Stochastic Volterra Integral Equations with Ranks as Scaling Limits of Parallel Infinite-Server Queues under Weighted Shortest Queue Policy

TOMOYUKI ICHIBA* AND GUODONG PANG**

Abstract. We study a queueing system with a fixed number of parallel service stations of infinite servers, each having a dedicated arrival process, and one flexible arrival stream that is routed to one of the service stations according to a "weighted" shortest queue policy. We consider the model with general arrival processes and general service time distributions. Assuming that the dedicated arrival rates are of order n and the flexible arrival rate is of order $\sqrt{n}$, we show that the diffusion-scaled queueing processes converge to a stochastic Volterra integral equation with "ranks" driven by a continuous Gaussian process. It reduces to the limiting diffusion with a discontinuous drift in the Markovian setting.

1. Introduction

We consider a system of parallel service stations, each of which has a dedicated arrival process and an infinite number of servers. There is also a flexible arrival stream which can be served by any of service stations, according to a "weighted" shortest queue routing policy. This model was previously studied in [9, 5, 15] when the arrival processes are Poisson and the service times are i.i.d. exponential. In this paper we study the model with general arrival processes and service times with general distributions. The model has many applications, such as CDMA cellular systems [20] and customer service systems. The model is also related to the studies of "load balancing" in the sense that the "weighted" shortest queue policy for the flexible arrivals balances the load of each of the service stations. We refer the readers to the recent survey on load balancing in [6], and [22, 23] for studies on joining the shortest queue policy in infinite-server queues.

We study the system behavior under heavy traffic, that is, when the arrival rates are scaled to grow to infinity, while the service times are unscaled. In the Markovian setting (Poisson arrivals and exponential service times), when the dedicated arrivals are of order n and the flexible arrival stream is of order $\sqrt{n}$, Fleming and Simon [9] conjectured the diffusion limit with a discontinuous drift, which was then formally proved by Chao [5]. Krylov and Lipster [15] considered a similar model with a modification in the service process when the queues are over a threshold, and proved a diffusion limit with both discontinuous drift and variance coefficients. The discontinuity in the drift is a consequence of the "weighted"

*Department of Statistics & Applied Probability, University of California, Santa Barbara, CA 93106

**Department of Computational Applied Mathematics and Operations Research, George R. Brown School of Engineering, Rice University, Houston, TX 77005

E-mail addresses: ichiba@pstat.ucsb.edu, gdpang@rice.edu.

Date: December 10, 2023.

Key words and phrases. Stochastic Volterra integral equation with ranks, "weighted" shortest queue routing policy, parallel infinite-server queues.

shortest queue policy, and in fact, the diffusion limit can be also regarded as a diffusion with "ranks" in the drift. The discontinuity in the drift prevents us from applying the standard techniques in establishing heavy-traffic scaling limits for queueing processes (such as the continuous mapping theorem applied to the integral mapping for standard Markovian many-server queues [17]). The idea to circumvent the discontinuity in the drift and/or variance coefficient in [5, 15] is to show that the time spent by the process in the set of their discontinuity is almost surely Lebesgue measure zero. The methods in [5, 15] rely heavily on semi-martingales (constructed from the Poisson arrival and exponential service processes). However, their approaches through the (super-)martingale characterization do not apply to the non-Markovian setting we are considering.

For $G/GI/\infty$ queues with general arrival process and service times, a functional central limit theorem (FCLT) is established in [13]. In that paper, the FCLT for the diffusion-scaled queueing process gives a Gaussian process limit, which has two independent components, capturing the randomness in the arrival and service processes, respectively. We adapt that approach to our setting, and in fact, will use directly the results on the convergence of these components. We will show that the limit in the FCLT for our model is a multidimensional stochastic Volterra integral equation with "ranks" in the integral term and driven by Gaussian processes (essentially the same Gaussian components as those in the limit for standard $G/GI/\infty$ queues with an additional Gaussian component resulting from the initial quantities), see equation (2.9). However, similarly to [5, 15], we have to tackle the issue of the "ranks" in the integral term of the limit as a consequence of the "weighted" shortest queue policy. We refer to [3] for the existence, uniqueness and perturbation results for the multidimensional stochastic Volterra integral equation driven by Brownian motions with Lipschitz continuous coefficients. Here, (2.9) does not satisfy the sufficient conditions for existence and uniqueness discussed in [3], because of discontinuity of the drift coefficients due to the "ranks".

We show that the limiting stochastic Volterra integral equation has a unique weak solution with continuous paths. In order to prove the existence of a weak solution, we employ the Girsanov change-of-measure theorem for Brownian motion and Brownian sheet (noting that the driving Gaussian processes are functionals of either Brownian motion or Brownian sheet). We include the terms with "ranks" in a construction of a semi-martingale that is a Brownian motion with a random drift, which will become a Brownian motion under a new measure. We also prove an important property of the limit process regarding the "ranks", as in [5, 15]: the cumulative time that the limit process lies at the boundary (that is, when any two "weighted" queues are equal) has Lebesgue measure zero with probability one. For that purpose, we again exploit the Girsanov theorem and representations under the new measure.

In order to prove the convergence of the diffusion-scaled queueing processes to the limit process, we exploit some existing convergence results for the driving Gaussian components for standard $G/GI/\infty$ queues [13]. However, from the representation (4.1), the integral component with "ranks" under the "weighted" shortest queue policy forbids us from applying the continuous mapping theorem directly. To tackle this issue, we exploit the property of the limit process at the boundary mentioned above.

Our limit process (2.9) is related to the works on diffusions with discontinuous drifts in various applications. For example, the weak uniqueness of the diffusions with piece-wise constant coefficients are established by Bass and Pardoux [2], weak uniqueness for diffusions with discontinuous coefficients by Krylov [14], and the strong uniqueness of the diffusions with rank-based coefficients are discussed in [10]. See e.g., [1, 7, 11] for the related stationary distributions and applications to the performance of functionally generated portfolios in financial markets. More generally, it is relevant to the theory of (stochastic) differential equations with discontinuous right hand side, see e.g., [8] for differential equations.

The remaining paper is organized as follows. In Section 2, we describe the model in detail and state the main results Theorems 2.1-2.2. In Section 3, we prove the existence and uniqueness (Theorem 3.1) of a weak solution to the limiting stochastic Volterra integral equation. We prove the convergence the diffusion-scaled processes in Section 4. We show how the limiting stochastic Volterra integral equation is reduced to the limiting diffusion with discontinuous drift in the Markovian case in the Appendix.

2. Model and Results

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ be a filtered probability space where all the random variables and processes are defined. We consider a queueing system with K service stations, each of which has its own dedicated arrival process and an infinite number of parallel servers. In addition, there is a flexible arrival stream, which can be served by any of the stations. Let $A_k = \{A_k(t) : t \geq 0\}$ be the dedicated arrival process at station $k = 1, \dots, K$, with arrival rate λ_k and arrival times $\tau_{k,i}$, $i \in \mathbb{N}$, and $A_0 = \{A_0(t) : t \geq 0\}$ be the flexible arrival process, with arrival rate λ_0 and arrival times $\tau_{0,i}$, $i \in \mathbb{N}$. Assume that these arrival processes are mutually independent. Let $X_k = \{X_k(t) : t \geq 0\}$ be the process counting the number of jobs in station k for $k = 1, \dots, K$, and denote the counting process of jobs in the K service stations by $X = (X_1, \dots, X_K)$.

For jobs initially in service in station k, let $\eta^0_{k,j}$, $j = 1, \dots, X_k(0)$, be their remaining service times, and for the newly arrivals from the stream A_k, let $\eta_{k,i}$, $i \in \mathbb{N}$, be their service times. For the jobs from the arrival stream A_0, let $\eta_{0,i}$, $i \in \mathbb{N}$, be their service times. We assume that the remaining service times $\{\eta^0_{k,j}\}_{k,j}$ are i.i.d. continuous random variables with cumulative distribution function (c.d.f.) F_0, and the service times $\{\eta_{k,i}\}_{k,i}$ are all i.i.d. with c.d.f. F. Let us denote the upper tail probabilities by $F^c_0 = 1 - F_0$ and $F^c = 1 - F$ for F_0 and F, respectively. Without loss of generality, assume that the mean of F is one.

