Heavy Traffic Diffusion Limit for a Closed Queueing Network with Single-Server and Infinite-Server Stations

Amir A. Alwan[a], Barış Ata[b]

[a]*Lubar College of Business, University of Wisconsin–Milwaukee*
[b]*Booth School of Business, The University of Chicago*

Abstract. This paper studies the limiting behavior of a closed queueing network with multiple single-server and infinite-server stations. Under a heavy traffic asymptotic regime—where the number of jobs and single-server service rates grow large while infinite-server rates remain fixed—we prove a weak convergence result for the queue length and idleness process vector, providing an approximation for the original system.

Keywords: stochastic process limits; queueing theory; heavy traffic analysis; closed queueing network; infinite-server queues

1. Introduction

This paper studies the limiting behavior of a closed queueing network consisting of multiple single-server and infinite-server stations under a heavy traffic asymptotic regime. Jobs circulate perpetually through the system according to a two-level probabilistic routing structure: from single-server stations to infinite-server stations, and from infinite-server stations back to the single-server stations.

The asymptotic regime is one in which both the number of jobs in the system and the service rates at the single-server stations grow large, while the service rates at the infinite-server stations remain fixed. In this regime, we further assume that the system is in heavy traffic, meaning that each single-server station is critically loaded: the nominal arrival rate to each station matches its service capacity. This reflects a setting in which congestion arises at the single-server stations, while the infinite-server stations act as delay buffers.

Our main result is a functional limit theorem for the diffusion-scaled queue length and idleness process vector, where the limit process solves a Skorokhod problem. To establish this, we prove existence, uniqueness, and continuity of the associated nonlinear regulator mapping—results which are of interest in their own right. Unlike several earlier works on the weak convergence of closed networks with infinite-server stations, which have employed martingale techniques, we use a continuous mapping argument.

The model was motivated in part by ride-hailing systems such as Uber and Lyft. In this context, the single-server stations represent regions of a city where drivers await ride requests, while the infinite-server stations capture travel times between regions. The service rates at the single-server stations represent customer arrival rates in each region, while the service rates at the infinite-server stations represent the travel time completion rates. After picking up a passenger (i.e., after a service completion at a single-server station), a driver enters a travel phase to reach the destination, and upon completion (i.e., after a service completion at an infinite-server station), becomes available again in a potentially new region. After picking up a passenger, a driver enters a travel phase to reach the destination, and upon completion, becomes available again in a potentially new region. The two-level routing structure in our model naturally reflects such origin-destination dynamics, and the presence of multiple infinite-server stations allows for heterogeneity in travel times across different region pairs. While ride-hailing serves as a useful motivation, the model and results apply more broadly to systems in which tasks are queued for assignment or dispatch at specific locations, and then processed or delayed in parallel elsewhere. By establishing a rigorous diffusion approximation for such systems, this paper lays the theoretical groundwork for future research on their dynamic control and optimization.

Ride-hailing systems are often modeled as open, with drivers entering and leaving the platform and hence a time-varying fleet size; for example, Özkan and Ward [22] models a ride-hailing system as an open two-sided matching system. Our model is closed and is most appropriate over time horizons in which the fleet size is approximately constant; over longer horizons, the diffusion approximation can be applied in a piecewise fashion by updating the fleet size between intervals. Demand can also be non-stationary, particularly in the presence of dynamic pricing. Although this paper focuses on time-homogeneous primitives, such non-stationarity can be captured through time-varying service rates at the single-server stations, and we expect the convergence results to extend to such a time-inhomogeneous setting, although the mathematical details are beyond the scope of this paper. In particular, second-order adjustments to the effective demand rates would appear as drift terms in the limit and naturally give rise to drift-rate control problems; see, e.g., Alwan et al. [1]. Extending the convergence results to open networks with external arrivals and departures, while allowing jobs to recirculate through the network, together with infinite-server travel-time nodes, is left for future work.

1.1. Literature Review

This paper contributes to the literature on heavy traffic diffusion approximations for queueing networks, with a particular focus on closed systems involving infinite-server stations. Classical work by Iglehart [17], Borovkov [8], and

Whitt and Glynn [26] establishes diffusion limits for open queueing networks with infinite-server queues. Other papers have studied the structure and properties of these limiting processes. For example, Glynn [13] establishes necessary and sufficient conditions under which the heavy-traffic limit of the $GI/G/\infty$ queue is Markovian. Foundational work on heavy-traffic approximations for stochastic flow networks with bottlenecks includes Chen and Mandelbaum [10, 11], which develop both fluid and diffusion limits. Various surveys, such as Glynn [14] and Pang et al. [23] discuss a variety of techniques for deriving diffusion approximations in queueing systems; see the references therein for a broader overview.

In the context of closed queueing networks with infinite-server stations, several papers establish diffusion limits under heavy traffic conditions. Notably, Kogan et al. [20], Krichagina [21], and Kogan and Lipster [19] analyze models with many single-server stations and a single infinite-server station. These papers have a single probability routing vector and use martingale methods to prove convergence. Our paper supplements this line of research by allowing an arbitrary number of infinite-server stations and a two-level probabilistic routing structure. (This enables more realistic modeling of origin-destination dynamics and heterogeneous travel times, which are important to service systems such as ride-hailing.) Our analysis also differs methodologically: we use a continuous mapping approach and establish existence, uniqueness, and continuity of a nonlinear regulator mapping—results that are not readily implied by existing work. To the best of our knowledge, this is the first paper to rigorously establish a diffusion limit for a closed queueing network with both multiple infinite-server stations and a two-level routing structure.

We note that the network studied here is a closed queueing network with Markovian routing and service times (i.e., a Gordon–Newell network) with a fixed number of jobs n, and therefore admits a product-form stationary distribution. Using this stationary distribution, however, requires computing the normalizing constant, i.e., the partition function $G(n)$, which is a sum over all ways of allocating the n jobs across stations; the number of such allocations grows combinatorially with the number of jobs and stations. This quickly becomes computationally difficult to evaluate in large networks. While this challenge has motivated work on asymptotic and algorithmic methods for approximating $G(n)$ in large closed networks, exact computation can still be burdensome when both the number of jobs and the number of stations are large; see, e.g., Birman and Kogan [7] and Kogan [18]. In contrast, the diffusion limit developed here provides a tractable process-level approximation for $\sqrt{n}$-scale fluctuations, capturing transient behavior and serving as a natural basis for diffusion-scale control and optimization in heavy traffic.

A closely related and important antecedent of our model is the one developed in Alwan et al. [1], which considers a ride-hailing system modeled as a closed queueing network with multiple single-server regions and a single infinite-server travel node. The objective is to maximize system profit by making dynamic pricing and dispatch control decisions. That paper formulates an approximating Brownian control problem (BCP) using formal limiting arguments in the spirit of Harrison [15, 16]. Under a complete resource pooling assumption, the BCP is reduced to a one-dimensional equivalent workload formulation, which is then solved analytically. However, the assumption of a single infinite-server node effectively enforces homogeneous travel times and origin-independent destination routing. This modeling simplification was necessary to permit a one-dimensional state space collapse, but such a reduction is not possible when the model includes multiple infinite-server nodes.

In contrast, our model accommodates multiple infinite-server stations and a two-level routing structure that allows for heterogeneous travel times and captures general origin-destination routing distributions. (Notably, Braverman et al. [9] considers a related model with empty-car routing control but studies it under fluid scaling, which avoids the complexities associated with diffusion-scaled fluctuations.) Our setting leads to a multidimensional diffusion limit. By rigorously proving a weak convergence result in this setting, our work provides the theoretical foundation for studying more complex control policies and generalizes the modeling framework introduced in Alwan et al. [1].

As mentioned earlier, the presence of multiple infinite-server stations does not permit a one-dimensional state space collapse, implying that the limiting diffusion control problem remains high-dimensional. However, recent advancements in computing power have renewed interest in studying high-dimensional diffusion control problems using approximate dynamic programming and neural network-based methods. One expects the dynamic control problem associated with the system here to involve drift and singular control of (reflecting) Brownian motion in the heavy-traffic limit. Such problems have been successfully addressed in dimensions up to 100 or more in recent work by Ata et al. [2, 4], and Ata and Xu [5]. While those papers consider different applications, we expect their computational methods to be applicable to ride-hailing applications as well. Although the study of dynamic control of ride-hailing systems in high dimensions is left for future work, our paper lays the mathematical groundwork for future research in stochastic control of such systems, including those where dimensionality precludes analytical tractability.

Beyond ride-hailing, our results are also relevant to other applications modeled using queueing networks with infinite-server stations. For example, Ata et al. [3] studies nonprofit volunteer management using a queueing network with a single-server station and multiple infinite-server stations. They formulate an approximating Brownian control problem but lack rigorous justifications for their approximations. Our results help close this gap by providing provable limit theorems in a similar setting.

The rest of the paper is organized as follows. Section 2 introduces the model primitives and makes a sample-path construction of the queue-length processes describing the network. Section 3 articulates the heavy traffic assumption and the asymptotic regime, and states the main result of the paper (Theorem 1). Section 4 develops the key tools needed to prove the main result. Section 5 is devoted to a proof of the main result, which involves proving weak convergence of the fluid-scaled processes. Relevant notation and technical preliminaries are given shortly. Additional proofs are provided in Appendices A, B, C, and D.

Notation and Technical Preliminaries. The set of positive integers is denoted by $\mathbb{N} := \{1, 2, 3, \dots\}$, and we define $[k] := \{1, 2, \dots, k\}$ for $k \in \mathbb{N}$. For $a, b \in \mathbb{R}$, we let $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$, and let $\lfloor a \rfloor$ denote the largest integer less than or equal to a. For $l \in [k]$, the lth unit basis vector in $\mathbb{R}^k$ is denoted by e_l, which has one in the lth component and zeros elsewhere. Moreover, for $l \in [k]$, the lth projection map $\pi_l : \mathbb{R}^k \to \mathbb{R}$ is given by $\pi_l(x) = x_l$, where x_l is the lth component of $x \in \mathbb{R}^k$. For $k \in \mathbb{N}$, the positive orthant in $\mathbb{R}^k$ is denoted by $\mathbb{R}^k_+ := \{x \in \mathbb{R}_k : x_l \geq 0 \text{ for all } l \in [k]\}$. For a function $f : X \to Y$ and a subset $S \subseteq X$, we denote by $f|_S$ the restriction of f to S. The indicator function of a subset $S \subseteq X$ is denoted by $\mathbb{1}_S$.

