^n (since $\gamma^{1,n}$ is positive) and in turn by T (since we are in a finite time horizon T and $|\rho^n| \le 1$).

Also, notice that Θ^n is a continuous function and since $\rho^n \le 1$, its modulus of continuity satisfies: $w_{\Theta^n}(\delta) := \sup_{|s-t|<\delta} |\int_s^t \rho^n_u du| \le \delta$ In addition, the modulus of continuity of $\gamma^{1,n}$ satisfy: $w_{\gamma^{1,n}}(\delta) \le \sup_{|s-t|\le\delta} 2\int_s^t \mathbf{E}[X]\lambda_u du \le 2 w_{\Lambda^*}(\delta)$, where Λ^* is given by $\Lambda^*(t) = \int_0^t \mathbf{E}[X]\lambda_u du$, which is uniformly continuous on $[0,T]$. As a consequence of all this we have: $\lim_{\delta\downarrow 0} \sup_n w_{X^{1,n}}(\delta) = 0$. Observe that the L^2-norm of the semimartingale $X^{1,n}$ satisfies:

$$\begin{aligned}
\mathbf{E}\left(X^{1,n}_t\right)^2 &= \mathbf{E}\left[\int_0^t V_n^2(u,x) p^R_c(du,dx)\right] \\
&\le \frac{1}{n^2}\int_0^t \int_{\mathbb{R}} (x \wedge (t-u))^2 p^R_c(du,dx) \\
&\le \frac{1}{n}\mathbf{E}[X^2]\int_0^t \lambda_u du \le \frac{1}{n}\mathbf{E}[X^2]\Lambda \longrightarrow 0.
\end{aligned}$$

Consequently for each $t \in [0,T]$, $X^{1,n}_t \xrightarrow{p} 0$. Thus the finite dimensional vectors $(X^{1,n}_{t_1}, \ldots, X^{1,n}_{t_n}) \xrightarrow{p} (0,\ldots,0)$. Recalling the tightness condition we have thus obtained that $X^{1,n}$ converges in distribution to the constant zero function and hence also in probability. Since the limiting function is continuous, the convergence is also in the uniform topology. Thus we have: $X^{1,n} \xrightarrow{ucp} 0$. By our previous considerations since $\rho^n \xrightarrow{ucp} \rho$ we have $\gamma^{1,n} \xrightarrow{ucp} \gamma^1$, where γ^1 is given by $\gamma^1_t = \int_0^t \mathbf{1}_{\{\rho_{u-}<1\}} \mathbf{E}[X \wedge (t-u)]\lambda_u du$. We have thus obtained $\Theta^n \xrightarrow{ucp} \gamma^1$. A similar trick as employed in the previous section guarantees almost sure convergence. This is because for any subsequence there is a further subsequence such that the convergence of $X^{1,n}$ and $\gamma^{1,n}$ happen almost surely. In addition, there exists an $\Omega_3 \subset \Omega$ with $P(\Omega_3) = 1$ such that for every $\omega \in \Omega_3$, the sequence $\{\Theta^n(\omega)\}_{n\ge1}$ is tight. Similar calculations as employed in the previous section now yield: $\Theta^n \longrightarrow \gamma^1$ almost surely, with uniform convergence over $[0,T]$. □

The following regarding the integrated process is an easy corollary of Theorem 3.10.

Corollary 3.13. *There exists a unique solution Θ to (18), where ρ is given by (13).*