For each station k, we associate a "weight" $\alpha_k > 0$, in order to evaluate the status of station k by a score $\alpha_k X_k(\cdot)$ and our routing policy depends on the scores. If the score $\alpha_k X_k$ of station k is lower than those of the other stations, then we route the newly arriving jobs to station k. More specifically, for the i^{th} job from the arrival stream A_0, at the arrival time $\tau_{0,i}$, it is routed to station k if the score of station k is the lowest among the other scores, that is,

$$\alpha_k X_k(\tau_{0,i}) < \min_{\ell \neq k} \alpha_\ell X_\ell(\tau_{0,i}) \,. \tag{2.1}$$

And if there are multiple stations that have the lowest scores, then job i is routed to the station with the smallest number. For example, if stations 1 and 2 have the same score, i.e., $\alpha_1 X_1(\tau_{0,i}) = \alpha_2 X_2(\tau_{0,i})$ at the time $\tau_{0,i}$ of the arrival of job i, then job i is routed to station 1. The ties of the scores are resolved this lexicographic way. For each $x = (x_1, \dots, x_K) \in \mathbb{R}^K_+$, define an indicator function

$$\delta_k(x) = \begin{cases} 1 & \text{if } k = \min\{j : \alpha_j x_j = \min_{1\leq \ell\leq K} \alpha_\ell x_\ell\}, \\ 0 & \text{otherwise}, \end{cases} \tag{2.2}$$

and the set $\mathcal{R}_k := \{x \in \mathbb{R}^K_+ : \delta_k(x) = 1\}$ for $k = 1, \dots, K$. Then we have $\delta_k(x) = \mathbf{1}_{\mathcal{R}_k}(x)$ for $x \in \mathbb{R}^K_+$, $k = 1, \dots, K$. Moreover, $\mathcal{R}_j \cap \mathcal{R}_k = \emptyset$ for $j \neq k$, and $\cup_{k=1}^K \mathcal{R}_k = \mathbb{R}^K_+$.

Since $\delta_k(c\,x) = \delta_k(x)$ for every $x \in \mathbb{R}_+^K,\ c > 0$, we see that each set $\mathcal{R}_k$ forms a cone for $k = 1, \dots, K$.

Then with these indicator functions, the process $X_k(t)$ can be described as

$$X_k(t) = \sum_{j=1}^{X_k(0)} \mathbf{1}_{\{\eta^0_{k,j} > t\}} + \sum_{i=1}^{A_k(t)} \mathbf{1}_{\{\tau_{k,i}+\eta_{k,i} > t\}} + \sum_{i=1}^{A_0(t)} \delta_k(X(\tau_{0,i}-))\mathbf{1}_{\{\tau_{0,i}+\eta_{0,i} > t\}}, \quad t \geq 0. \tag{2.3}$$

We consider a sequence of such queueing systems, indexed by n, and the scaling limit. We first make the following assumption on the arrival rates.

Assumption 1. *Assume that for $k = 1, \dots, K$, $\lambda_k^n/n \to \lambda_k$ as $n \to \infty$, and $\lambda_0^n/\sqrt{n} \to \lambda_0$ as $n \to \infty$.*

Let $\bar{X}_k^n := n^{-1}X_k^n$ for $k = 1, \dots, K$. The following FLLN holds. We use $D(\mathbb{R}_+, \mathbb{R}^d)$ to denote $\mathbb{R}^d$-valued càdlàg functions, endowed with the Skorohod J_1 topology written as $(D(\mathbb{R}_+, \mathbb{R}^d), J_1)$. When $d = 1$, we write $D = D(\mathbb{R}_+, \mathbb{R})$ and use (D^d, J_1) to denote the d-fold product topology. When the space is restricted on the fixed time interval $[0, T]$ for some $T > 0$, we write $D_{[0,T]} := D([0, T], \mathbb{R})$ and use $(D^d_{[0,T]}, J_1)$ to denote the d-fold product topology.

Theorem 2.1. *Under Assumption 1, if there are deterministic constants $\bar{X}_k(0), k = 1, \dots, K$ such that $(\bar{X}_1^n(0), \dots, \bar{X}_K^n(0)) \Rightarrow (\bar{X}_1(0), \dots, \bar{X}_K(0))$ in $\mathbb{R}_+^K$ as $n \to \infty$,*

$$(\bar{X}_1^n, \dots, \bar{X}_K^n) \Rightarrow (\bar{X}_1, \dots, \bar{X}_K) \quad in \quad (D^K, J_1) \quad as \quad n \to \infty, \tag{2.4}$$

where

$$\bar{X}_k(t) = \bar{X}_k(0)F_0^c(t) + \lambda_k \int_0^t F^c(t-s)\mathrm{d}s, \quad t \geq 0. \tag{2.5}$$

The limits have a steady state $\bar{X}^* = (\bar{X}_1^*, \dots, \bar{X}_K^*)$ with $\bar{X}_k^* = \lambda_k$ for each $k = 1, \dots, K$.

Observe that in the fluid limit, the "weighted" shortest queue policy in (2.1) is irrelevant as indicated in (2.5), and the associated steady state values $\bar{X}_k^* = \lambda_k$. This is evident since the extra load $A_0^n(t)$ is of order $\sqrt{n}$ while the fluid scale is of order n. It is then necessary to consider the queueing dynamics in the diffusion scale.

Let $\hat{X}_k^n := \sqrt{n}(\bar{X}_k^n - \bar{X}_k^*)$ for $k = 1, \dots, K$. Denote $\hat{X}^n = (\hat{X}_1^n, \dots, \hat{X}_K^n)$. Note that we center the process $\hat{X}_k^n$ by its equilibrium point. It is then clear that if $\alpha_k = 1/\lambda_k$,

$$\delta_k(X^n(t)) = \delta_k(\hat{X}^n(t)), \quad t \geq 0.$$

We choose this specific value of $\alpha_k = 1/\lambda_k$ to establish the following FCLT. Note that since all the service times are i.i.d. with mean one, the value λ_k is the (offered) load at each station. Thus, the routing criterion chooses the station with minimum ratio of the current state and the steady state. When λ_k's are equal for all k, the routing policy becomes the so-called "joining the shortest queue" (JSQ) policy. Thus, the routing policy can be regarded as a "weighted" JSQ policy with the weights being the reciprocal of the offered load.

Let $\hat{A}_k^n(t) := \frac{1}{\sqrt{n}}(A_k^n(t) - \lambda_k^n t)$ for $t \geq 0$ and $k = 1, \dots, K$, and let $\hat{A}_0^n(t) = \frac{1}{\sqrt{n}}A_0^n(t)$ for $t \geq 0$. We make the following assumption for these scaled arrival processes.

Assumption 2. *The following hold for the arrival processes:*

(i) *In addition to the conditions in Assumption 1, for each $k = 1, \dots, K$, there exists $\hat{\lambda}_k \in \mathbb{R}$ such that $\hat{\lambda}_k^n := \sqrt{n}(\lambda_k^n/n - \lambda_k) \to \hat{\lambda}_k$ as $n \to \infty$.*

(ii) *There exist mutually independent Brownian motion* $(\hat{A}_1, \dots, \hat{A}_K)$ *such that*

$$(\hat{A}^n_1, \dots, \hat{A}^n_K) \Rightarrow (\hat{A}_1, \dots, \hat{A}_K) \quad in \quad (D^K, J_1) \quad as \quad n \to \infty$$

where $\hat{A}_k \stackrel{\mathrm{d}}{=} c_k \hat{B}_k(t)$ *for the variance coefficient* $c_k > 0$ *and a standard Brownian motion* $\hat{B}_k$ *(mutually independent over* k*).*

(iii) $\hat{A}^n_0 \Rightarrow \lambda_0 e$ *in* (D, J_1) *as* $n \to \infty$*, where* $e(t) \equiv t$ *for* $t \ge 0$*. Moreover, associating measures on* $[0, T]$ *to the nonnegative, non-decreasing, càdlàg functions* $\hat{A}^n_0$ *and* $\lambda_0 e$*, we assume the total variation distance between* $\mathrm{d}\hat{A}^n_0$ *and* $\lambda_0 \mathrm{d}e$ *converges weakly to* 0*, as* $n \to \infty$ *for every* $T > 0$*.*

In the second condition, when the arrival processes A_k 's are mutually independent renewal processes with the interarrival times having mean λ_k^{-1} and variance σ_k^2, if A^n_k is defined by scaling the interarrival times by n^{-1}, then $\hat{A}_k(t) = \sqrt{\lambda_k^3 \sigma_k^2} \hat{B}_k(t)$ for standard Brownian motion $\hat{B}_k(t)$ (see, e.g., [21, Chapter 13.7]).

We also make the following assumption on the initial condition.