For $k \in \mathbb{N}$, we denote by $D^k \equiv D\left([0, \infty), \mathbb{R}^k\right)$ the set of all functions $x : [0, \infty) \to \mathbb{R}^k$ that are right continuous on $[0, \infty)$ and have left limits on $(0, \infty)$. We denote by $C^k \equiv C\left([0, \infty), \mathbb{R}^k\right)$ the set of all functions $x : [0, \infty) \to \mathbb{R}^k$ that are continuous. The identically zero function is denoted by $\mathbf{0}$. Similarly, for $k \in \mathbb{N}$ and $T > 0$, we denote by $D^k_T \equiv D\left([0, T], \mathbb{R}^k\right)$ the set of all functions $x : [0, T] \to \mathbb{R}^k$ that are right continuous on $[0, T)$ and have left limits on $(0, T]$. When the space D^k_T is endowed with the norm

$$\|x\|_{T,k} := \max_{l \in [k]} \sup_{t \in [0,T]} |x_l(t)|,$$

it is a Banach space. When $k = 1$, we write $D^1 = D$, $D^1_T = D_T$, and $\|\cdot\|_{T,1} = \|\cdot\|_T$. The one-sided reflection map on D is the pair of functions $(\psi, \phi) : D \to D^2$ defined as follows:

$$\psi(x)(t) := \sup_{s \in [0,t]} [-x(s)]^+, \quad t \geq 0, \tag{1}$$

$$\phi(x)(t) := x(t) + \psi(x)(t), \quad t \geq 0. \tag{2}$$

The space D^k is a topological space when endowed with the Skorokhod J_1 topology. All random variables in this paper are defined on a common probability space $(\Omega, \mathcal{F}, P)$. We denote by $\mathcal{M}^k$ the Borel σ-algebra on D^k induced by the Skorokhod J_1 topology. All stochastic processes in this paper are measurable functions from $(\Omega, \mathcal{F}, P)$ to $(D^k, \mathcal{M}^k)$ for some appropriate dimension k. For a sequence of stochastic processes $\{\xi^n\}_{n=1}^{\infty}$ and a stochastic process ξ, all with sample paths in D^k almost surely, we write $\xi^n \Rightarrow \xi$ as $n \to \infty$ to mean that the sequence of probability measures on $(D^k, \mathcal{M}^k)$ induced by the processes ξ^n converge weakly to the probability measure on $(D^k, \mathcal{M}^k)$ induced by the stochastic process ξ; see Billingsley [6] and Whitt [25] for further details.

2. Queueing Network Model

This section describes the closed queueing network model. In Section 2.1, we introduce the system primitives. Then, in Section 2.2, we describe the system performance measures and their dynamics.

2.1. Model Primitives

We consider a closed queueing network consisting of a fixed number of jobs circulating among J single-server stations and K infinite-server stations. Each single-server station $j \in [J]$ has a buffer where jobs of that class are stored and a single server with service rate of $\mu_j > 0$. Each infinite-server station $k \in [K]$ has an unlimited number of parallel servers, each with a service rate of $\eta_k > 0$. For future analysis, it is useful to define $\eta := \max_{k \in [K]} \eta_k$.

Jobs move through the system according to a two-level probabilistic routing structure: from single-server stations to infinite-server stations, and from infinite-server stations back to single-server stations. When a job completes service at single-server station $j \in [J]$, it is routed to infinite-server station $k \in [K]$ with probability $p_{jk} \in [0, 1]$. Similarly, when a job completes service at infinite-server station $k \in [K]$, it is routed to single-server station $j \in [J]$ with probability $q_{kj} \in [0, 1]$. To formalize this, we define stochastic matrices

$$P = (p_{jk}) \in \mathbb{R}^{J \times K}, \quad Q = (q_{kj}) \in \mathbb{R}^{K \times J} \tag{3}$$

representing routing from single-server stations to infinite-server stations and from infinite-server stations to single-server stations, respectively.

The stochastic evolution of the system is driven by the following primitives. For $j \in [J]$, let $N_j = \{N_j(t) : t \geq 0\}$ be a unit-rate Poisson process governing service completions at single-server station j. For $k \in [K]$, let $M_k = \{M_k(t) : t \geq 0\}$ be a unit-rate Poisson process governing service completions at infinite-server station k. To model the probabilistic routing of jobs in the network, let $\phi_j = \{\phi_j(l) : l \geq 1\}$ for $j \in [J]$ and $\psi_k = \{\psi_k(l) : l \geq 1\}$ for $k \in [K]$ denote sequences of i.i.d. random (routing) vectors. Their probability distributions are given by

$$P(\phi_j(1) = e_k) = p_{jk}, \quad P(\psi_k(1) = e_j) = q_{kj},$$

where e_k and e_j are the kth and jth standard unit basis vector in $\mathbb{R}^K$ and $\mathbb{R}^J$, respectively. We assume that all stochastic primitives are mutually independent.

To track the cumulative routing of jobs, for $j \in [J]$ and $k \in [K]$, we define the cumulative routing processes $\Phi_{jk} = \{\Phi_{jk}(m) : m \geq 1\}$ and $\Psi_{kj} = \{\Psi_{kj}(m) : m \geq 1\}$ as follows:

$$\Phi_{jk}(m) := \sum_{l=1}^{m} \phi_{jk}(l), \quad \Psi_{kj}(m) := \sum_{l=1}^{m} \psi_{kj}(l), \tag{4}$$

where $\phi_{jk}(l)$ and $\psi_{kj}(l)$ are the kth and jth components of $\phi_j(l)$ and $\psi_k(l)$, respectively. That is, $\Phi_{jk}(m)$ represents the total number of jobs that are routed from single-server station j to infinite-server station k among the first m jobs served by server j. The interpretation of $\Psi_{kj}(m)$ is similar.

2.2. State Dynamics

For $j \in [J]$ we denote by $Q_j(t)$ the number of jobs in the buffer of the jth single-server station at time t. Similarly, for $k \in [K]$ we denote by $V_k(t)$ the number of jobs in the kth infinite-server station at time t. To describe the state dynamics of Q_j and V_k, let $Q_j(0)$ for $j \in [J]$ and $V_k(0)$ for $k \in [K]$ be almost surely nonnegative random variables representing the initial distribution of the number of jobs in the network. We assume that these random variables are independent of all other stochastic primitives. Then, for $j \in [J]$ and $k \in [K]$, we have that

$$Q_j(t) = Q_j(0) + A_j(t) - D_j(t), \quad t \geq 0, \tag{5}$$

$$V_k(t) = V_k(0) + E_k(t) - F_k(t), \quad t \geq 0, \tag{6}$$

where $A_j = \{A_j(t) : t \geq 0\}$ and $D_j = \{D_j(t) : t \geq 0\}$ denote the arrival and departure processes for single-server station j. Similarly, $E_k = \{E_k(t) : t \geq 0\}$ and $F_k = \{F_k(t) : t \geq 0\}$ denote the arrival and departure processes for infinite-server station k. To be more specific, the arrival and departure processes are defined as follows:

$$A_j(t) := \sum_{k=1}^{K} \Psi_{kj}\big(F_k(t)\big), \qquad t \geq 0, \tag{7}$$

$$E_k(t) := \sum_{j=1}^{J} \Phi_{jk}\big(D_j(t)\big), \qquad t \geq 0, \tag{8}$$

$$D_j(t) := N_j\big(\mu_j T_j(t)\big), \qquad t \geq 0, \tag{9}$$

$$F_k(t) := M_k\Big(\eta_k \int_0^t V_k(s)\, ds\Big), \quad t \geq 0, \tag{10}$$

where Φ_{jk} and Ψ_{kj} are given by (4), and $T_j = \{T_j(t) : t \geq 0\}$ is a process that represents the scheduling policy for the server at the jth single-server station. To be more specific, $T_j(t)$ is the cumulative amount of time the server is working up to time t at single-server station j. The corresponding idleness process $I_j = \{I_j(t) : t \geq 0\}$ for $j \in [J]$ of the server at single-server station j is defined as follows:

$$I_j(t) := t - T_j(t), \quad t \geq 0. \tag{11}$$

By (5)–(10), it is straightforward to verify that

$$\sum_{j=1}^{J} Q_j(t) + \sum_{k=1}^{K} V_k(t) = \sum_{j=1}^{J} Q_j(0) + \sum_{k=1}^{K} V_k(0) \tag{12}$$

almost surely for all $t \geq 0$. Throughout our analysis, we restrict to scheduling policies T that satisfy the following conditions almost surely for all $j \in [J]$:

$$Q_j(t) \in [0, \infty) \text{ for all } t \geq 0, \tag{13}$$

$$\int_0^\infty \mathbb{1}_{\{Q_j(t) > 0\}}\, dI_j(t) = 0, \tag{14}$$

$$I_j \text{ is continuous and nondecreasing with } I_j(0) = 0, \tag{15}$$

$$I_j(t) - I_j(s) \leq t - s \text{ for all } 0 \leq s \leq t < \infty. \tag{16}$$

Equation (13) reflects the obvious physical restriction that the servers can only work on jobs when the queues are not empty. Equation (14) implies that the server idleness at each single-server station does not increase as long as the queue is not empty. In other words, we restrict attention to (otherwise arbitrary) work-conserving scheduling policies. Finally, (15)–(16) are natural consequences of the interpretation of I_j as server idleness and are standard in the heavy traffic literature.

To lighten the notation, we henceforth work pathwise on a full-measure set on which the initial conditions, the scheduling-policy constraints, and the primitives' usual path properties hold for all $t \geq 0$; we therefore omit the "almost surely" terminology throughout.

3. Heavy Traffic Assumption and Main Result

As is standard in heavy traffic asymptotic analysis, we consider a sequence of queueing networks—as described in Section 2—indexed by the system parameter n. We study this sequence of systems in the heavy traffic asymptotic regime as $n \to \infty$. Throughout, we attach a superscript of n to the various quantities of interest to indicate that they correspond to the nth system.

We consider a regime in which both the number of jobs and the service rates at the single-server stations grow large with the system parameter n, and where the system satisfies a critical loading condition. However, we assume that the service rates at the infinite-server stations do not vary with the system parameter n. This regime is formalized by the following heavy traffic assumption. To state it, define

$$m_k := \eta_k^{-1} \sum_{j=1}^{J} \mu_j p_{jk}, \quad k \in [K]. \tag{17}$$

Assumption 1 (Heavy Traffic Assumption)**.** *The service rates at the single-server stations and infinite-server stations, respectively, vary with n as follows: $\mu_j^n = n\mu_j$ for $j \in [J]$ and $\eta_k^n = \eta_k$ for $k \in [K]$. Moreover, the following critical loading conditions hold:*

$$\sum_{k=1}^{K} \sum_{i=1}^{J} \mu_i p_{ik} q_{kj} = \mu_j \quad \textit{for all} \quad j \in [J], \tag{18}$$

$$\sum_{k=1}^{K} m_k = 1. \tag{19}$$

Roughly speaking, condition (18) assumes that every single-server is fully utilized, whereas (19) roughly says that almost all jobs are in the infinite-server stations; see Alwan et al. [1] and Ata et al. [3] for similar assumptions in ride-hailing and volunteer engagement applications, respectively. For simplicity, Assumption 1 focuses on the zero-drift case. However, this assumption can be refined to give a nonzero drift term in the heavy-traffic limit. In particular, suppose that the service rates at the single-server stations in the sequence satisfy

$$\sqrt{n}\Big[\sum_{k=1}^{K}\sum_{i=1}^{J}\mu_i p_{ik} q_{kj} \;-\; n^{-1}\mu_j^n\Big] \to c_j \in \mathbb{R} \quad \text{as} \quad n \to \infty,$$

for all $j \in [J]$. Under this condition, our main result, i.e., Theorem 1, remains valid, except that the limiting diffusion for the single-server queue-length process acquires an additional drift term $c_j t$.

To shed light on (18), note that $\sum_{i=1}^{J} n\mu_i p_{ik}$ represents the total rate of jobs leaving the single-server stations to infinite-server station k. Thus, $\sum_{k=1}^{K} q_{kj} \sum_{i=1}^{J} n\mu_i p_{ik}$ represents the total rate of jobs entering single-server station j from the infinite-server stations. Because $n\mu_j$ is the service rate of the server in single-server station j, (18) states that the rate of jobs entering the single-server stations are balanced out by the service rates at the stations.