3.4.2. Blocked arrivals

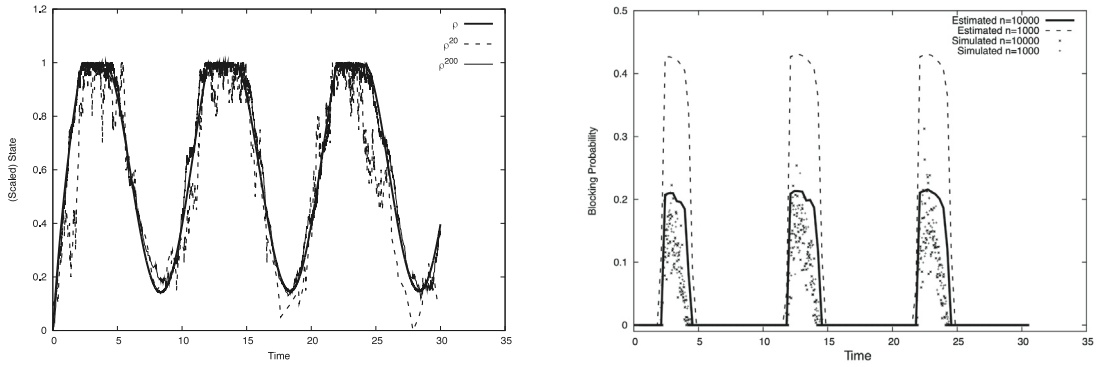
Blocking probabilities and congestion measurement are the most important measures of performance in loss models. Computing these quantities in non-stationary models is rather hard to do, however, requiring approximations [14,21]. In this section we briefly note how the fluid approximations established in this paper can be utilized.

The following calculation provides a way of approximating blocking probabilities using the functional strong law obtained in Theorem 3.4. Following (5), we define the stochastic process,

$$\hat{\rho}^n_t := \frac{1}{n}\sum_{i=1}^{\infty} \mathbf{1}_{\{\rho_{T_i-}<1\}}\mathbf{1}_{\{S_i+T_i>t\}}\mathbf{1}_{\{T_i<t\}} \;\forall\, t \in [0,T]. \tag{19}$$

Observe that $\hat{\rho}^n$ is formally equivalent to ρ^n in (5), but uses the FSLLN limit $\rho_\cdot$ from (7) in lieu of ρ^n on the right hand side of (5). Now, using the thinning property of Poisson processes it follows that for each $t \in [0,T]$, $n\hat{\rho}^n_t$ must be distributed as a Poisson random variable with mean

$$\int_0^t \int_{t-u}^{\infty} \mathbf{1}_{\{\rho_u<1\}} \nu(x) n \lambda_u dx du = n \int_0^t \mathbf{1}_{\{\rho_u<1\}} \bar{F}(t-u) \lambda_u du = n\rho_t.$$



(a) Convergence of ρ^n to ρ for an initially empty system.

(b) Comparison of the estimated and simulated blocking probabilities.

Fig. 1. ρ^n has been simulated for $n = 20$ and $n = 200$ with service times drawn from Lognormal$(-0.5, 1)$ while the intensity of arrivals is sinusoidal with $\lambda_u = \frac{2}{3}(1 + \sin(\frac{2\pi u}{10}))$. ρ has been approximated using a mollified version of the indicator function in relation (8) (note ρ does not reach 1 as a consequence).

Therefore, as a consequence of Theorem 3.4, we claim that $n\rho_t^n \overset{d}{\approx} \text{Poi}(n\rho_t)$, where Poi(1) is a Poisson random variable with mean 1. Of course the approximation made above is formal and only a careful second order functional central limit analysis of ρ^n would allow a rigorous justification of the approximation. This is outside the scope of the current paper. However, for practical purposes the blocking probabilities can be roughly estimated as $\mathbf{P}\left(\rho_t^n = 1\right) \approx \mathbf{P}\left(\frac{\text{Poi}(n\rho_t)}{n} = 1\right)$. Fig. 1(b) compares the simulated blocking probabilities for the setting described in the caption. Observe that for large enough n, the estimated blocking probability is remarkably close to the simulated blocking probability (averaged over 100 simulated sample paths).

4. Conclusions

The results in this note complement extant results establishing many-server fluid limits, by considering the nonstationary, non-Markovian loss model setting, and by using a bespoke semimartingale representation of the fraction of occupied servers. The primary result shows that the fraction of occupied servers converges to the unique solution of a Volterra integral equation. As a consequence of our main result, we also establish a fluid limit to the integrated number of occupied servers and the fraction of arriving jobs that are blocked on arrival. We anticipate our results, proofs of which are quite simple, should be broadly useful. In future work we anticipate extending the analysis to include functional central limit theorems. However, the analysis is much harder than the fluid limit results in this note and merit a separate paper.