Assumption 3. *There exists a random vector* $(\hat{X}_1(0), \dots, \hat{X}_K(0)) \in \mathbb{R}^K$ *such that*

$$(\hat{X}^n_1(0), \dots, \hat{X}^n_K(0)) \Rightarrow (\hat{X}_1(0), \dots, \hat{X}_K(0)) \quad in \quad \mathbb{R}^K_+ \quad as \quad n \to \infty.$$

Under this condition, given the scaling of $\hat{X}_k$, it is also clear that $\bar{X}_k(0) = \bar{X}^*_k = \lambda_k$ for each $k = 1, \dots, K$.

For the FCLT below, we also assume that $F_0(t) = F_e(t) = \int_0^t F^c(s)\mathrm{d}s$, the equilibrium (stationary excess) distribution of F. Recall that we have assumed that the mean for the c.d.f. F is one. In the scaling limits, we have the following three mutually independent driving noises:
(i) some K-dimensional, independent Gaussian processes $\hat{X}_{\cdot,0} := (\hat{X}_{1,0}, \dots, \hat{X}_{K,0})'$,
(ii) another K-dimensional, independent Gaussian process, $\hat{X}_{\cdot,1} := (\hat{X}_{1,1}, \dots, \hat{X}_{K,1})'$ and
(iii) the K-dimensional, independent Brownian motions $(\hat{A}_1, \dots, \hat{A}_K)$ with strictly positive variance rates $(c_1, \dots, c_K)$ from Assumption 2.

The processes $\hat{X}_{k,0}$ and $\hat{X}_{k,1}$ are independent continuous Gaussian processes, independent of $\hat{A}_k$, with mean zero and covariance functions: for $t, t' \ge 0$,

$$\mathrm{Cov}(\hat{X}_{k,0}(t), \hat{X}_{k,0}(t')) = \lambda_0 \big(F^c_e(t \vee t') - F^c_e(t) F^c_e(t')\big), \tag{2.6}$$

$$\mathrm{Cov}(\hat{X}_{k,1}(t), \hat{X}_{k,1}(t')) = \lambda_k \int_0^{t \wedge t'} \big(F^c(t \vee t' - s) - F^c(t - s) F^c(t' - s)\big) \mathrm{d}s. \tag{2.7}$$

In addition, the processes $\hat{X}_{k,0}$ and $\hat{X}_{k,1}$ are independent of $\hat{X}_{k',0}$ and $\hat{X}_{k',1}$ as well as $\hat{A}_{k'}$ for every $k' \neq k$. The process $\int_0^t F^c(t-s)\mathrm{d}\hat{A}_k(s)$ is also a continuous Gaussian process with covariance function: for $k = 1, \dots, K$,

$$\mathrm{Cov}\Big(\int_0^t F^c(t-s)\mathrm{d}\hat{A}_k(s), \int_0^{t'} F^c(t'-s)\mathrm{d}\hat{A}_k(s)\Big) = c_k \int_0^{t \wedge t'} F^c(t-s) F^c(t'-s)\mathrm{d}s.$$

As we shall see later in (3.2) and in (3.5), respectively, $\hat{X}_{k,0}$ and $\hat{X}_{k,1}$ can be represented as a time-changed, Brownian bridge driven by another Brownian motion, and as a time-changed, Kiefer process driven by a Brownian sheet, independent from the Brownian motions.

Theorem 2.2 (FCLT)**.** *Under Assumptions 1, 2 and 3, we have*

$$(\hat{X}_1^n, \dots, \hat{X}_K^n) \Rightarrow (\hat{X}_1, \dots, \hat{X}_K) \quad in \quad (D^K, J_1) \quad as \quad n \to \infty, \tag{2.8}$$

where $\hat{X} = (\hat{X}_1, \dots, \hat{X}_K)$ is the unique weak solution to the following system of stochastic Volterra integral equations with ranks, driven by independent, continuous Gaussian processes $(\hat{X}_{1,0}, \dots, \hat{X}_{K,0})$, $(\hat{X}_{1,1}, \dots, \hat{X}_{K,1})$ and Brownian motions $(\hat{A}_1, \dots, \hat{A}_K)$: for $k = 1, \dots, K$,

$$\begin{aligned} \hat{X}_k(t) = \hat{X}_k(0) F_e^c(t) + \hat{\lambda}_k F_e(t) + \lambda_0 \int_0^t \delta_k(\hat{X}(s)) F^c(t-s) \mathrm{d}s \\ + \int_0^t F^c(t-s) \mathrm{d}\hat{A}_k(s) + \hat{X}_{k,0}(t) + \hat{X}_{k,1}(t), \end{aligned} \tag{2.9}$$

where F_e is the equilibrium distribution function that satisfies $F_0(\cdot) = F_e(\cdot) = \int_0^{\cdot} F^c(s)\mathrm{d}s$.

Remark 2.1. *If we denote the last three Gaussian processes in (2.9) by $\hat{Y}_k(t) = \int_0^t F^c(t-s)\mathrm{d}\hat{A}_k(s) + \hat{X}_{k,0}(t) + \hat{X}_{k,1}(t)$ and assume F has a density f and $F(0) = 0$, then we can write*

$$\begin{aligned} \hat{X}_k(t) &= \hat{X}_k(0) F_e^c(t) + \hat{\lambda}_k F_e(t) + \lambda_0 \int_0^t \delta_k(\hat{X}(s)) F^c(t-s) \mathrm{d}s + \hat{Y}_k(t) \\ &= \int_0^t \left(-\hat{X}_k(0) F^c(s) + \hat{\lambda}_k F^c(s) + \lambda_0 \delta_k(\hat{X}(s)) - \lambda_0 \int_0^s \delta_k(\hat{X}(s)) f(s-u) \mathrm{d}u \right) \mathrm{d}s + \hat{Y}_k(t) \end{aligned} \tag{2.10}$$

for every $t \geq 0$. Observe that in the "drift", there is not only a discontinuous term $\lambda_0 \delta_k(\hat{X}(t))$, but also a memory of the process $\hat{X}(t)$ in the term $\lambda_0 \int_0^t \delta_k(\hat{X}(s)) f(t-s)\mathrm{d}s$.

Remark 2.2 (Diffusion with discontinuous drift)**.** *When the c.d.f. F is given by $F(t) = 1 - e^{-t}$, $t \geq 0$, it can be shown that the stochastic equation in (2.10) reduces to the diffusion with discontinuous drift studied in [5] and [15] (with some modification on the coefficients and also in the drift due to the absence of queue thresholds, see also the conjecture in [9]), that is,*

$$\mathrm{d}\hat{X}_k(t) = (\hat{\lambda}_k + \lambda_0 \delta_k(\hat{X}(t)) - \hat{X}_k(t))\, \mathrm{d}t + \sqrt{\lambda_k + c_k^2}\, \mathrm{d}\tilde{B}_k(t)\,, \tag{2.11}$$

for standard Brownian motion $(\tilde{B}_1, \dots, \tilde{B}_K)$. The proof of this property is given in the Appendix.

Remark 2.3 (Atlas models)**.** *We remark that in the completely symmetric case, that is, λ_k's are all equal, we have $\hat{\lambda}_k$'s are also all equal, and assuming that $\hat{X}_k(0)$ have the same distribution for all k, then the limit process $\hat{X}_k(t)$ in (2.9) becomes symmetric over k with the ranks $\delta_k(\hat{X}(t))$ only ranking the coordinates without weights at each time t. When $\alpha_1 = \cdots = \alpha_K$, this resembles the first-order Atlas model in stochastic portfolio theory [7], which is a diffusion with both the drift and variance coefficients depending only on ranking, driven by Brownian motion.*

3. Existence and uniqueness of a weak solution to the limiting stochastic Volterra integral equation with ranks

In this section, we prove that the stochastic Volterra integral equation in (2.9) has a unique weak solution that has continuous paths.