To shed light on (19), note that $n\sum_{j=1}^{J}\mu_j p_{jk}$ represents the arrival rate to the kth infinite-server station, whereas its service rate is η_k. Based on intuition from the classical $M/M/\infty$ queue, we expect the steady-state average queue length at the kth infinite-server to be $n\sum_{j=1}^{J}\mu_j p_{jk}/\eta_k$. It follows that the expected fraction of jobs at the kth infinite-server is m_k. Therefore, (19) states that almost all jobs are at the infinite-server stations. This is consistent with the first condition: It is well known that in large balanced-flow systems in heavy traffic, the queue-length processes are of second-order relative to the system size. Thus, the total number of jobs in the single-server stations are of order $\sqrt{n}$, which implies that the total number of jobs at the infinite-server stations are of order n, since the network is closed.

To facilitate the analysis to follow, we next define the following diffusion- and fluid-scaled processes:

Diffusion-Scaled Processes: For $j \in [J]$ and $k \in [K]$, we define the following diffusion-scaled processes:

$$\hat{Q}_j^n(t) := n^{-1/2} Q_j^n(t), \qquad t \geq 0, \tag{20}$$

$$\hat{V}_k^n(t) := n^{-1/2}\big(V_k^n(t) - nm_k\big), \qquad t \geq 0, \tag{21}$$

$$\hat{I}_j^n(t) := \sqrt{n} I_j^n(t), \qquad t \geq 0, \tag{22}$$

$$\hat{T}_j^n(t) := \sqrt{n} T_j^n(t), \qquad t \geq 0, \tag{23}$$

$$\hat{\Phi}_{jk}^n(t) := n^{-1/2}\big(\Phi_{jk}(\lfloor nt \rfloor) - p_{jk} nt\big), \quad t \geq 0, \tag{24}$$

$$\hat{\Psi}_{kj}^n(t) := n^{-1/2}\big(\Psi_{kj}(\lfloor nt \rfloor) - q_{kj} nt\big), \quad t \geq 0, \tag{25}$$

$$\hat{N}_j^n(t) := n^{-1/2}\big(N_j(nt) - nt\big), \qquad t \geq 0, \tag{26}$$

$$\hat{M}_k^n(t) := n^{-1/2}\big(M_k(nt) - nt\big), \qquad t \geq 0. \tag{27}$$

Fluid-Scaled Processes: For $j \in [J]$ and $k \in [K]$, we define the following fluid-scaled processes:

$$\bar{Q}_j^n(t) := n^{-1} Q_j^n(t), \qquad t \geq 0, \tag{28}$$

$$\bar{V}_k^n(t) := n^{-1/2} \hat{V}_k^n(t) \qquad t \geq 0, \tag{29}$$

$$\bar{\bar{V}}_k^n(t) := n^{-1} V_k^n(t), \qquad t \geq 0, \tag{30}$$

$$\bar{N}_j^n(t) := n^{-1} N_j(nt), \qquad t \geq 0 \tag{31}$$

$$\bar{M}_k^n(t) := n^{-1} M_k(nt), \quad t \geq 0. \tag{32}$$

By (5)–(6), (20)–(32), and Assumption 1, it is straightforward verify that for $j \in [J]$, $k \in [K]$, and all $t \geq 0$ the following equalities hold:

$$\hat{Q}_j^n(t) = \hat{\xi}_j^n(t) + \sum_{k=1}^{K} q_{kj}\eta_k \int_0^t \hat{V}_k^n(s)\,ds + \mu_j \hat{I}_j^n(t), \tag{33}$$

$$\hat{V}_k^n(t) = \hat{\zeta}_k^n(t) - \eta_k \int_0^t \hat{V}_k^n(s)\,ds - \sum_{j=1}^{J} p_{jk}\mu_j \hat{I}_j^n(t), \tag{34}$$

where

$$\begin{aligned}\hat{\xi}_j^n(t) := \hat{Q}_j^n(0) &+ \sum_{k=1}^{K} \hat{\Psi}_{kj}^n\Big(\bar{M}_k^n\big(\eta_k \int_0^t \bar{\bar{V}}_k^n(s)\,ds\big)\Big) \\ &- \hat{N}_j^n\big(\mu_j T_j^n(t)\big) + \sum_{k=1}^{K} q_{kj} \hat{M}_k^n\Big(\eta_k \int_0^t \bar{\bar{V}}_k^n(s)\,ds\Big),\end{aligned} \tag{35}$$

$$\begin{aligned}\hat{\zeta}_k^n(t) := \hat{V}_k^n(0) &+ \sum_{j=1}^{J} \hat{\Phi}_{jk}^n\big(\bar{N}_j^n(\mu_j T_j^n(t))\big) \\ &- \hat{M}_k^n\Big(\eta_k \int_0^t \bar{\bar{V}}_k^n(s)\,ds\Big) + \sum_{j=1}^{J} p_{jk} \hat{N}_j^n(\mu_j T_j^n(t)).\end{aligned} \tag{36}$$

Also, by (28)–(29) and (33)–(34), it is straightforward to verify that for all $t \geq 0$ the following equalities hold:

$$\bar{Q}_j^n(t) = \bar{\xi}_j^n(t) + \sum_{k=1}^{K} q_{kj}\eta_k \int_0^t \bar{V}_k^n(s)\,ds + \mu_j I_j^n(t), \tag{37}$$

$$\bar{V}_k^n(t) = \bar{\zeta}_k^n(t) - \eta_k \int_0^t \bar{V}_k^n(s)\,ds - \sum_{j=1}^{J} p_{jk}\mu_j I_j^n(t), \tag{38}$$

where

$$\bar{\xi}_j^n(t) := n^{-1/2}\hat{\xi}_j^n(t) \quad \text{and} \quad \bar{\zeta}_k^n(t) := n^{-1/2}\hat{\zeta}_k^n(t). \tag{39}$$

We make the following regularity assumption on the initial conditions:

Assumption 2 (Joint Convergence of the Initial Conditions)**.** *As $n \to \infty$, $(\hat{Q}^n(0), \hat{V}^n(0)) \Rightarrow (Q(0), V(0))$.*

To facilitate the statement of our main result, let (ξ^*, ζ^*) be a $(J+K)$-dimensional Brownian motion with initial state $(Q(0), V(0))$ and covariance matrix $\Sigma \in \mathbb{R}^{(J+K)\times(J+K)}$, where Σ is given as follows: For $i, j \in [J]$ and $k, l \in [K]$ such that $i \neq j$ and $k \neq l$, we have

$$\begin{aligned}\Sigma_{j,j} \quad = \quad & \sum_{k=1}^{K} q_{kj}(1-q_{kj})\eta_k m_k \\ & + \mu_j + \sum_{k=1}^{K} q_{kj}^2 \eta_k m_k,\end{aligned} \tag{40}$$

$$\begin{aligned}\Sigma_{J+k,J+k} &= \sum_{j=1}^{J} p_{jk}(1-p_{jk})\mu_j \\ &\qquad + \eta_k m_k + \sum_{j=1}^{J} p_{jk}^2 \mu_j, \end{aligned}\tag{41}$$

$$\Sigma_{i,j} = \sum_{k=1}^{K} q_{ki} q_{kj} \eta_k m_k, \tag{42}$$

$$\Sigma_{j,J+k} = -p_{jk}\mu_j - q_{kj}\eta_k m_k, \tag{43}$$

$$\Sigma_{J+l,J+k} = \sum_{j=1}^{J} p_{jl} p_{jk} \mu_j. \tag{44}$$

Theorem 1. *As $n \to \infty$, $(\hat{Q}^n, \hat{I}^n, \hat{V}^n) \Rightarrow (Q^*, I^*, V^*)$, where (Q^*, I^*, V^*) is a $(2J+K)$-dimensional process with continuous sample paths in $\mathbb{R}_+^{2J} \times \mathbb{R}^K$ that satisfies the following equalities for all $j \in [J]$, $k \in [K]$, and $t \geq 0$:*

$$Q_j^*(t) = \xi_j^*(t) + \sum_{k=1}^{K} q_{kj}\eta_k \int_0^t V_k^*(s)\,ds + \mu_j I_j^*(t), \tag{45}$$

$$V_k^*(t) = \zeta_k^*(t) - \eta_k \int_0^t V_k^*(s)\,ds - \sum_{j=1}^{J} p_{jk}\mu_j I_j^*(t), \tag{46}$$

$$I_j^*(t) = \mu_j^{-1}\psi\Big(\xi_j^* + \sum_{k=1}^{K} q_{kj}\eta_k \int_0^{\cdot} V_k^*\,ds\Big)(t), \tag{47}$$

$$\int_0^\infty \mathbb{1}_{\{Q_j^*(t)>0\}}\,dI_j^*(t) = 0. \tag{48}$$

4. Auxiliary Results

This section establishes the existence of (suitably defined) continuous functions that will aid in the proof of Theorem 1 via a continuous mapping argument. To that end, let $\xi \in D^J$ and $\zeta \in D^K$ be functions such that

$$\sum_{j=1}^{J} \xi_j(t) + \sum_{k=1}^{K} \zeta_k(t) = 0, \quad \text{for all } t \geq 0, \tag{49}$$

$$\xi_j(0) \geq 0, \qquad \text{for all } j \in [J]. \tag{50}$$

Given such functions ξ and ζ, consider the following system of equations for $j \in [J]$, $k \in [K]$, and $t \geq 0$:

$$x_j(t) = \xi_j(t) + \sum_{k=1}^{K} q_{kj}\eta_k \int_0^t y_k(s)\,ds + \mu_j u_j(t) \geq 0, \tag{51}$$

$$y_k(t) = \zeta_k(t) - \eta_k \int_0^t y_k(s)\,ds - \sum_{j=1}^{J} p_{jk}\mu_j u_j(t), \tag{52}$$

$$\sum_{j=1}^{J} x_j(t) + \sum_{k=1}^{K} y_k(t) = 0, \tag{53}$$

$$u_j \text{ is nondecreasing with } u_j(0) = 0, \tag{54}$$

$$\int_0^\infty \mathbb{1}_{\{x_j(t)>0\}}\,du_j(t) = 0. \tag{55}$$

The following result establishes the existence and uniqueness of a triple (x, u, y) satisfying the above equations.

Proposition 1. *For every $(\xi, \zeta) \in D^{J+K}$ satisfying (49)–(50), there exists a unique $(x, u, y) \in D^{2J+K}$ satisfying (51)–(55).*

Proof. See Appendix B. □

The next result is immediate from Proposition 1.