Acknowledgments

PC gratefully acknowledges support from a Purdue Research Foundation dissertation fellowship, USA. HH was partly supported by the National Science Foundation, USA through grant CMMI/1636069. The authors thank the anonymous referee for their insightful comments, in particular on approximating the instantaneous blocking probability.

References

[1] D. Applebaum, Lévy Processes and Stochastic Calculus, Cambridge University Press, 2009.
[2] G. Bet, R. van der Hofstad, J.S. van Leeuwaarden, Heavy-traffic analysis through uniform acceleration of queues with diminishing populations, Math. Oper. Res. (2019).
[3] P. Billingsley, Convergence of Probability Measures, John Wiley & Sons, 2013.
[4] P. Brémaud, Point Processes and Queues: Martingale Dynamics, 50, Springer, 1981.
[5] P. Chakraborty, H. Honnappa, Strong embeddings for transitory queueing models, 2019, arXiv preprint arXiv:1906.06740.
[6] S. Halfin, W. Whitt, Heavy-traffic limits for queues with many exponential servers, Oper. Res. 29 (3) (1981) 567–588.
[7] H. Honnappa, R. Jain, A.R. Ward, A queueing model with independent arrivals, and its fluid and diffusion limits, Queueing Syst. 80 (1–2) (2015) 71–103.
[8] H. Kaspi, K. Ramanan, et al., Law of large numbers limits for many-server queues, Ann. Appl. Probab. 21 (1) (2011) 33–114.
[9] T. Kiffe, A discontinuous Volterra integral equation, J. Integral Equ. (1979) 193–200.
[10] Y. Liu, W. Whitt, The $G_t/GI/s_t+GI$ many-server fluid queue, Queueing Syst. 71 (4) (2012) 405–444.
[11] Y. Liu, W. Whitt, A many-server fluid limit for the Gt/GI/st+ GI queueing model experiencing periods of overloading, Oper. Res. Lett. 40 (5) (2012) 307–312.
[12] Y. Liu, W. Whitt, et al., Many-server heavy-traffic limit for queues with time-varying parameters, Ann. Appl. Probab. 24 (1) (2014) 378–421.
[13] W.A. Massey, W. Whitt, An analysis of the modified offered-load approximation for the nonstationary Erlang loss model, Ann. Appl. Probab. (1994) 1145–1160.
[14] J. Pender, Nonstationary loss queues via cumulant moment approximations, Probab. Engrg. Inform. Sci. 29 (1) (2015) 27–49.
[15] J. Pender, Y.M. Ko, Approximations for the queue length distributions of time-varying many-server queues, INFORMS J. Comput. 29 (4) (2017) 688–704.
[16] P. Protter, Stochastic Integration and Differential Equations, Springer, Berlin, Heidelberg, 2005.
[17] J. Reed, et al., The /GI/N queue in the Halfin–Whitt regime, Ann. Appl. Probab. 19 (6) (2009) 2211–2269.
[18] J.A. Wellner, A Glivenko-Cantelli theorem for empirical measures of independent but non-identically distributed random variables, Stochastic Process. Appl. 11 (3) (1981) 309–312.
[19] J.A. Wellner, et al., A Glivenko-Cantelli theorem and strong laws of large numbers for functions of order statistics, Ann. Statist. 5 (3) (1977) 473–480.
[20] W. Whitt, Time-varying queues, Queueing Models Serv. Manage. 1 (2) (2018).
[21] W. Whitt, J. Zhao, Many-server loss models with non-poisson time-varying arrivals, Nav. Res. Logist. 64 (3) (2017) 177–202.
[22] J. Zhang, Fluid models of many-server queues with abandonment, Queueing Syst. 73 (2) (2013) 147–193.