Theorem 3.1. *The stochastic Volterra integral equation* (2.9) *has a unique weak solution in* $\mathbb{C}(\mathbb{R}_+, \mathbb{R}^K)$*. Namely, on some filtered probability space* $(\Omega, \mathbb{P}, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0})$*, a continuous, adapted,* K*-dimensional process* $\hat{X} = (\hat{X}_1, \dots, \hat{X}_K)'$*, an independent,* K*-dimensional Gaussian process* $\hat{X}_{\cdot,0} := (\hat{X}_{1,0}, \dots, \hat{X}_{K,0})'$*, another independent,* K*-dimensional Gaussian process* $\hat{X}_{\cdot,1} := (\hat{X}_{1,1}, \dots, \hat{X}_{K,1})'$ *determined by the covariance functions* (2.6)*, and an independent,* K*-dimensional Brownian motion* $\hat{A}.$ *with variance rates* $(c_1, \dots, c_K)$ *specified in Assumption 2 satisfy* (2.9) *almost surely, and the solution is unique in the sense of probability law. Moreover, almost surely,*

$$\int_0^\infty \sum_{k,\ell=1,\, k\neq\ell}^{K} \mathbf{1}_{\{\alpha_k \hat{X}_k(t) = \alpha_\ell \hat{X}_\ell(t)\}} \mathrm{d}t \;=\; 0\,, \tag{3.1}$$

where $\{\alpha_k\}_{1\leq k\leq K}$ *are the associated weights in* (2.1)*.*

Before proceeding to the proof, we make the following observations on the Gaussian processes $\hat{X}_{\cdot,0} := (\hat{X}_{1,0}, \dots, \hat{X}_{K,0})'$ and $\hat{X}_{\cdot,1} := (\hat{X}_{1,1}, \dots, \hat{X}_{K,1})'$. The process $\hat{X}_{k,0}$ is equivalent in distribution to a time-changed Brownian bridge $\hat{W}_k^0$:

$$\hat{X}_{k,0}(t) = \lambda_k^{1/2} W^0(F_e(t)) = \lambda_k^{1/2} \hat{W}_k^0(1 - e^{-t})\,. \tag{3.2}$$

Recall that the Brownian bridge W_k^0 is the unique strong solution to the one-dimensional SDE [12]:

$$\mathrm{d}\hat{W}_k^0(t) = -\frac{\hat{W}_k^0(t)}{1-t}\mathrm{d}t + \mathrm{d}\breve{B}_k(t)\,,$$

where $\breve{B}_k(t)$ is an independent standard Brownian motion. Thus we can write

$$\begin{aligned}
\hat{X}_{k,0}(t) &= \lambda_k^{1/2}\bigg(-\int_0^{1-e^{-t}} \frac{W_k^0(s)}{1-s}\mathrm{d}s + \breve{B}_k(1-e^{-t}) \bigg) \\
&= -\int_0^t \hat{X}_{k,0}(s)\mathrm{d}s + \lambda_k^{1/2}\breve{B}_k(1-e^{-t})\,,
\end{aligned} \tag{3.3}$$

for standard Brownian motions $\breve{B}_k(t)$, independent of $\hat{B}_k(t)$'s in Assumption 2 (ii). The second equality follows from a calculation using change of variables for the integrable term.

Next, the Gaussian process $\hat{X}_{k,1}$ is equivalent in distribution to a time-changed Kiefer process, $\hat{\mathcal{K}}_k(t,x)$, that is,

$$\hat{X}_{k,1}(t) = -\int_0^t\int_0^t \mathbf{1}_{s+x\leq t}\mathrm{d}\hat{\mathcal{K}}_k(\lambda_k s, F(x)) = -\int_0^t\int_0^t \mathbf{1}_{s+x\leq t}\mathrm{d}\hat{\mathcal{K}}_k(\lambda_k s, 1-e^{-x}). \tag{3.4}$$

Recall that, similar to Brownian bridge, the Kiefer process $\hat{\mathcal{K}}_k(t,x)$ can be written in terms of Brownian sheets $\hat{W}_k(t,x)$, that is, $\hat{\mathcal{K}}_k(t,x)$ is the unique strong solution to the SDE [19]:

$$\mathcal{K}(t,x) = -\int_0^x \frac{\mathcal{K}(t,y)}{1-y}\mathrm{d}y + \hat{W}_k(t,x).$$

So we have

$$\begin{aligned}
\hat{X}_{k,1}(t) &= -\int_0^t\int_0^t \mathbf{1}_{s+x\leq t}\mathrm{d}_{s,x}\bigg(-\int_0^{1-e^{-x}} \frac{\mathcal{K}(\lambda_k s, y)}{1-y}dy + \hat{W}_k(\lambda_k s, 1-e^{-x}) \bigg) \\
&= -\int_0^t \hat{X}_{k,1}(s)\mathrm{d}s + \int_0^t\int_0^t \mathbf{1}_{s+x\leq t}\mathrm{d}\hat{W}_k(\lambda_k s, 1-e^{-x})\,,
\end{aligned} \tag{3.5}$$

where $\hat{W}_k$'s are mutually independent Brownian sheets, which are also independent of the standard Brownian motions $\hat{B}_k$'s in (3.3) and B_k's in Assumption 2 (ii).

Proof of Theorem 3.1. Let us consider a probability space $(\Omega, \mathbb{P}, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0})$ under which the independent Brownian motions $(\hat{A}_1, \dots, \hat{A}_K)$ in Assumption 2, independent Brownian bridges $(\hat{W}^0_1, \dots, \hat{W}^0_K)$ in (3.3) and independent Kiefer processes $(\hat{\mathcal{K}}_1, \dots, \hat{\mathcal{K}}_K)$ in (3.4) are defined. For $k = 1, \dots, K$, let

$$\hat{R}_k(t) := \hat{\lambda}_k t + c_k \hat{B}_k(t) + \lambda_0 \int_0^t \delta_k(\hat{X}(s))\mathrm{d}s \tag{3.6}$$

and

$$\hat{S}_k(t) := \hat{X}_k(0) F^c_e(t) + \hat{X}_{k,0}(t) + \hat{X}_{k,1}(t)\,. \tag{3.7}$$

Denote $\hat{R} = (\hat{R}_1, \dots, \hat{R}_K)'$ and $\hat{S} = (\hat{S}_1, \dots, \hat{S}_K)'$. Then $\hat{R}$ is a semi-martingale with respect to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$, which is a Brownian motion with a random drift. The process $\hat{S}_k$ is a continuous Gaussian process starting at $\hat{X}_k(0)$ for $k = 1, \dots, K$. Recall the representations of $\hat{X}_{k,0}(t)$ and $\hat{X}_{k,1}(t)$ in (3.2) and (3.4), respectively, using Brownian bridge $\hat{W}^0_k(t)$ and Kiefer process $\hat{\mathcal{K}}_k(t,x)$ (through the Brownian sheet $\hat{W}_k$. Also let $\hat{X}_{\cdot,0} = (\hat{X}_{1,0}, \dots, \hat{X}_{K,0})'$ and $\hat{X}_{\cdot,1} = (\hat{X}_{1,1}, \dots, \hat{X}_{K,1})'$. Similarly for $\hat{W}^0$, $\hat{\mathcal{K}}$ and $\hat{W}$. Note the mutual independence between $\hat{B}(t)$, $\hat{X}_0(t)$ and $\hat{X}_1(t)$, and hence the mutual independence between $\hat{B}(t)$, $\hat{W}^0$ and $\hat{W}$.

Then with (3.6)-(3.7), the stochastic integral equation (2.9) can be rewritten as

$$\hat{X}_k(t) = \int_0^t F^c(t-s)\mathrm{d}\hat{R}_k(s) + \hat{S}_k(t) \tag{3.8}$$

for $k = 1, \dots, K$, $t \geq 0$, where the semimartingale $\hat{R}$ depends on $\hat{X}$ as in (3.6) and the continuous Gaussian process $\hat{S}$ is independent of $\hat{B}$.

We shall first construct a weak solution. To construct a solution to this stochastic equation, we use change of measure in such a way that the semi-martingale $\hat{R}_k(s)$ becomes a Brownian motion under a new measure.

(Step 1: Construction) We consider a K-dimensional standard Brownian motion $\widetilde{\beta} := (\widetilde{\beta}_1, \dots, \widetilde{\beta}_K)$ and the independent Gaussian process $\hat{S}$ in (3.7) constructed from the independent Brownian bridges $\hat{W}^0_\cdot$ and the independent Kiefer processes $\hat{\mathcal{K}}_\cdot$ through the Brownian sheets, independent of $\widetilde{\beta}$, on a filtered probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \{\widetilde{\mathcal{F}}(t)\}_{t\geq 0}, \widetilde{\mathbb{P}})$, and we define $\xi := (\xi_1, \dots, \xi_K)$, $M := (M_1, \dots, M_K)$,

$$\xi_k(t) := \hat{S}_k(t) + \int_0^t F^c(t-s) c_k \mathrm{d}\widetilde{\beta}_k(s)\,, \quad M_k(t) \;:=\; \widetilde{\beta}_k(t) - \int_0^t \Big(\frac{\hat{\lambda}_k}{c_k} + \frac{\lambda_0}{c_k}\delta_k(\xi(s))\Big)\mathrm{d}s\,, \tag{3.9}$$

with the indicator function δ_k in (2.2) for $k = 1, \dots, K$, $t \geq 0$, and

$$Z(t) \;=\; \exp\left(\sum_{k=1}^K \int_0^t \Big(\frac{\hat{\lambda}_k}{c_k} + \frac{\lambda_0}{c_k}\delta_k(\xi(s))\Big)\mathrm{d}\widetilde{\beta}_k(s) - \frac{1}{2}\sum_{k=1}^K \int_0^t \Big(\frac{\hat{\lambda}_k}{c_k} + \frac{\lambda_0}{c_k}\delta_k(\xi(s))\Big)^2 \mathrm{d}s \right); \quad t \geq 0\,.$$

Here, M is a K-dimensional, drifted Brownian motion with at most linearly growing drifts.