Corollary 1. *There exists a function $f : D^{J+K} \to D^{2J+K}$ such that whenever $(\xi, \zeta) \in D^{J+K}$ satisfies (49)–(50), $f(\xi, \zeta) \in D^{2J+K}$ satisfies (51)–(55).*

The following result is useful in the proof of Proposition 1. To state it, given functions $\xi \in D^J$ and $\zeta \in D^K$, consider the following equation for $k \in [K]$ and $t \geq 0$:

$$\begin{aligned} y_k(t) &= \zeta_k(t) - \eta_k \int_0^t y_k(s)\,ds \\ &\qquad - \sum_{j=1}^{J} p_{jk}\psi\Big(\xi_j + \sum_{l=1}^{K} q_{lj}\eta_l \int_0^{\cdot} y_l(s)\,ds\Big)(t). \end{aligned}\tag{56}$$

Lemma 1. *For each $(\xi, \zeta) \in D^{J+K}$, there exists a unique $y \in D^K$ satisfying (56).*

Proof. See Appendix A. □

Below we provide a description of the function f from Corollary 1. Let $f_3 : D^{J+K} \to D^K$ be the mapping that sends $(\xi, \zeta) \in D^{J+K}$ to the unique $y \in D^K$ satisfying (56); see Lemma 1. Then, following the proof of Proposition 1, let $f_1 : D^{J+K} \to D^J$ and $f_2 : D^{J+K} \to D^J$ be the mappings defined as follows:

$$\begin{aligned} &f_1(\xi, \zeta) \\ &:= \Big(\phi\big(\pi_j \circ \xi + \sum_{l=1}^{K} q_{lj}\eta_l \int_0^{\cdot} \big(\pi_l \circ f_3(\xi, \zeta)\big)(s)\,ds\big)\Big)_{j \in [J]}, \end{aligned}\tag{57}$$

$$\begin{aligned} &f_2(\xi, \zeta) \\ &:= \Big(\mu_j^{-1}\psi\big(\pi_j \circ \xi + \sum_{l=1}^{K} q_{lj}\eta_l \int_0^{\cdot} \big(\pi_l \circ f_3(\xi, \zeta)\big)(s)\,ds\big)\Big)_{j \in [J]}. \end{aligned}\tag{58}$$

Let $f := (f_1, f_2, f_3)$. Then, whenever $(\xi, \zeta) \in D^{J+K}$ satisfies (49)–(50), $f(\xi, \zeta) \in D^{2J+K}$ satisfies (51)–(55). The next result establishes continuity of f.

Proposition 2. *The function $f : D^{J+K} \to D^{2J+K}$ is continuous when both the domain and range are endowed with the Skorokhod J_1 topology.*

Proof. See Appendix C. □

Proposition 3. *The function f maps C^{J+K} into C^{2J+K}.*

Proof. Following the same argument as in the proof of Lemma 1, but now with $(\xi, \zeta) \in C^{J+K}$, we can construct a sequence $\{y^n : n = 0, 1, \dots\}$ in C^K (via the method of successive approximations) whose limit $y \in C^K$ is the unique solution to (56). It follows that $f_3(C^{J+K}) \subseteq C^K$. It then follows from (57) and (58), together with the fact that the operations involved preserve continuity, that $f_1(C^{J+K}) \subseteq C^J$ and $f_2(C^{J+K}) \subseteq C^J$, respectively. Therefore, $f(C^{J+K}) \subseteq C^{2J+K}$, as desired. □

5. Main Convergence Results

This section contains the main convergence results of this paper, culminating with a proof of Theorem 1. In Section 5.1, we prove convergence of the fluid scaled processes. (These results are necessary because several of the fluid scaled processes serve as random time changes in the diffusion-scaled equations.) In Section 5.2, we prove convergence of the process $(\hat{\xi}^n, \hat{\zeta}^n)$. This, combined with a continuous mapping argument, allows us to complete the proof of Theorem 1.

5.1. Convergence of Fluid Scaled Processes

We begin by establishing weak convergence of the fluid-scaled processes.

Lemma 2. *As $n \to \infty$, $(\bar{\xi}^n, \bar{\zeta}^n) \Rightarrow \mathbf{0} \in D^{J+K}$.*

Proof. To prove that $(\bar{\xi}^n, \bar{\zeta}^n) \Rightarrow \mathbf{0}$ as $n \to \infty$, it suffices to show that $\bar{\xi}^n_j \Rightarrow 0$ as $n \to \infty$ for all $j \in [J]$ and $\bar{\xi}^n_k \Rightarrow 0$ as $n \to \infty$ for all $k \in [K]$; see, e.g., Whitt [25, Theorem 11.4.5]. In turn, it suffices to show that for all $T > 0$,

$$\|\bar{\xi}^n_j\|_T \Rightarrow 0 \quad \text{and} \quad \|\bar{\zeta}^n_k\|_T \Rightarrow 0 \quad \text{as} \quad n \to \infty \tag{59}$$

for all $j \in [J]$ and $k \in [K]$; see Lemma 6 in Appendix D for a proof of this claim. By (35)–(36), the triangle inequality, and the fact that $\int_0^t \bar{V}^n_k(s)\,ds \le t$ and $T^n_j(t) \le t$ for all $t \ge 0$, it follows that for all $T > 0$,

$$\begin{aligned} \|\hat{\xi}^n_j\|_T \le{}& \|\hat{Q}^n_j(0)\|_T + \sum_{k=1}^{K} \|\hat{\Psi}^n_{kj}\big(\bar{M}^n_k(\eta_k \cdot)\big)\|_T \\ &+ \|\hat{N}^n_j(\mu_j \cdot)\|_T + \sum_{k=1}^{K} \|\hat{M}^n_k(\eta_k \cdot)\|_T, \end{aligned} \tag{60}$$

$$\begin{aligned} \|\hat{\zeta}^n_k\|_T \le{}& \|\hat{V}^n_k(0)\|_T + \sum_{j=1}^{J} \|\hat{\Phi}^n_{jk}\big(\bar{N}^n_j(\mu_j \cdot)\big)\|_T \\ &+ \|\hat{M}^n_k(\eta_k \cdot)\|_T + \sum_{j=1}^{J} \|\hat{N}^n_j(\mu_j \cdot)\|_T. \end{aligned} \tag{61}$$

By Donsker's theorem, the functional central limit theorem for renewal processes, the random time change theorem, and the continuous mapping theorem, it is straightforward to show that the right-hand sides of (60) and (61) converge weakly to nondegenerate limits; see, e.g., Billingsley [6] and Glynn [14]. By this and the fact that $\bar{\xi}^n_j = n^{-1/2}\hat{\xi}^n_j$ and $\bar{\zeta}^n_k = n^{-1/2}\hat{\zeta}^n_k$, we obtain (59). This is a standard argument, so the detailed proof is omitted. □

Lemma 3. *As $n \to \infty$, $(\bar{Q}^n, I^n, \bar{V}^n) \Rightarrow \mathbf{0} \in D^{J+K}$*

Proof. It is straightforward to show that the process $(\bar{\xi}^n, \bar{\zeta}^n)$ defined by (39) satisfies

$$\bar{\xi}^n_j(0) = \bar{Q}^n_j(0) \ge 0, \qquad \text{for all } j \in [J], \tag{62}$$

$$\sum_{j=1}^{J} \bar{\xi}^n_j(t) + \sum_{k=1}^{K} \bar{\zeta}^n_k(t) = 0, \quad \text{for all } t \ge 0. \tag{63}$$

Furthermore, by (14)–(15) and (28), I^n is nondecreasing with $I^n(0) = 0$ and satisfies

$$\int_0^\infty \mathbb{1}_{\{\bar{Q}^n_j(t) > 0\}}\, dI^n_j(t) = \int_0^\infty \mathbb{1}_{\{Q^n_j(t) > 0\}}\, dI^n_j(t) = 0. \tag{64}$$

Since $(\bar{\xi}^n, \bar{\zeta}^n) \in D^{J+K}$ pathwise, it follows from (37)–(38), (62)–(64), and Proposition 1 that $(\bar{Q}^n, I^n, \bar{V}^n) = (f_1(\bar{\xi}^n, \bar{\zeta}^n), f_2(\bar{\xi}^n, \bar{\zeta}^n), f_3(\bar{\xi}^n, \bar{\zeta}^n))$. Then, by Proposition 2, Lemma 2, and the continuous mapping theorem,

$$\begin{aligned} (\bar{Q}^n, I^n, \bar{V}^n) &= \big(f_1(\bar{\xi}^n, \bar{\zeta}^n), f_2(\bar{\xi}^n, \bar{\zeta}^n), f_3(\bar{\xi}^n, \bar{\zeta}^n)\big) \\ &\Rightarrow \big(f_1(\mathbf{0}), f_2(\mathbf{0}), f_3(\mathbf{0})\big). \end{aligned}$$

It now suffices to show that $\bar{V} := f_3(\mathbf{0}) = \mathbf{0}$, for then it follows from (57)–(58) that $f_1(\mathbf{0}) = \mathbf{0}$ and $f_2(\mathbf{0}) = \mathbf{0}$. To that end, it follows from (56), the definition of $\bar{V}$, and the triangle inequality that, for any fixed $T > 0$, all $t \in [0, T]$, and all $k \in [K]$, we have

$$\begin{aligned} \|\bar{V}_k\|_t &\le \eta \int_0^t \|\bar{V}_k\|_s\, ds + \sum_{j=1}^{J}\sum_{l=1}^{K} \eta_l \Big\| \int_0^{\cdot} \bar{V}_l(s)\, ds \Big\|_t \\ &\le 2\eta J K \int_0^t \max_{k \in [K]} \|\bar{V}_k\|_s\, ds. \end{aligned}$$

Therefore, it follows that

$$\max_{k \in [K]} \|\bar{V}_k\|_t \le 2\eta J K \int_0^t \max_{k \in [K]} \|\bar{V}_k\|_s\, ds. \tag{65}$$

By Gronwall's inequality (see, e.g., Pang et al. [23, Lemma 4.1]) and (65), it follows that $\max_{k \in [K]} \|\bar{V}_k\|_T = 0$. Since T was arbitrary, it follows that $\bar{V} \equiv 0$. □

Corollary 2. *As $n \to \infty$, $T^n \Rightarrow e \in C^J$, where $e(t) := (t, \ldots, t)$ for $t \ge 0$.*

Proof. By the definition in (11), $T^n = e - I^n$. The result then follows by Lemma 3 since $I^n \Rightarrow \mathbf{0}$ as $n \to \infty$. □

5.2. Convergence of Diffusion Scaled Processes

The next result establishes weak convergence of the diffusion-scaled "primitive" processes $\hat{\xi}^n$ and $\hat{\zeta}^n$:

Lemma 4. *As $n \to \infty$, $(\hat{\xi}^n, \hat{\zeta}^n) \Rightarrow (\xi^*, \zeta^*)$, where (ξ^*, ζ^*) is $(J + K)$-dimensional Brownian motion with initial state $(Q(0), V(0))$ and covariance matrix Σ given by (40)–(44).*

Proof. By Lemma 3, Corollary 2, Donsker's theorem, the functional central limit theorem for renewal processes, and the continuous mapping theorem, we get weak convergence as $n \to \infty$ for the following fluid-scaled processes $j \in [J]$ and $k \in [K]$:

$$\bar{M}^n_k(\eta_k \cdot) \Rightarrow \eta_k e, \quad \bar{N}^n_j(\mu_j \cdot) \Rightarrow \mu_j e, \quad T^n_j \Rightarrow e, \quad \bar{\bar{V}}^n_k \Rightarrow m_k,$$

where $e : [0, \infty) \to [0, \infty)$ denote the one-dimensional identity map $e(t) = t$ for $t \geq 0$. Similarly, we get weak convergence as $n \to \infty$ for the following diffusion-scaled processes for $j \in [J]$ and $k \in [K]$:

$$\begin{aligned} \hat{\Psi}^n_{kj} &\Rightarrow \sqrt{q_{kj}(1-q_{kj})}\, B_{kj}, \quad & \hat{\Phi}^n_{kj} &\Rightarrow \sqrt{p_{jk}(1-p_{jk})}\, \tilde{B}_{jk}, \\ \hat{M}^n_k(\eta_k \cdot) &\Rightarrow \sqrt{\eta_k}\, B_k, & \hat{N}^n_j(\mu_j \cdot) &\Rightarrow \sqrt{\mu_j}\, \tilde{B}_j, \end{aligned}$$

where B_{kj}, $\tilde{B}_{jk}$, B_k, and $\tilde{B}_j$ are independent standard Brownian motions. (These convergence arguments are routine and therefore omitted for brevity.) Furthermore, the mapping $H : D \to D$, defined by $H(x)(t) := \int_0^t x(s)\, ds$ for $(x, t) \in D \times [0, \infty)$, is continuous in the Skorokhod J_1 topology (see, e.g., Pang et al. [23, page 229]), which implies that

$$H(\bar{\bar{V}}^n_k) \Rightarrow H(m_k) = m_k e \quad \text{as} \quad n \to \infty.$$