Then the stochastic exponential Z is a continuous martingale under the probability measure $\widetilde{\mathbb{P}}$, and hence, for a fixed $T > 0$, we define a new probability measure $\widetilde{\mathbb{Q}}$ by

$$\frac{\mathrm{d}\widetilde{\mathbb{Q}}}{\mathrm{d}\widetilde{\mathbb{P}}}\Big|_{\widetilde{\mathcal{F}}(T)} := Z(T)\,.$$

Applying the Girsanov theorem (Theorem 3.5.1 of [12] for Brownian motions and also see e.g., Proposition 1.6 of [16] for Brownian sheets), we see M is a K-dimensional, standard Brownian motion, independent of $\hat{S}$, under the probability measure $\widetilde{\mathbb{Q}}$. Here, by the Girsanov theorem, we remove the drifts of the drifted Brownian motion but we do not shift the Brownian sheets that drive the independent Gaussian processes $\hat{S}$. Thus, under $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \{\widetilde{\mathcal{F}}(t)\}_{t\ge 0}, \widetilde{\mathbb{Q}})$, the adapted continuous process ξ, the continuous Gaussian process $\hat{S}$, and the Brownian motion M satisfy the equation

$$\xi_k(t) \,=\, \hat{S}_k(t) + \int_0^t F^c(t-s)c_k \mathrm{d}M_k(s) + \int_0^t \hat{\lambda}_k F^c(t-s)\mathrm{d}s + \lambda_0 \int_0^t F^c(t-s)\delta_k(\xi(s))\mathrm{d}s$$

for $k = 1, \dots, K$, $0 \le t \le T$. Thus, ξ, M, $\hat{S}$ in (3.9) satisfy the system (2.9) of stochastic Volterra integral equations for $0 \le t \le T$ under the new measure $\widetilde{\mathbb{Q}}$. Since $T > 0$ is arbitrary, by the above construction, there is a weak solution for $t \ge 0$ to the system (2.9) of the stochastic Volterra integral equations.

(Step 2: Uniqueness) The joint distribution of the weak solution $(\hat{X}, \hat{A}, \hat{X}_{\cdot,0}, \hat{X}_{\cdot,1})$ to (2.9) is uniquely determined by the Girsanov change of measure as in the proof of Proposition 5.3.10 of [12], that is, first we localize the problem by defining the sequence of the first passage times for $\hat{X}$ to the sphere of integer radiuses, centered at the origin, secondly we apply the Girsanov change of measure with those stopping times, and then we take the limits. Note that the Gaussian processes $(\hat{X}_{\cdot,0}, \hat{X}_{\cdot,1})$ are independent of the Brownian motion $\hat{A}$. When the initial values $\hat{X}(0)$ are randomized, we may determine the distribution as in Corollary 5.3.11 of [12].

(Step 3: proof of (3.1)) To show (3.1), we first show that for the processes ξ in (3.9) under $\widetilde{\mathbb{P}}$,

$$\int_0^T \sum_{k,\ell=1,\, k\neq \ell}^K \mathbf{1}_{\{\alpha_k\xi_k(t) = \alpha_\ell \xi_\ell(t)\}}\mathrm{d}t \,=\, 0 \tag{3.10}$$

and then once again, we apply the Girsanov theorem to show that (3.10) holds under $\widetilde{\mathbb{Q}}$ as in Step 1, and hence, the weak solution $\hat{X}$ satisfies (3.1). Thus, it suffices to show (3.10) under $\widetilde{\mathbb{P}}$ where $\hat{S}$ and $\widetilde{\beta}$ are independent. Note that since the tail probability F^c and positive constants c_k are deterministic, each integral $\int_0^t F^c(t-s)c_k\mathrm{d}\widetilde{\beta}_k(s)$ is normally distributed with mean 0 and variance $\int_0^t (F^c(t-s)c_k)^2\mathrm{d}s$ for each $k = 1, \dots, K$, independent of $\hat{S}$ under $\widetilde{\mathbb{P}}$.

It follows that for every $k \neq \ell$, $t \ge 0$, and for every fixed $(\theta_1, \dots, \theta_K) \in \mathbb{R}^K$,

$$\widetilde{\mathbb{P}}\Big(\alpha_k\Big(\theta_k + \int_0^t F^c(t-s)c_k\mathrm{d}\widetilde{\beta}_k(s)\Big) \,=\, \alpha_\ell\Big(\theta_\ell + \int_0^t F^c(t-s)c_\ell\mathrm{d}\widetilde{\beta}_\ell(s)\Big)\Big) \,=\, 0\,,$$

and hence, by the tower property of the conditional probability and by the independence, we have for any $k \neq \ell$ and $t \in [0, T]$,

$$\begin{aligned}
&\widetilde{\mathbb{P}}(\alpha_k \xi_k(t) \, = \, \alpha_\ell \xi_\ell(t)) \\
&= \widetilde{\mathbb{E}}\Big[\widetilde{\mathbb{P}}\Big(\alpha_k\Big(\hat{S}_k(t) + \int_0^t F^c(t-s) c_k \mathrm{d}\widetilde{\beta}_k(s)\Big) \, = \, \alpha_\ell\Big(\hat{S}_\ell(t) + \int_0^t F^c(t-s) c_\ell \mathrm{d}\widetilde{\beta}_\ell(s)\Big) \Big| \hat{S}(t)\Big)\Big] \\
&= \widetilde{\mathbb{E}}\Bigg[\widetilde{\mathbb{P}}\Big(\alpha_k\Big(\theta_k + \int_0^t F^c(t-s) c_k \mathrm{d}\widetilde{\beta}_k(s)\Big) \, = \, \alpha_\ell\Big(\theta_\ell + \int_0^t F^c(t-s) c_\ell \mathrm{d}\widetilde{\beta}_\ell(s)\Big)\Big)\Big|_{\substack{\theta_k = \hat{S}_k(t), \\ \theta_\ell = \hat{S}_\ell(t)}}\Bigg] \\
&= 0 \, .
\end{aligned} \tag{3.11}$$

Here, $\widetilde{\mathbb{E}}$ represents the expectation under $\widetilde{\mathbb{P}}$. Thus, we obtain (3.10) under $\widetilde{\mathbb{P}}$, because

$$\widetilde{\mathbb{E}}\Big[\int_0^T \sum_{k,\ell=1,\, k\neq\ell}^K \mathbf{1}_{\{\alpha_k\xi_k(t) = \alpha_\ell\xi_\ell(t)\}} \mathrm{d}t\Big] \, = \, \int_0^T \sum_{k,\ell=1,\, k\neq\ell}^K \widetilde{\mathbb{P}}(\alpha_k\xi_k(t) \, = \, \alpha_\ell\xi_\ell(t))\mathrm{d}t \, = \, 0 \, .$$

By the reasoning in the previous paragraph, we claim the property (3.1). □

Corollary 3.1. *Recall the conic set $\mathcal{R}_k$ defined from the indicator function δ_k in (2.2) and $\delta_k(\cdot) \, = \, \mathbf{1}_{\mathcal{R}_k}(\cdot)$ for $k = 1, \ldots, K$. We denote the closure of $\mathcal{R}_k$ by $\overline{\mathcal{R}_k}$ for every k. The stochastic equation*

$$\begin{aligned}
\hat{X}_k(t) = \hat{X}_k(0) F_e^c(t) + \hat{\lambda}_k F_e(t) + \lambda_0 \int_0^t \mathbf{1}_{\overline{\mathcal{R}}_k}(\hat{X}(s)) F^c(t-s) \mathrm{d}s \\
+ \int_0^t F^c(t-s) \mathrm{d}\hat{A}_k(s) + \hat{X}_{k,0}(t) + \hat{X}_{k,1}(t), \quad k \, = \, 1, \ldots, K
\end{aligned} \tag{3.12}$$

has a unique weak solution in $C(\mathbb{R}_+, \mathbb{R}^K)$.

Proof. Thanks to (3.1), the integrals $\int_0^\cdot \mathbf{1}_{\overline{\mathcal{R}}_k}(\hat{X}(s)) F^c(t-s)\mathrm{d}s$ and $\int_0^\cdot \delta_k(\hat{X}(s)) F^c(t-s)\mathrm{d}s$ are the same almost surely. □

4. Proof for the convergence to the limit

We have the representation

$$\begin{aligned}
\hat{X}_k^n(t) = \hat{X}_k^n(0) F_0^c(t) + \hat{\lambda}_k^n \int_0^t F^c(t-s)\mathrm{d}s + \int_0^t \delta_k(\hat{X}^n(s-)) F^c(t-s) \mathrm{d}\hat{A}_0^n(s) \\
+ \int_0^t F^c(t-s)\mathrm{d}\hat{A}_k^n(s) + \hat{X}_{k,0}^n(t) + \hat{X}_{k,1}^n(t) + \hat{X}_{k,2}^n(t),
\end{aligned} \tag{4.1}$$

where

$$\hat{X}_{k,0}^n(t) := \frac{1}{\sqrt{n}} \sum_{j=1}^{X_k^n(0)} \big(\mathbf{1}_{\eta_{k,j}^0 > t} - F_e^c(t)\big) = -\frac{1}{\sqrt{n}} \sum_{j=1}^{X_k^n(0)} \big(\mathbf{1}_{\eta_{k,j}^0 \le t} - F_e(t)\big),$$

$$\hat{X}_{k,1}^n(t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{A_k^n(t)} \big(\mathbf{1}_{\tau_{k,i}^n + \eta_{k,i} > t} - F^c(t - \tau_{k,i}^n)\big),$$

$$\hat{X}^n_{k,2}(t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{A^n_0(t)} \delta_k(X^n(\tau^n_{0,i}-))\big(\mathbf{1}_{\tau^n_{0,i}+\eta_{0,i}>t} - F^c(t-\tau^n_{0,i})\big).$$

Note that this assumption of $F_0 = F_e$ is essential since we are centering the process $\bar{X}^n_k$ by its equilibrium $\bar{X}^*_k$, and in the derivation of the representation of $X^n_k(t)$, the term $-\sqrt{n}\lambda_k F_0(t)$ cancels out the term $\sqrt{n}\lambda_k \int_0^t F^c(t-s)\mathrm{d}s$ only if $F_0 = F_e$.

The joint convergence in the following lemma follows directly from the existing results for each component in [13, 18] for the heavy-traffic analysis of $G/GI/\infty$ queues and the mutual independence of the corresponding limits.

Lemma 4.1. *Under Assumptions 1, 2 and 3,*

$$\left(\int_0^{\cdot} F^c(\cdot - s)\mathrm{d}\hat{A}^n_k(s), \hat{X}^n_{k,0}, \hat{X}^n_{k,1}\right)_{k=1,\dots,K} \Rightarrow \left(\int_0^{\cdot} F^c(\cdot - s)\mathrm{d}\hat{A}_k(s), \hat{X}_{k,0}, \hat{X}_{k,1}\right)_{k=1,\dots,K},$$

in (D^{3K}, J_1) as $n \to \infty$, where the limits $\hat{X}_{k,0}$ and $\hat{X}_{k,1}$ are given in Theorem 2.2.

Lemma 4.2. *Under Assumptions 1 and 2,*

$$\hat{X}^n_{k,2} \Rightarrow 0 \quad in \quad (D, J_1) \quad as \quad n \to \infty. \tag{4.2}$$

Proof. For each t, by conditioning on the filtration generated by the arrival process A^n_0, we have

$$\begin{aligned}
\mathbb{E}\big[\big(\hat{X}^n_{k,2}(t)\big)^2\big] &= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{A^n_0(t)} \delta_k(X^n(\tau^n_{0,i}-))F(t-\tau^n_{0,i})F^c(t-\tau^n_{0,i})\right] \\
&\le \frac{1}{\sqrt{n}}\mathbb{E}\left[\int_0^t F(t-s)F^c(t-s)\mathrm{d}\frac{A^n_0(s)}{\sqrt{n}}\right] \\
&\to 0 \quad \text{as} \quad n \to \infty,
\end{aligned}$$

where the convergence follows from Assumption 2 that $\frac{A^n_0}{\sqrt{n}} \Rightarrow \lambda_0 e$ in D.

Next, we consider the increment, for $t, u \ge 0$,

$$\begin{aligned}
\big|\hat{X}^n_{k,2}(t+u) - \hat{X}^n_{k,2}(t)\big| \le{}& \frac{1}{\sqrt{n}} \sum_{i=1}^{A^n_0(t)} \mathbf{1}_{t<\tau^n_{0,i}+\eta_{0,i}\le t+u} \\
&+ \frac{1}{\sqrt{n}} \sum_{i=1}^{A^n_0(t)} (F(t+u-\tau^n_{0,i}) - F(t-\tau^n_{0,i})) \\
&+ \frac{1}{\sqrt{n}} \sum_{i=A^n_0(t)+1}^{A^n_0(t+u)} \delta_k(X^n(\tau^n_{0,i}-))\big|\mathbf{1}_{\tau^n_{0,i}+\eta_{0,i}>t+u} - F^c(t+u-\tau^n_{0,i})\big|.
\end{aligned} \tag{4.3}$$

For the first term, since it is nondecreasing in u, we have, for $\delta > 0$ and $\epsilon > 0$,

$$\mathbb{P}\left(\sup_{u\in[0,\delta]} \frac{1}{\sqrt{n}} \sum_{i=1}^{A^n_0(t)} \mathbf{1}_{t<\tau^n_{0,i}+\eta_{0,i}\le t+u} > \epsilon/3\right)$$

$$
\begin{aligned}
&\le \frac{9}{\epsilon^2}\mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{A_0^n(t)} \mathbf{1}_{t<\tau_{0,i}^n+\eta_{0,i}\le t+\delta}\right)^2\right] \\
&\le \frac{18}{\epsilon^2}\mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{A_0^n(t)} \mathbf{1}_{t<\tau_{0,i}^n+\eta_{0,i}\le t+\delta} - (F(t+\delta-\tau_{0,i}^n) - F(t-\tau_{0,i}^n))\right)^2\right] \\
&\quad + \frac{18}{\epsilon^2}\mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{A_0^n(t)} (F(t+\delta-\tau_{0,i}^n) - F(t-\tau_{0,i}^n))\right)^2\right] \\
&= \frac{18}{\epsilon^2}\mathbb{E}\left[\frac{1}{\sqrt{n}}\int_0^t (F(t+\delta-s)-F(t-s))(1-(F(t+\delta-s)-F(t-s)))\mathrm{d}\frac{A_0^n(s)}{\sqrt{n}}\right] \\
&\quad + \frac{18}{\epsilon^2}\mathbb{E}\left[\left(\int_0^t (F(t+\delta-s)-F(t-s))\mathrm{d}\frac{A_0^n(s)}{\sqrt{n}}\right)^2\right]. \qquad (4.4)
\end{aligned}
$$

Here the first term converges to zero as $n\to\infty$, and the second term satisfies

$$
\begin{aligned}
&\frac{1}{\delta}\limsup_{N\to\infty}\sup_{t\in[0,T]}\mathbb{E}\left[\left(\int_0^t (F(t+\delta-s)-F(t-s))\mathrm{d}\frac{A_0^n(s)}{\sqrt{n}}\right)^2\right] \\
&\le \frac{1}{\delta}\sup_{t\in[0,T]}\lambda_0^2\left(\int_0^t (F(t+\delta-s)-F(t-s))\mathrm{d}s\right)^2 \\
&= \frac{1}{\delta}\sup_{t\in[0,T]}\lambda_0^2\left(\int_t^{t+\delta} F(s)\mathrm{d}s - \int_0^{\delta} F(s)\mathrm{d}s\right)^2 \\
&\le \lambda_0^2\delta, \qquad (4.5)
\end{aligned}
$$

which converges to zero as $\delta\to 0$.