By the above weak convergence results, Whitt [25, Theorems 11.4.4 and 11.4.5], Assumption 2, and the independence of the stochastic model primitives, it follows that the (joint) processes $(\hat{Q}^n(0), \hat{V}^n(0), \hat{\Psi}^n, \hat{\Phi}^n, \hat{N}^n, \ldots, \hat{M}^n)$ and $(T^n, \bar{\bar{V}}^n, \bar{N}^n, \bar{M}^n)$ converge weakly as $n \to \infty$ to their appropriate limits. From this, (35)–(36), the random time change theorem, and the continuous mapping theorem, it follows that $(\hat{\xi}^n, \hat{\zeta}^n)$ converges weakly as $n \to \infty$ to $(J + K)$-dimensional Brownian motion (ξ^*, ζ^*) with initial state $(Q(0), V(0))$ and covariance matrix Σ given by (40)–(44). Because it is straightforward, albeit tedious, to derive the entries of the covariance matrix Σ, we omit the details. □

Proof of Theorem 1. It is straightforward to show that the process $(\hat{\xi}^n, \hat{\zeta}^n)$ defined by (35)–(36) satisfies

$$\hat{\xi}^n_j(0) = \hat{Q}^n_j(0) \geq 0, \qquad \text{for all } j \in [J], \tag{66}$$

$$\sum_{j=1}^{J} \hat{\xi}^n_j(t) + \sum_{k=1}^{K} \hat{\zeta}^n_k(t) = 0, \quad \text{for all } t \geq 0. \tag{67}$$

Moreover, as in the proof of Lemma 3, it is straightforward to show that $\hat{I}^n_j$ is nondecreasing with $\hat{I}^n_j(0) = 0$ and satisfies

$$\int_0^\infty \mathbb{1}_{\{\hat{Q}^n_j(t) > 0\}}\, d\hat{I}^n_j(t) = 0. \tag{68}$$

Since $(\hat{\xi}^n, \hat{\zeta}^n) \in D^{J+K}$ pathwise, it follows from (33)–(34), (66)–(68), and Proposition 1 that $(\hat{Q}^n, \hat{I}^n, \hat{V}^n) = f(\hat{\xi}^n, \hat{\zeta}^n)$. Then, by Proposition 2, Lemma 4, and the continuous mapping theorem, it follows that as $n \to \infty$,

$$\begin{aligned} (\hat{Q}^n, \hat{I}^n, \hat{V}^n) &= \left(f_1(\hat{\xi}^n, \hat{\zeta}^n), f_2(\hat{\xi}^n, \hat{\zeta}^n), f_3(\hat{\xi}^n, \hat{\zeta}^n) \right) \\ &\Rightarrow (Q^*, I^*, V^*), \end{aligned}$$

where $Q^* := f_1(\xi^*, \zeta^*)$, $I^* := f_2(\xi^*, \zeta^*)$, and $V^* := f_3(\xi^*, \zeta^*)$. Moreover, since inequalities are preserved under weak convergence, Lemma 4 and (66)–(67) imply that

$$\xi^*_j(0) = Q_j(0) \geq 0, \qquad \text{for all } j \in [J], \tag{69}$$

$$\sum_{j=1}^{J} \xi^*_j(t) + \sum_{k=1}^{K} \zeta^*_k(t) = 0, \quad \text{for all } t \geq 0. \tag{70}$$

Therefore, since $(\xi^*, \zeta^*) \in C^{J+K}$ pathwise by Lemma 4, it follows from (69)–(70), Corollary 1, and Proposition 3 that (Q^*, I^*, V^*) satisfies (45)–(48) and has continuous sample paths. This completes the proof. □

References

[1] Alwan, A. A., Ata, B., Zhou, Y. (2024). A Queueing Model of Dynamic Pricing and Dispatch Control for Ride-Hailing Systems Incorporating Travel Times. *Queueing Systems*, **106**(1–2):1–66.

[2] Ata, B., Harrison, J. M., Si, N. (2024). Singular Control of (Reflected) Brownian Motion: A Computational Method Suitable for Queueing Applications. *Queueing Systems*, **108**(3):215–251.

[3] Ata, B., Tongarlak, M. H., Lee, D., Field, J. (2024). A Dynamic Model for Managing Volunteer Engagement. *Operations Research*, **72**(5):1958–1975.

[4] Ata, B., Harrison, J. M., Si, N. (2025). Drift Control of High-Dimensional RBM: A Computational Method Based on Neural Networks. *Stochastic Systems*, **15**(2):111–146.

[5] Ata, B., Xu, Y. (2025). Dynamic Control of Stochastic Matching Systems in Heavy Traffic: An Effective Computational Method for High-Dimensional Problems. Working Paper.

[6] Billingsley, P. (1999). Convergence of Probability Measures. John Wiley & Sons, New York.

[7] Birman, A., Kogan, Y. (1992). Asymptotic Evaluation of Closed Queueing Networks with Many Stations. *Communications in Statistics. Stochastic Models*, **8**(3):543–563.

[8] Borovkov, A. A. (1967). On Limit Laws for Service Processes in Multi-Channel Systems. *Sibirskii Matematicheskii Zhurnal*, **8**(5):983–1004.

[9] Braverman, A., Dai, J. G., Liu, X., Ying, L. (2019). Empty-Car Routing in Ridesharing Systems. *Operations Research*, **67**(5):1437–1452.

[10] Chen, H., Mandelbaum, A. (1991). Discrete Flow Networks: Bottleneck Analysis and Fluid Approximations. *Mathematics of Operations Research*, **16**(2):223–446.

[11] Chen, H., Mandelbaum, A. (1991). Stochastic Discrete Flow Networks: Diffusion Approximations and Bottlenecks. *The Annals of Probability*, **19**(4):1463–1519.

[12] Ethier, S., Kurtz, T. (2005). Markov Processes: Characterization and Convergence. John Wiley & Sons, New York.

[13] Glynn, P. W. (1982). On the Markov Property of the $GI/G/\infty$ Limit. *Advances in Applied Probability*, **14**(2):191–194.

[14] Glynn, P. W. (1990). Chapter 4: Diffusion Approximations. In Handbooks on OR & MS, Vol. 2, Stochastic Models. North-Holland, Amsterdam, 145–198.

[15] Harrison, J. M. (1988). Brownian Models of Queueing Networks with Heterogeneous Customer Populations. In Stochastic Differential Systems, Stochastic Control Theory and Applications. Springer, 147–186.

[16] Harrison, J. M. (2003). A Broader View of Brownian Networks. *The Annals of Applied Probability*, **13**(3):1119–1150.

[17] Iglehart, D. L. (1965). Limiting Diffusion Approximations for the Many Server Queue and the Repairman Problem. *Journal of Applied Probability*, **2**(2):429–441.

[18] Kogan, Y. (1992). Another Approach to Asymptotic Expansions for Large Closed Queueing Models. *Operations Research Letters*, **11**(5):317–321.

[19] Kogan, Y., Lipster, R. (1993). Limit Non-Stationary Behavior of Large Closed Queueing Networks with Bottlenecks. *Queueing Systems*, **14**(1–2):33–55.

[20] Kogan, Y., Liptser, R., Smorodinskii, A. V. (1986). Gaussian Diffusion Approximation of Closed Markov Models of Computer Networks. *Problems of Information Transmission*, **22**(1):38–51.

[21] Krichagina, E. V. (1992). Asymptotic Analysis of Queueing Networks. *Stochastics and Stochastics Reports*, **40**(1–2):43–76.

[22] Özkan, E., Ward, A. R. (2020). Dynamic Matching for Real-Time Ride Sharing. *Stochastic Systems*, **10**(1):29–70.

[23] Pang, G., Talreja, R., Whitt, W. (2007). Martingale Proofs of Many-Server Heavy-Traffic Limits for Markovian Queues. *Probability Surveys*, **4**:193–267.

[24] Reed, J., Ward, A. R. (2004). A Diffusion Approximation for a Generalized Jackson Network with Reneging. *Proceedings of the 42nd Annual Conference on Communication, Control, and Computing*, **2**:983–994.

[25] Whitt, W. (2002). Stochastic-Process Limits. Springer-Verlag, New York.

[26] Whitt, W., Glynn, P. W. (1991). A New View on the Heavy-Traffic Limit Theorem for Infinite-Server Queues. *Advances in Applied Probability*, **23**(1):188–209.

Appendix

A. Proof of Lemma 1

We prove that for each $T > 0$, there exists a unique $y \in D_T^K$ satisfying (56) for all $t \in [0, T]$, then extend this solution to D^K in the obvious way. To improve the readability of the argument, we break the proof into a few separate steps, organized as subsections.

A.1. Existence of an element in D_K^T satisfying (56) for all $t \in [0, T]$

We prove existence via the method of successive approximations; see, e.g., Reed and Ward [24] for a similar proof approach. In particular, we construct a sequence that is Cauchy in D_K^T under the sup norm, and then argue that the limit of the sequence (which exists by completeness of D_K^T) satisfies (56).