For the second term in (4.3), it satisfies (4.5). Now for the third term in (4.3), it can be bounded by

$$
\frac{1}{\sqrt{n}}(A_0^n(t+u) - A_0^n(t)),
$$

which is nondecreasing in u, so that the supremum over $u\in[0,\delta]$ is bounded by $\frac{1}{\sqrt{n}}(A_0^n(t+\delta) - A_0^n(t))$. Then, by the convergence of $\frac{A_0^n}{\sqrt{n}} \Rightarrow \lambda_0 e$ in D, we obtain that for small enough δ,

$$
\limsup_{N\to\infty}\mathbb{P}\left(\sup_{u\in[0,\delta]}\frac{1}{\sqrt{n}}(A_0^n(t+u) - A_0^n(t)) > \epsilon/3\right) = 0.
$$

Thus, by the Corollary on page 83 of Billingsley [4], we have shown that

$$
\frac{1}{\delta}\limsup_{N\to\infty}\sup_{t\in[0,T]}\mathbb{P}\left(\sup_{u\in[0,\delta]}\left|\hat{X}_{k,2}^n(t+u) - \hat{X}_{k,2}^n(t)\right| > \epsilon/3\right) \to 0 \quad \text{as} \quad \delta\to 0.
$$

This completes the proof. $\square$

Proof of Theorem 2.2. We first observe that (4.1) can be rewritten as

$$\begin{aligned}\hat{Z}_k^n(t) &:= \hat{X}_k^n(t) - \int_0^t \delta_k(\hat{X}^n(s-))F^c(t-s)\mathrm{d}\hat{A}_0^n(s) \\ &= \hat{X}_k^n(0)F_0^c(t) + \hat{\lambda}_k^n \int_0^t F^c(t-s)\mathrm{d}s \\ &\qquad + \int_0^t F^c(t-s)\mathrm{d}\hat{A}_k^n(s) + \hat{X}_{k,0}^n(t) + \hat{X}_{k,1}^n(t) + \hat{X}_{k,2}^n(t)\end{aligned} \tag{4.6}$$

for $k = 1, \ldots, K$, $0 \le t \le T$, and write $\hat{Z}^n := (\hat{Z}_1^n, \ldots \hat{Z}_k^n)$. Because of the tightness of the sequence $(\hat{X}_k^n(0), \hat{\lambda}_k^n)_{n\ge1}$ in $\mathbb{R}^2$ and the tightness of the sequence

$$\Big((\hat{A}_0^n(\cdot), (\hat{A}_k^n(\cdot), \hat{X}_{k,0}^n(\cdot), \hat{X}_{k,1}^n(\cdot), \hat{X}_{k,2}^n(\cdot)), k = 1, \ldots, K)\Big)_{n\ge1}$$

in $(D_{[0,T]}^{4K+1}, J_1)$, we claim the tightness of the sequence

$$\Big(\hat{X}^n(\cdot), \hat{Z}^n(\cdot), \hat{A}_0^n(\cdot)\Big)_{n\ge1} = \Big(\hat{X}_1^n(\cdot), \ldots, \hat{X}_K^n(\cdot), \hat{Z}_1^n(\cdot), \ldots, \hat{Z}_K^n(\cdot), \hat{A}_0^n(\cdot)\Big)_{n\ge1} \tag{4.7}$$

in $(D_{[0,T]}^{2K+1}, J_1)$.

Now let us take a weak limit point $(\hat{X}^\infty(\cdot), \hat{Z}^\infty(\cdot), \hat{A}_0^\infty(\cdot))$ of the sequence (4.7) in $(D_{[0,T]}^{2K+1}, J_1)$. Without loss of generality, we may assume that the whole sequence may converge weakly to this limit point. By the Skorokhod representation theorem for the separable metric space $(D_{[0,T]}^{2K+1}, J_1)$, we may take almost sure convergence

$$\lim_{n\to\infty} (\hat{X}^n(t), \hat{Z}^n(t), \hat{A}_0^n(t)) = (\hat{X}^\infty(t), \hat{Z}^\infty(t), \hat{A}_0^\infty(t)) \tag{4.8}$$

for all but countably many t on $[0, T]$, by extending the probability space, if necessary. The filtration for the corresponding probability space is taken to be the one generated by all these processes.

The limits of the right side of (4.6), that is, the limit of $\hat{Z}^n(\cdot) = (\hat{Z}_1^n(\cdot), \ldots, \hat{Z}_K^n(\cdot))$ can be represented as

$$\begin{aligned}\hat{Z}_k^\infty(t) &= \hat{X}_k(0)F_0^c(t) + \hat{\lambda}_k \int_0^t F^c(t-s)\mathrm{d}s \\ &\qquad + \int_0^t F^c(t-s)\mathrm{d}\hat{A}_k(s) + \hat{X}_{k,0}(t) + \hat{X}_{k,1}(t)\end{aligned} \tag{4.9}$$

for $1 \le k \le K$, $0 \le t \le T$, and it is continuous on $[0, T]$. Thus, the almost sure convergence of $\hat{Z}^n$ to $\hat{Z}^\infty$ in $(D_{[0,T]}^K, J_1)$ is uniform on $[0, T]$. With the same reasoning, the almost sure convergence $\lim_{n\to\infty} \hat{A}_0^n(t) = \lambda_0 t$ is also uniform on $[0, T]$, that is,

$$\lim_{n\to\infty} \sup_{0\le t\le T} |\hat{A}_0^n(t) - \lambda_0 t| = 0 .$$

Moreover, since the right hand of (4.9) is continuous almost surely in $t \in [0, T]$, so is $\hat{Z}_k^\infty$ for each k. Note that the integrand $\delta_k(\hat{X}^n(s-))F^c(\cdot - s)$ of $\int_0^\cdot \delta_k(\hat{X}^n(s-))F^c(\cdot - s)\mathrm{d}\hat{A}_0^n(s)$ is bounded and F^c is differentiable with a bounded derivative. Then, the sequence

$\{\int_0^t \delta_k(\hat{X}^n(s-))F^c(t-s)\mathrm{d}\hat{A}_0^n(s), t \in [0,T], n \geq 1\}$ of absolutely continuous functions on $[0,T]$ is uniformly equicontinuous. Thus, the almost sure limit

$$\Phi_k(\cdot) := \lim_{n\to\infty} \int_0^{\cdot} \delta_k(\hat{X}^n(s-))F^c(\cdot - s)\mathrm{d}\hat{A}_0^n(s) \;=\; \lambda_0 \lim_{n\to\infty} \int_0^{\cdot} \delta_k(\hat{X}^n(s-))F^c(\cdot - s)\mathrm{d}s \quad (4.10)$$

is uniformly converging by the Arzelà-Ascoli theorem, and hence, it is also continuous in $t \in [0,T]$. Thus, the limit $\hat{X}_k^{\infty}(t)$ is continuous in t, because of (4.6) and of continuity of $\hat{Z}^{\infty}$. That is, $\hat{X}^{\infty}(t) = \hat{X}^{\infty}(t-)$ for every $t \in [0,T]$. Then we have the uniform convergence

$$\lim_{n\to\infty} \sup_{0\leq t\leq T} |\hat{X}^n(t) - \hat{X}^{\infty}(t)| \;=\; 0\,,$$

and $\hat{X}^{\infty}$ satisfies

$$\hat{X}_k^{\infty}(t) = \Phi_k(t) + \hat{Z}_k^{\infty}(t)\,, \quad k \;=\; 1,\ldots,K\,,\; t \geq 0\,, \quad (4.11)$$

where Φ_k and $\hat{Z}_k^{\infty}$ are given in (4.10) and (4.9), respectively.

Now we claim that each Φ_k is absolutely continuous with respect to Lebesgue measure by an application of the Riesz representation theorem for bounded linear functionals. Hence, by a similar argument of the change of measure as in the proof of Theorem 3.1, an analogue of (3.1) holds for $\hat{X}^{\infty}$, that is,

$$\int_0^T \sum_{k,\ell=1,\,k\neq\ell}^{K} \mathbf{1}_{\{\alpha_k\hat{X}_k^{\infty}(t)\,=\,\alpha_\ell\hat{X}_\ell^{\infty}(t)\}}\mathrm{d}t \;=\; \int_0^T \mathbf{1}_{\cup_{k=1}^{K}\partial\mathcal{R}_k}(\hat{X}^{\infty}(t))\mathrm{d}t \;=\; 0\,. \quad (4.12)$$

Here, $\partial\mathcal{R}_k$ is the boundary of $\mathcal{R}_k$, i.e., $\partial\mathcal{R}_k = \{x \in \mathbb{R}_+^K : \alpha_k x_k = \alpha_\ell x_\ell \text{ for some } \ell\}$ for $k = 1,\ldots,K$.