Let $y_k^0 \equiv 0$, and define $y_k^n \in D$ for $n \in \mathbb{N}$ iteratively as follows for each $k \in [K]$:

$$\begin{aligned} y_k^n := \xi_k - \eta_k \int_0^{\cdot} y_k^{n-1}(s)\,ds \\ - \sum_{j=1}^{J} p_{jk}\psi\Big(\xi_j + \sum_{l=1}^{K} q_{lj}\eta_l \int_0^{\cdot} y_l^{n-1}\,ds\Big), \end{aligned} \tag{71}$$

Then, the sequence $\{(y_1^n|_{[0,T]}, \dots, y_K^n|_{[0,T]}) : n = 0, 1, \dots\}$ of elements in D_T^K defined by (71) is a Cauchy sequence with respect to the sup norm; see Lemma 5 at the end of this subsection for a proof of this claim. Therefore, by completeness of $(D_T^K, \|\cdot\|_{T,K})$, it follows that as $n \to \infty$,

$$\big(y_1^n|_{[0,T]}, \dots, y_K^n|_{[0,T]}\big) \to \big(y_{1,T}^{\infty}, \dots, y_{K,T}^{\infty}\big) \in D_T^K. \tag{72}$$

We claim that $(y_{1,T}^{\infty}, \dots, y_{K,T}^{\infty})$ satisfies (56) for all $t \in [0, T]$. To show this, consider the mapping $L : D_T^K \to D_T^K$ defined as follows:

$$\begin{aligned} &(y_1, \dots, y_K) \\ &\mapsto \Big(\zeta_k - \eta_k \int_0^{\cdot} y_k(s)\,ds \\ &\qquad - \sum_{j=1}^{J} p_{jk}\psi\Big(\xi_j + \sum_{l=1}^{K} q_{lj}\eta_l \int_0^{\cdot} y_l(s)\,ds\Big)\Big)_{k\in[K]}. \end{aligned}$$

Then, for $y, \tilde{y} \in D_T^K$, it follows by the definition of L, the triangle inequality, and Whitt [25, Lemma 13.5.1] that

$$\begin{aligned} &\|L(y) - L(\tilde{y})\|_{T,K} \\ &\quad\le \max_{k\in[K]} \Big\{\eta T\|y_k - \tilde{y}_k\|_T + \sum_{j=1}^{J}\sum_{l=1}^{K} \eta T\|y_l - \tilde{y}_l\|_T\Big\} \\ &\quad\le \eta T(JK + 1)\|y - \tilde{y}\|_{T,K}, \end{aligned}$$

implying that L is Lipschitz continuous. It then follows from (71)–(72) that

$$\begin{aligned} \big(y_{1,T}^{\infty}, \dots, y_{K,T}^{\infty}\big) &\leftarrow \big(y_1^{n+1}|_{[0,T]}, \dots, y_K^{n+1}|_{[0,T]}\big) \\ &= L\big(y_1^n|_{[0,T]}, \dots, y_K^n|_{[0,T]}\big) \\ &\to L\big(y_{1,T}^{\infty}, \dots, y_{K,T}^{\infty}\big), \end{aligned}$$

as $n \to \infty$. By uniqueness of limits in metric spaces, it follows that $L(y_{1,T}^{\infty}, \dots, y_{K,T}^{\infty}) = (y_{1,T}^{\infty}, \dots, y_{K,T}^{\infty})$, implying that $(y_{1,T}^{\infty}, \dots, y_{K,T}^{\infty})$ satisfies (56) for all $t \in [0, T]$.

We conclude this subsection with a proof showing that the sequence defined by (71) is Cauchy:

Lemma 5. *For each $T > 0$, the sequence*

$$\big\{\big(y_1^n|_{[0,T]}, \dots, y_K^n|_{[0,T]}\big) : n = 0, 1, \dots\big\}$$

defined by (71) is Cauchy in D_K^T with respect to $\|\cdot\|_{T,K}$.

Proof. Fix $\delta \in (0, T)$ such that $2\delta\eta JK < 1$. (This choice of δ will be used crucially later.) First, we claim that

$$\|y_k^n - y_k^{n-1}\|_\delta \le (2\delta\eta JK)^{n-1}C_\delta, \tag{73}$$

for all $n \in \mathbb{N}$ and $k \in [K]$, where $C_t := \max_{k\in[K]}\|\zeta_k\|_t + \sum_{j=1}^{J}\|\psi(\xi_j)\|_t$ for $t > 0$. (Note that $C_{t_1} \le C_{t_2}$ for all $0 \le t_1 \le t_2 < \infty$.) To prove this, note that when $n = 1$,

$$\begin{aligned} \|y_k^1 - y_k^0\|_\delta &= \Big\|\zeta_k - \sum_{j=1}^{J} p_{jk}\psi(\xi_j)\Big\|_\delta \\ &\le \|\zeta_k\|_\delta + \sum_{j=1}^{J}\|\psi(\xi_j)\|_\delta \le C_\delta. \end{aligned}$$

for all $k \in [K]$. Moreover, note that when $n \ge 2$,

$$\begin{aligned} &\|y_k^n - y_k^{n-1}\|_\delta \\ &\quad\le \eta_k\delta\|y_k^{n-1} - y_k^{n-2}\|_\delta \\ &\qquad + \sum_{j=1}^{J} p_{jk}\Big\|\sum_{l=1}^{K} q_{lj}\eta_l \int_0^{\cdot} \big(y_l^{n-1}(s) - y_l^{n-2}(s)\big)\,ds\Big\|_\delta \\ &\quad\le 2\delta\eta J \sum_{l=1}^{K}\|y_l^{n-1} - y_l^{n-2}\|_\delta, \end{aligned}$$

for all $k \in [K]$, where the first inequality follows from Whitt [25, Lemma 13.5.1] and (71). By performing a similar estimate of $\|y_l^{n-1} - y_l^{n-2}\|_\delta$ in the above display, it follows that

$$\|y_k^n - y_k^{n-1}\|_\delta \le (2\delta\eta J)^2 K \sum_{l=1}^{K} \|y_l^{n-2} - y_l^{n-3}\|_\delta,$$

for all $k \in [K]$. By continuing in this way, we obtain (73). Next, we claim that for each $k \in [K]$ that

$$\|y_k^n - y_k^{n-1}\|_{m\delta} \le m^2 n^m (2\delta\eta J K)^{n-1} C_{m\delta}. \tag{74}$$

for all $m, n \in \mathbb{N}$. Fixing $k \in [K]$, we consider the $n = 1$ and $n \ge 2$ cases separately. When $n = 1$, we have that $\|y_k^1 - y_k^0\|_{m\delta} \le C_{m\delta} \le m^2 C_{m\delta}$ for all $m \in \mathbb{N}$ by (73), so that (74) holds. When $n \ge 2$, we proceed by (strong) induction on m. The base case of $m = 1$ holds immediately by (73). For the inductive step, assume that for all $n \ge 2$,

$$\|y_k^n - y_k^{n-1}\|_{r\delta} \le r^2 n^r (2\delta\eta J K)^{n-1} C_{r\delta}, \tag{75}$$

for all $r \in [m]$. Thus, by (75), it follows that for $n \ge 2$,

$$\begin{aligned}
&\|y_k^n - y_k^{n-1}\|_{(m+1)\delta} \\
&\le \eta_k \sum_{r=1}^{m+1} \int_{(r-1)\delta}^{r\delta} \|y_k^{n-1} - y_k^{n-2}\|_{r\delta}\, ds \\
&\qquad + \sum_{j=1}^{J} p_{jk} \Big\| \sum_{l=1}^{K} q_{lj} \eta_l \int_0^{\cdot} |y_l^{n-1}(s) - y_l^{n-2}(s)|\, ds \Big\|_{(m+1)\delta} \\
&= \delta\eta \Big[\sum_{r=1}^{m+1} \|y_k^{n-1} - y_k^{n-2}\|_{r\delta} \\
&\qquad\qquad + J \sum_{l=1}^{K} \sum_{r=1}^{m+1} \|y_l^{n-1} - y_l^{n-2}\|_{r\delta} \Big] \\
&\le 2\delta\eta J \sum_{l=1}^{K} \sum_{r=1}^{m} r(n-1)^r (2\delta\eta J K)^{n-2} C_{m\delta} \\
&\qquad\qquad + 2\delta\eta J \sum_{l=1}^{K} \|y_l^{n-1} - y_l^{n-2}\|_{(m+1)\delta} \\
&= (2\delta\eta J K)^{n-1} C_{r\delta} \sum_{r=1}^{m} r(n-1)^r \\
&\qquad\qquad + 2\delta\eta J \sum_{l=1}^{K} \|y_l^{n-1} - y_l^{n-2}\|_{(m+1)\delta} \\
&\le (2\delta\eta J K)^{n-1} C_{m\delta} m^2 n^m \\
&\qquad\qquad + 2\delta\eta J \sum_{l=1}^{K} \|y_l^{n-1} - y_l^{n-2}\|_{(m+1)\delta}.
\end{aligned}$$

Continuing from the previous display, it follows from (73) that for $n = 2$,

$$\begin{aligned}
&\|y_k^2 - y_k^1\|_{(m+1)\delta} \\
&\quad \le (2\delta\eta J K) C_{m\delta} m^2 2^m + 2\delta\eta J \sum_{l=1}^{K} \|y_l^1 - y_l^0\|_{(m+1)\delta} \\
&\quad \le (2\delta\eta J K) C_{(m+1)\delta} (m^2 2^m + 1).
\end{aligned}$$

Continuing in this way, it follows that for $n \ge 2$,

$$\begin{aligned}
&\|y_k^n - y_k^{n-1}\|_{(m+1)\delta} \\
&\quad \le (2\delta\eta J K)^{n-1} C_{(m+1)\delta} \big(m^2 \sum_{i=2}^{n} i^m + 1 \big) \\
&\quad \le (m+1) n^{m+1} (2\delta\eta J K)^{n-1} C_{(m+1)\delta}.
\end{aligned}$$

This completes the inductive step. This proves that (74) holds for all $m, n \in \mathbb{N}$.

In particular, by (74), we have that

$$\begin{aligned}
&\|y_k^n - y_k^{n-1}\|_T \\
&\quad \le \|y_k^n - y_k^{n-1}\|_{\lceil \delta^{-1} T \rceil \delta} \\
&\quad \le \lceil \delta^{-1} T \rceil^2 n^{\lceil \delta^{-1} T \rceil} (2\delta J K)^{n-1} C_{\lceil \delta^{-1} T \rceil \delta},
\end{aligned}$$

for all $n \in \mathbb{N}$ and $k \in [K]$, implying that

$$\begin{aligned}
&\|y_k^n - y_k^{n-1}\|_{T,K} \\
&\quad \le \lceil \delta^{-1} T \rceil^2 n^{\lceil \delta^{-1} T \rceil} (2\delta J K)^{n-1} C_{\lceil \delta^{-1} T \rceil \delta}.
\end{aligned} \tag{76}$$

Thus, to prove that the sequence $\{(y_1^n|_{[0,T]}, \dots, y_K^n|_{[0,T]}) : n = 0, 1, \dots\}$ defined by (71) is Cauchy in D_T^K, it suffices to show that the right-hand side of (76) converges to zero as $n \to \infty$. But, by our choice of δ, note that

$$\begin{aligned}
&\limsup_{n\to\infty} \left| \frac{\lceil \delta^{-1} T \rceil^2 (n+1)^{\lceil \delta^{-1} T \rceil} (2\delta J K)^n C_{\lceil \delta^{-1} T \rceil \delta}}{\lceil \delta^{-1} T \rceil^2 n^{\lceil \delta^{-1} T \rceil} (2\delta J K)^{n-1} C_{\lceil \delta^{-1} T \rceil \delta}} \right| \\
&\quad = \limsup_{n\to\infty} \Big[\frac{(n+1)^{\lceil \delta^{-1} T \rceil}\, 2\delta J K}{n^{\lceil \delta^{-1} T \rceil}} \Big] = 2\delta J K < 1,
\end{aligned}$$

implying that $\sum_{n=1}^{\infty} \lceil \delta^{-1} T \rceil^2 n^{\lceil \delta^{-1} T \rceil} (2\delta J K)^{n-1} C_T < \infty$ by the ratio test. Hence, the right-hand side of (76) converges to zero as $n \to \infty$, completing the proof. □

A.2. Uniqueness of the element in D_K^T satisfying (56) for all $t \in [0, T]$

We show that $(y_{1,T}^\infty, \dots, y_{K,T}^\infty) \in D_T^K$ given by (72) is the unique element in D_T^K satisfying (56) for all $t \in [0, T]$. Suppose that $(y_1, \dots, y_K), (\tilde{y}_1, \dots, \tilde{y}_K) \in D_T^K$ both satisfy (56) for all $t \in [0, T]$. We partition the interval $[0, T]$ into a finite number of subintervals and show that both solutions agree on each of these subintervals. To that end, define

$$m := \inf \big\{ n \ge 1 : \frac{n}{2} (2JK\eta)^{-1} > T \big\}.$$

First, letting $t_1 := \frac{1}{2} (2JK\eta)^{-1}$, we have that

$$\begin{aligned}
&\max_{k\in[K]} \|y_k - \tilde{y}_k\|_{t_1} \\
&\quad \le \max_{k\in[K]} \eta \int_0^{t_1} \|y_k - \tilde{y}_k\|_{t_1}\, ds \\
&\qquad\qquad + \sum_{j=1}^{J} \sum_{l=1}^{K} \eta \int_0^{t_1} \|y_l - \tilde{y}_l\|_{t_1}\, ds
\end{aligned}$$

$$\begin{aligned} &\leq 2JK\eta t_1 \max_{k\in[K]} \|y_k - \tilde{y}_k\|_{t_1} \\ &= \frac{1}{2} \max_{k\in[K]} \|y_k - \tilde{y}_k\|_{t_1}. \end{aligned}$$

Hence, it follows that $\max_{k\in[K]} \|y_k - \tilde{y}_k\|_{t_1} = 0$, implying that $y \equiv \tilde{y}$ on $[0, t_1]$. Second, letting $t_2 := (2JK\eta)^{-1}$, we have that

$$\begin{aligned} &\max_{k\in[K]} \|y_k - \tilde{y}_k\|_{t_2} \\ &\quad = \max_{k\in[K]} \eta \Big\| \int_0^{t_1} |y_k(s) - \tilde{y}_k(s)|\, ds \\ &\qquad\qquad + \int_{t_1}^{\cdot} |y_k(s) - \tilde{y}_k(s)|\, ds \Big\|_{t_2} \\ &\qquad + \eta J \sum_{l=1}^{K} \Big\| \int_0^{t_1} |y_l(s) - \tilde{y}_l(s)|\, ds \\ &\qquad\qquad + \int_{t_1}^{\cdot} |y_l(s) - \tilde{y}_l(s)|\, ds \Big\|_{t_2} \\ &\quad \leq \max_{k\in[K]} \left\{ \eta t_1 \|y_k - \tilde{y}_k\|_{t_1} + (t_2 - t_1)\|y_k - \tilde{y}_k\|_{t_2} \right\} \\ &\qquad + \eta J \sum_{l=1}^{K} \left[t_1 \|y_l - \tilde{y}_l\|_{t_1} + (t_2 - t_1)\|y_l - \tilde{y}_l\|_{t_2} \right] \\ &\quad \leq 2\eta JK(t_2 - t_1) \max_{k\in[K]} \|y_k - \tilde{y}_k\|_{t_2} \\ &\quad = \frac{1}{2} \max_{k\in[K]} \|y_k - \tilde{y}_k\|_{t_2}, \end{aligned}$$

where the third inequality follows from the fact that $y \equiv \tilde{y}$ on $[0, t_1]$. Hence, it follows that $\|y_k - \tilde{y}_k\|_{t_2} = 0$, implying that $y \equiv \tilde{y}$ on $[0, t_2]$.

Continuing in an iterative fashion, the same argument shows that $y \equiv \tilde{y}$ on $[0, t_n]$ for all $n \in [m-1]$, where $t_n := \frac{n}{2}(2JK\eta)^{-1}$ for $n \in [m-1]$. If $t_{m-1} = T$, then we are done. However, if $t_{m-1} < T$, then we can let $t_m := T$ and use the same argument to show that $y \equiv \tilde{y}$ on $[0, t_m]$.

A.3. Extension to a unique element in D^K satisfying (56) for all $t \in [0, \infty)$

The previous two subsections have shown that, for each $T > 0$, there exists a unique $(y^\infty_{1,T}, \dots, y^\infty_{K,T}) \in D^K_T$ satisfying (56) for all $t \in [0, T]$. Using these solutions, we construct an element in D^K that uniquely satisfies (56) for all $t \in [0, \infty)$. To that end, define $(y^\infty_1, \dots, y^\infty_K) \in D^K$ by

$$\big(y^\infty_1, \dots, y^\infty_K\big)(t) := \big(y^\infty_{1,T}, \dots, y^\infty_{K,T}\big)(t), \quad t \in [0, T].$$

for each $T > 0$. We claim that $(y^\infty_1, \dots, y^\infty_K)$ is well-defined and is the unique element in D^K satisfying (56) for all $t \in [0, \infty)$. To prove that it is well-defined, we must show that whenever $t \in [0, T_1] \cap [0, T_2]$, we have

$$\big(y^\infty_{1,T_1}, \dots, y^\infty_{K,T_1}\big)(t) = \big(y^\infty_{1,T_2}, \dots, y^\infty_{K,T_2}\big)(t). \tag{77}$$

Without loss of generality, suppose that $T_1 \leq T_2$. Then $(y^\infty_{1,T_2}, \dots, y^\infty_{K,T_2})|_{[0,T_1]} \in D^K_{T_1}$ satisfies (56) for all $t \leq T_1$. By uniqueness,

$$\big(y^\infty_{1,T_1}, \dots, y^\infty_{K,T_1}\big) = \big(y^\infty_{1,T_2}, \dots, y^\infty_{K,T_2}\big)\Big|_{[0,T_1]},$$

which implies that

$$\begin{aligned} \big(y^\infty_{1,T_1}, \dots, y^\infty_{K,T_1}\big)(t) &= \big(y^\infty_{1,T_2}, \dots, y^\infty_{K,T_2}\big)\Big|_{[0,T_1]}(t) \\ &= \big(y^\infty_{1,T_2}, \dots, y^\infty_{K,T_2}\big)(t), \end{aligned}$$

for all $t \in [0, T_1]$, proving (77) holds. Finally, by the construction, it is obvious that $(y^\infty_1, \dots, y^\infty_K)$ uniquely satisfies (56) for all $t \in [0, \infty)$. □

B. Proof of Proposition 1

Fix $(\xi, \zeta) \in D^{J+K}$ satisfying (49)–(50). We first prove existence. By Lemma 1, there exists a $y \in D^K$ satisfying (56). Then, for $j \in [J]$, define

$$u_j(t) := \mu_j^{-1} \psi\Big(\xi_j + \sum_{l=1}^{K} q_{lj}\eta_l \int_0^t y_l(s)\, ds\Big), \quad t \geq 0, \tag{78}$$

$$x_j(t) := \phi\Big(\xi_j + \sum_{l=1}^{K} q_{lj}\eta_l \int_0^t y_l(s)\, ds\Big), \qquad t \geq 0. \tag{79}$$

Since $y \in D^K$, it follows that $u \in D^J$ and $x \in D^J$, so that $(x, u, y) \in D^{2J+K}$. To prove existence, it suffices to show that (x, u, y) satisfies (51)–(55). Equation (51) follows from (1)–(2) and (78)–(79). Equation (52) follows from (56)–(78). Equation (53) follows from (49), (51)–(52), and the fact that the matrices in (3) are stochastic. Equation (54) follows from (1), (49), and (78). Finally, for $j \in [J]$, define

$$z_j(t) := \xi_j + \sum_{l=1}^{K} q_{lj}\eta_l \int_0^t y_l(s)\, ds, \quad t \geq 0. \tag{80}$$

Thus, $u_j = \mu_j^{-1}\psi(z_j)$ and $x_j = \phi(z_j)$. But, by (1)–(2), it follows that $\int_0^\infty \mathbb{1}_{\{x_j>0\}}\, d(u_j)(t) = 0$. Therefore, (55) holds.

We now prove uniqueness. Suppose $(x, u, y), (\tilde{x}, \tilde{u}, \tilde{y}) \in D^{2J+K}$ both satisfy (51)–(55). Then, by (51), we have that

$$x_j = z_j + \mu_j u_j \geq 0, \quad \tilde{x}_j = \tilde{z}_j + \mu_j \tilde{u}_j \geq 0, \tag{81}$$

for $j \in [J]$, where z_j and $\tilde{z}_j$ are given by (80), with y_j and $\tilde{y}_j$, respectively. Since (x, u) and $(\tilde{x}, \tilde{u})$ both satisfy (54)–(55), it follows that $(x, \mu_j u_j)$ and $(\tilde{x}, \mu_j \tilde{u}_j)$ also both satisfy (54)–(55). By this and (81), for $j \in [J]$ we have that

$$\mu_j u_j = \psi(z_j) \quad \text{and} \quad \mu_j \tilde{u}_j = \psi(\tilde{z}_j). \tag{82}$$

It then follows from (52), (80), (82), and Lemma 1 that $y_k = \tilde{y}_k$ for all $k \in [K]$. By uniqueness of y, it follows from (80) and (82) that $u_j = \tilde{u}_j$ for all $j \in [J]$. Finally, by uniqueness of y and u, it follows from (80)–(81) that $x_j = \tilde{x}_j$ for all $j \in [J]$. This completes the proof.

C. Proof of Proposition 2

It is easy to show that the function $f = (f_1, f_2, f_3)$ is continuous if the functions f_1, f_2, and f_3 are continuous. On the other hand, it is straightforward to prove continuity of f_1 and f_2 once the continuity of f_3 is established. Thus, we only prove continuity of f_3 and omit the continuity proofs of f_1, f_2, and f for the purposes of brevity.

In the Skorokhod $\boldsymbol{J}_1$ topology (see, e.g., Billingsley [6] and Whitt [25]), a sequence $\{x^n\}_{n=1}^{\infty}$ in $\boldsymbol{D}^k$ converges to an element $x \in \boldsymbol{D}^k$, written $x^n \to x$, as $n \to \infty$, if $d_T^k(x^n|_{[0,T]}, x|_{[0,T]}) \to 0$ as $n \to \infty$ for all continuity points $T > 0$ of x, where $d_T^k : \boldsymbol{D}_T^k \times \boldsymbol{D}_T^k \to [0, \infty)$ is given by

$$d_T^k(x, y) := \inf_{\lambda \in \Lambda_T} \big\{ \|x \circ \lambda - y\|_{T,k} \vee \|\lambda - e\|_T \big\},$$

where $e : [0, T] \to [0, T]$ is the identity map, and

$$\Lambda_T := \big\{ \lambda : [0, T] \to [0, T] \\ \mid \lambda \text{ is an increasing homeomorphism} \big\}.$$

To that end, suppose that $(\xi^n, \zeta^n) \to (\xi, \zeta)$ in $\boldsymbol{D}^{J+K}$ as $n \to \infty$ and let $\tilde{T} > 0$ be a continuity point of $f_3(\xi, \zeta)$. To prove that f_3 is continuous with respect to the Skorokhod $\boldsymbol{J}_1$ topology, we must show that

$$\lim_{n\to\infty} d_{\tilde{T}}^K \big(f_3(\xi^n, \zeta^n)|_{[0,\tilde{T}]}, f_3(\xi, \zeta)|_{[0,\tilde{T}]} \big) = 0.$$

Since $(\xi, \zeta) \in \boldsymbol{D}^{J+K}$, it has at most countably many points of discontinuity; see, e.g., Ethier and Kurtz [12, Lemma 5.1]. It follows that there exists some $T > \tilde{T}$ that is a continuity point of (ξ, ζ). Therefore, by Billingsley [6, Lemma 1, page 167], it suffices to show that

$$\lim_{n\to\infty} d_T^K \big(f_3(\xi^n, \zeta^n)|_{[0,T]}, f_3(\xi, \zeta)|_{[0,T]} \big) = 0. \tag{83}$$

However, we note that (83) can equivalently be written with $\tilde{d}_T^k$ in place of d_T^k (see, e.g., page 226 of Pang et al. [23] and Billingsley [6]), where $\tilde{d}_T^k : \boldsymbol{D}_T^k \times \boldsymbol{D}_T^k \to [0, \infty)$ is given by

$$\tilde{d}_T^k(x, y) := \inf_{\lambda \in \tilde{\Lambda}_T} \big\{ \|x \circ \lambda - y\|_{T,k} \vee \|\dot{\lambda} - 1\|_T \big\},$$

where $\dot{\lambda}$ is the derivative of λ, 1 is the constant function taking the value one everywhere, and

$$\tilde{\Lambda}_T := \big\{ \lambda \in \Lambda_T : \lambda \text{ is absolutely continuous} \\ \text{w.r.t. Lebesgue measure} \big\}.$$

Therefore, the remainder of the proof aims at proving (83) with $\tilde{d}_T^K$ in place of d_T^K. To avoid overly cumbersome notation, we write $f_3(\xi^n, \zeta^n)$ and $f_3(\xi, \zeta)$ to mean $f_3(\xi^n, \zeta^n)|_{[0,T]}$ and $f_3(\xi, \zeta)|_{[0,T]}$, respectively.