Since $\delta_k(\cdot)$ is an indicator function of conic set in (2.2), the almost sure convergence (4.8) of $\hat{X}^n$ implies that the almost convergence

$$\lim_{n\to\infty} \delta_k(\hat{X}^n(s-)) \;=\; \delta_k(\hat{X}^{\infty}(s-)) \;=\; \delta_k(\hat{X}^{\infty}(s)) \quad (4.13)$$

holds if $\hat{X}^{\infty}(s)$ is not on the boundary $\partial\mathcal{R}_k$ of the set $\mathcal{R}_k$ for $k = 1,\ldots,K$. Thus, thanks to (4.12), the almost sure limit $\Phi_k(t)$ of $\int_0^t \delta_k(\hat{X}^n(s-))F^c(t-s)\mathrm{d}\hat{A}_0^n(s)$ in (4.6) is

$$\lim_{n\to\infty} \int_0^t \delta_k(\hat{X}^n(s-))F^c(t-s)\mathrm{d}\hat{A}_0^n(s) \;=\; \lambda_0 \int_0^t \delta_k(\hat{X}^{\infty}(s-))F^c(t-s)\mathrm{d}s\,, \quad 0 \leq t \leq T\,. \quad (4.14)$$

Hence, we claim that (4.11) is reduced to

$$\hat{Z}_k^{\infty}(t) \;=\; \hat{X}_k^{\infty}(t) - \lambda_0 \int_0^t \delta_k(\hat{X}^{\infty}(s))F^c(t-s)\mathrm{d}s\,, \quad 0 \leq t \leq T\,, 1 \leq k \leq K\,.$$

In other words, because of the representation (4.9), the weak limit point $\hat{X}^{\infty}(\cdot)$ satisfies the stochastic equation (2.9). Thanks to the weak uniqueness in Theorem 3.1, the distribution of $\hat{X}^{\infty}(\cdot)$ is uniquely determined. Therefore, $\hat{X}^n(\cdot)$ converges weakly to $\hat{X}(\cdot)$ that satisfies the stochastic equation (2.9), as $n \to \infty$. □

5. Appendix

Proof of Remark 2.2. Recall the expression of $\hat{X}_k(t)$ in (2.10). We have

$$
\begin{aligned}
&-\hat{X}_k(0)F^c(t) + \hat{\lambda}_k F^c(t) + \lambda_0\delta_k(\hat{X}(t)) - \lambda_0 \int_0^t \delta_k(\hat{X}(s))f(t-s)\mathrm{d}s \\
&= -\hat{X}_k(0)e^{-t} + \hat{\lambda}_k e^{-t} + \lambda_0\delta_k(\hat{X}(t)) - \lambda_0 \int_0^t \delta_k(\hat{X}(s))e^{-(t-s)}\mathrm{d}s \\
&= \hat{\lambda}_k + \lambda_0\delta_k(\hat{X}(t)) - \left(\hat{X}_k(0)e^{-t} + \hat{\lambda}_k(1-e^{-t}) + \lambda_0 \int_0^t \delta_k(\hat{X}(s))e^{-(t-s)}\mathrm{d}s\right). \qquad (5.1)
\end{aligned}
$$

Denoting

$$
\hat{X}^A_k(t) = \int_0^t F^c(t-s)\mathrm{d}\hat{A}_k(s) = \int_0^t e^{-(t-s)}c_k\mathrm{d}\hat{B}_k(t)\,,
$$

we obtain

$$
\hat{X}^A_k(t) = -\int_0^t \hat{X}^A_k(s)\mathrm{d}s + c_k\hat{B}_k(t)\,. \qquad (5.2)
$$

Recall the representations of $\hat{X}_{k,0}$ in (3.3) and $\hat{X}_{k,1}$ in (3.5).

For the stochastic terms in (5.2), (3.3) and (3.5), since they are mutually independent, we obtain that the covariance function of their summation at times $0 \le t < t'$ is equal to

$$
\begin{aligned}
&\mathrm{Cov}(c_kB_k(t), c_kB_k(t')) + \mathrm{Cov}(\lambda_k^{1/2}\hat{B}_k(1-e^{-t}), \lambda_k^{1/2}\hat{B}_k(1-e^{-t'})) \\
&\quad + \mathrm{Cov}\left(\int_0^t\int_0^t \mathbf{1}_{s+x\le t}\mathrm{d}\hat{W}_k(\lambda_k s, 1-e^{-x}), \int_0^{t'}\int_0^{t'} \mathbf{1}_{s+x\le t'}\mathrm{d}\hat{W}_k(\lambda_k s, 1-e^{-x})\right) \\
&= c_k^2 t + \lambda_k(1-e^{-t}) + \lambda_k(t-(1-e^{-t})) = (\lambda_k + c_k^2)t\,.
\end{aligned}
$$

That is, the three stochastic terms in (5.2), (3.3) and (3.5) are equivalent in distribution with a Brownian motion with variance coefficient $\lambda_k + c_k^2$. In the case of the renewal arrival process, $c_k^2 = \lambda_k^3\sigma_k^2 = \lambda_k\mathrm{SCV}_k$ with $\mathrm{SCV}_k = \lambda_k^2\sigma_k^2$ being the squared coefficient of variation of the interarrival times, so the variance coefficient $\lambda_k + c_k^2 = \lambda_k(1+\mathrm{SCV}_k)$. If in addition, the arrival processes are Poisson, then $\mathrm{SCV}_k = 1$, and the variance coefficient $\lambda_k + c_k^2 = 2\lambda_k$.

Finally, combining (5.1), (5.2), (3.3) and (3.5), we obtain the expression in (2.11). □

Acknowledgement

T. Ichiba was supported by NSF Grant DMS 2008427 and G. Pang was supported by NSF Grant DMS 2108683/2216765.

References

[1] S. A. Almada Monter, M. Shkolnikov, and J. Zhang. Dynamics of observables in rank-based models and performance of functionally generated portfolios. *Ann. Appl. Probab.*, 29(5):2849–2883, 2019.

[2] R. F. Bass and È. Pardoux. Uniqueness for diffusions with piecewise constant coefficients. *Probability Theory and Related Fields*, 76(4):557–572, 1987.

[3] M. A. Berger and V. J. Mizel. Volterra equations with Itô integrals I. *The Journal of Integral Equations*, 2(3):187–245, 1980.

[4] P. Billingsley. *Convergence of probability measures*. John Wiley & Sons, 1999.

[5] Y.-J. Chao. Weak convergence of a sequence of semimartingales to a diffusion with discontinuous drift and diffusion coefficients. *Queueing systems*, 42(2):153–188, 2002.

[6] M. V. der Boor, S. C. Borst, J. S. Van Leeuwaarden, and D. Mukherjee. Scalable load balancing in networked systems: A survey of recent advances. *SIAM Review*, 64(3):554–622, 2022.

[7] E. R. Fernholz. *Stochastic portfolio theory*. Springer, 2002.

[8] A. F. Filippov. *Differential equations with discontinuous righthand sides*, volume 18 of *Mathematics and its Applications (Soviet Series)*. Kluwer Academic Publishers Group, Dordrecht, 1988. Translated from the Russian.

[9] P. J. Fleming and B. Simon. Heavy traffic approximations for a system of infinite servers with load balancing. *Probability in the Engineering and Informational Sciences*, 13(3):251–273, 1999.

[10] T. Ichiba, I. Karatzas, and M. Shkolnikov. Strong solutions of stochastic equations with rank-based coefficients. *Probability Theory and Related Fields*, 156(1-2):229–248, 2013.

[11] T. Ichiba, V. Papathanakos, A. Banner, I. Karatzas, and R. Fernholz. Hybrid atlas models. *Ann. Appl. Probab.*, 21(2):609–644, 2011.

[12] I. Karatzas and S. E. Shreve. *Brownian motion and stochastic calculus*. Springer Science & Business Media, 1991.

[13] E. V. Krichagina and A. A. Puhalskii. A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Systems*, 25(1-4):235–280, 1997.

[14] N. V. Krylov. On weak uniqueness for some diffusions with discontinuous coefficients. *Stochastic Processes and their applications*, 113(1):37–64, 2004.

[15] N. V. Krylov and R. Liptser. On diffusion approximation with discontinuous coefficients. *Stochastic Processes and Their Applications*, 102(2):235–264, 2002.

[16] D. Nualart and E. Pardoux. Markov field properties of solutions of white noise driven quasi-linear parabolic PDEs. *Stochastics Stochastics Rep.*, 48(1-2):17–44, 1994.

[17] G. Pang, R. Talreja, and W. Whitt. Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys*, 4:193–267, 2007.

[18] G. Pang and W. Whitt. Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems*, 65:325–364, 2010.

[19] G. R. Shorack and J. A. Wellner. *Empirical processes with applications to statistics*. SIAM, 2009.

[20] A. J. Viterbi. *CDMA: principles of spread spectrum communication*. Addison Wesley Longman Publishing Co., Inc., 1995.

[21] W. Whitt. *Stochastic-process limits: an introduction to stochastic-process limits and their application to queues*. Springer, 2002.

[22] H. Yao and C. Knessl. On the infinite server shortest queue problem: symmetric case. *Stochastic models*, 21(1):101–132, 2005.

[23] H. Yao and C. Knessl. On the infinite server shortest queue problem: Non-symmetric case. *Queueing Systems*, 52:157–177, 2006.