Since $T > 0$ is a continuity point of (ξ, ζ) and $(\xi^n, \zeta^n) \to (\xi, \zeta)$ in $\boldsymbol{D}^{J+K}$ as $n \to \infty$, there exists a sequence of homeomorphisms $\lambda^n \in \tilde{\Lambda}_T$ such that

$$\|(\xi, \zeta) \circ \lambda^n - (\xi^n, \zeta^n)\|_{T,J+K} \vee \|\dot{\lambda}^n - 1\|_T \to 0 \tag{84}$$

as $n \to \infty$. Then, letting $y := f_3(\xi, \zeta)$ and $y^n := f_3(\xi^n, \zeta^n)$, it follows from the definition of f_3 and the triangle inequality that for any $0 \leq t \leq T$,

$$\begin{aligned}
&\max_{k\in[K]} \|y_k \circ \lambda^n - y_k^n\|_t \\
&\quad \leq \max_{k\in[K]} \Bigg[\|\zeta_k \circ \lambda^n - \zeta_k^n\|_t \\
&\qquad + \eta \Bigg\| \int_0^{\lambda^n(\cdot)} y_k(s)\,ds - \int_0^{\cdot} y_k^n(s)\,ds \Bigg\|_t \\
&\qquad + \sum_{j=1}^{J} \Bigg\| \psi\Bigg(\xi_j + \sum_{l=1}^{K} q_{lj}\eta_j \int_0^{\cdot} y_l(s)\,ds \Bigg) \circ \lambda^n \\
&\qquad\qquad - \psi\Bigg(\xi_j^n + \sum_{l=1}^{K} q_{lj}\eta_l \int_0^{\cdot} y_l^n(s)\,ds \Bigg) \Bigg\|_t \Bigg].
\end{aligned} \tag{85}$$

We next bound each term on the right-hand side of (85). First, let $\boldsymbol{M}_T := \max_{k\in[K]} \|y_k\|_T < \infty$. Then, by the chain rule, it follows that

$$\begin{aligned}
&\Bigg\| \int_0^{\lambda^n(\cdot)} y_k(s)\,ds - \int_0^{\cdot} y_k^n(s)\,ds \Bigg\|_t \\
&\quad = \Bigg\| \int_0^{\cdot} y_k(\lambda^n(s))\,\dot{\lambda}^n(s)\,ds - \int_0^{\cdot} y_k^n(s)\,ds \Bigg\|_t \\
&\quad \leq \Bigg\| \int_0^{\cdot} y_k(\lambda^n(s))\big(\dot{\lambda}^n(s) - 1\big)\,ds \Bigg\|_t \\
&\qquad + \Bigg\| \int_0^{\cdot} \big(y_k(\lambda^n(s)) - y_k^n(s) \big)\,ds \Bigg\|_t \\
&\quad \leq T\boldsymbol{M}_T \|\dot{\lambda} - 1\|_T + \int_0^t \|y_k \circ \lambda^n - y_k^n\|_s\,ds.
\end{aligned} \tag{86}$$

Similarly, by the chain rule, it follows that

$$\begin{aligned}
&\Bigg\| \psi\Bigg(\xi_j + \sum_{l=1}^{K} q_{lj}\eta_j \int_0^{\cdot} y_l(s)\,ds \Bigg) \circ \lambda^n \\
&\qquad - \psi\Bigg(\xi_j^n + \sum_{l=1}^{K} q_{lj}\eta_l \int_0^{\cdot} y_l^n(s)\,ds \Bigg) \Bigg\|_t \\
&\quad \leq \|\xi_j \circ \lambda^n - \xi_j^n\|_T \\
&\qquad + \eta \sum_{l=1}^{K} \Bigg\| \int_0^{\lambda^n(\cdot)} y_l(s)\,ds - \int_0^{\cdot} y_l^n(s)\,ds \Bigg\|_t \\
&\quad \leq \|\xi_j \circ \lambda^n - \xi_j^n\|_T \\
&\qquad + \eta \sum_{l=1}^{K} \Bigg\| \int_0^{\cdot} y_l(\lambda^n(s))\,\dot{\lambda}^n(s)\,ds - \int_0^{\cdot} y_l^n(s)\,ds \Bigg\|_t \\
&\quad \leq \|\xi_j \circ \lambda^n - \xi_j^n\|_T + \eta K T \boldsymbol{M}_T \|\dot{\lambda}^n - 1\|_T \\
&\qquad + \eta \sum_{l=1}^{K} \int_0^t \|y_l \circ \lambda^n - y_l^n\|_s\,ds,
\end{aligned} \tag{87}$$

where the first inequality holds by Whitt [25, Lemma 13.5.1] and the fact that

$$\psi\Bigg(\xi_j + \sum_{l=1}^{K} q_{lj}\eta_j \int_0^{\cdot} y_l(s)\,ds \Bigg) \circ \lambda^n$$

$$= \psi\Big(\xi_j \circ \lambda^n + \sum_{l=1}^{K} q_{lj}\eta_j \int_0^{\lambda^n(\cdot)} y_l(s)\,ds\Big);$$

see, e.g., Whitt [25, Lemma 13.5.2]. Combining these estimates, it follows from (85)–(87) that

$$\begin{aligned}
&\max_{k\in[K]} \|y_k \circ \lambda^n - y_k^n\|_t \\
&\quad\le 2J\,\|(\xi,\zeta)\circ\lambda^n - (\xi^n,\zeta^n)\|_{T,J+K} \\
&\qquad + (JK+1)\,\eta T M_T \|\dot{\lambda}^n - 1\|_T \\
&\qquad + 2\eta JK \int_0^t \max_{k\in[K]} \|y_k\circ\lambda^n - y_k^n\|_s\,ds \qquad (88)
\end{aligned}$$

for all $0 \le t \le T$. Fixing $\epsilon > 0$, it follows from (84) that there exists an $N \in \mathbb{N}$ such that for all $n \ge N$,

$$2J\,\|(\xi,\zeta)\circ\lambda^n - (\xi^n,\zeta^n)\|_{T,J+K} < \frac{\epsilon}{2e^{2\eta JKT}}, \qquad (89)$$

$$(JK+1)\,\eta T M_T \|\dot{\lambda}^n - 1\|_T < \frac{\epsilon}{2e^{2\eta JKT}}, \qquad (90)$$

$$\|\dot{\lambda}^n - 1\|_T < \epsilon. \qquad (91)$$

Then, by (88)–(90), it follows that for all $n \ge N$,

$$\begin{aligned}
&\max_{k\in[K]} \|y_k \circ \lambda^n - y_k^n\|_t \\
&\quad< \frac{\epsilon}{e^{2\eta JKT}} + 2\eta JK \int_0^t \max_{k\in[K]} \|y_k\circ\lambda^n - y_k^n\|_s\,ds
\end{aligned}$$

for all $0 \le t \le T$. By Gronwall's inequality (see, e.g., Pang et al. [23, Lemma 4.1]) and the above displayed inequality, it follows that

$$\max_{k\in[K]} \|y_k \circ \lambda^n - y_k^n\|_t < e^{2\eta JK(t-T)}\epsilon \le \epsilon \qquad (92)$$

for all $0 \le t \le T$. Therefore, by (91)–(92), we have that $\|y\circ\lambda^n - y^n\|_{T,K} \vee \|\dot{\lambda}^n - 1\|_T < \epsilon$ for all $n \ge N$. This completes the proof.

D. Miscellaneous Proofs

Below is the proof of a result used in the proof of Lemma 2 in Section 5.

Lemma 6. *If $\{X^n\}_{n=1}^{\infty}$ is a random sequence in $\mathbf{D}$ such that $\|X^n\|_T \Rightarrow 0$ as $n \to \infty$ for all $T > 0$, then $X^n \Rightarrow \mathbf{0}$ as $n \to \infty$.*

Proof. Note that $\|X^n\|_T \Rightarrow 0$ as $n \to \infty$ for all $T > 0$ is equivalent to $\|X^n\|_T \xrightarrow{p} 0$ as $n \to \infty$ for all $T > 0$, where the notation $\xrightarrow{p}$ is shorthand for "converges in probability." We next show that $X^n \xrightarrow{p} \mathbf{0}$ as $n \to \infty$. This amounts to showing that for all $0 < \epsilon < 1$,

$$\lim_{n\to\infty} P\Big(\int_0^\infty e^{-t}\,\big[d_t(X^n, 0)\wedge 1\big]\,dt > \epsilon\Big) = 0;$$

see, e.g., Whitt [25, Chapter 3, Section 3]. But observe that

$$\begin{aligned}
&P\Big(\int_0^\infty e^{-t}\big[\inf_{\lambda\in\Lambda_t}\{\|X^n\circ\lambda\|_t \vee \|\lambda - e\|_t\}\wedge 1\big]\,dt > \epsilon\Big) \\
&\le P\Big(\int_0^\infty e^{-t}\,\big[\|X^n\|_t \wedge 1\big]\,dt > \epsilon\Big) \\
&\le P\Big(\int_0^T e^{-t}\|X^n\|_t\,dt + \int_T^\infty e^{-t}\,dt > \epsilon\Big) \qquad (93)
\end{aligned}$$

for all $T > 0$. Now fix $T > 0$ to be such that $\int_T^\infty e^{-t}\,dt = \epsilon/2$. It then follows from (93) that

$$\begin{aligned}
&P\Big(\int_0^T e^{-t}\|X^n\|_t\,dt + \int_T^\infty e^{-t}\,dt > \epsilon\Big) \\
&\quad= P\Big(\int_0^T e^{-t}\|X^n\|_t\,dt > \frac{\epsilon}{2}\Big) \\
&\quad\le P\Big(\|X^n\|_T > \frac{\epsilon}{2}\big(1-\frac{\epsilon}{2}\big)^{-1}\Big). \qquad (94)
\end{aligned}$$

Since $\|X^n\|_T \xrightarrow{p} 0$ as $n \to \infty$, it follows from (94) that $X^n \xrightarrow{p} \mathbf{0}$ as $n \to \infty$. Since convergence in probability implies convergence in distribution, it follows that $X^n \Rightarrow \mathbf{0}$ as $n \to \infty$, which completes the proof. □