

Multi-scale diffusion approximations for stochastic networks in random environments under heavy-traffic

Ankita Sen^{1,2} and N. Selvaraju^{2*}

¹Department of Data and Decision Sciences, Technion – Israel Institute of Technology, Haifa 3200003, Israel
E-mail address: atikna.math@gmail.com

²Department of Mathematics, Indian Institute of Technology Guwahati, Guwahati 781039, India
E-mail address: nselvaraju@iitg.ac.in
URL: <https://fac.iitg.ac.in/nselvaraju/>

Abstract. This paper studies a generalized Jackson network (GJN) with time-varying arrival, service, and abandonment mechanisms, which are governed by an independently evolving external random environment. The GJN is also equipped with time-dependent staffing and state dependent routing mechanism. The additional involvement of the external process to the main queue dynamics leads to a new many-server heavy-traffic regime under time-varying framework. By introducing a scaling parameter β (> 0), the transition speed of the external process is accelerated by N^β factor inside the heavy-traffic regime (fast switching regime). We establish the fluid (or deterministic) approximation and diffusion approximation inside the new time-varying many-server heavy-traffic asymptotic regime under suitable assumptions. Due to the fluctuation of fluid approximation with respect to time-varying staffing, the system enters into three possible states, which are referred to as *subcritical*, *critical*, or *supercritical*. In each of the regimes, the diffusion limiting process is characterized as a multi-dimensional Ornstein-Uhlenbeck process with time-varying drift and diffusion parameters. Based on the value of β (i.e., whether $\beta < 1$, $\beta = 1$, $\beta > 1$), the diffusion approximation of the system length process leads to a trichotomy. Since the GJN dynamics is entirely influenced by the evolution of the external process, we derive further the Gaussian approximations of the external process individually within a general asymptotic framework (*central limit theorem regime*) adapting a suitable *averaging principle*.

1. Introduction

1.1. *Background and Motivation:* General stochastic networks, such as the Jackson queueing network, serve as foundational models for complex and large scale service systems comprising multiple

Received by the editors June 1st, 2024; accepted February 24th, 2026; published under license [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/).

2010 *Mathematics Subject Classification.* 60K25, 60K37, 60F17, 60J25, 60G44.

Key words and phrases. Multi-scale diffusion approximations, GJN network, time-varying rate functions, Markovian random environment, averaging principle, martingale problems, multi-dimensional Ornstein-Uhlenbeck process.

The second author (N. Selvaraju) was supported by Grant No. MTR/2021/000425 from Science and Engineering Research Board (SERB), Government of India.

*Corresponding Author, <https://orcid.org/0000-0002-7184-4193>

interconnected service stations, which are widely applicable to cloud computing and telecommunication infrastructures, transportation and logistics systems, etc. However, the exact analysis of general queueing networks is often intractable, especially in high volume or high utilization regimes. This necessitates asymptotic methods that approximate system behaviours when the network operates under heavy traffic—that is, when demand closely matches service capacity. Nevertheless, from both practical and theoretical perspectives, many of the above stochastic network systems often deal with time-varying service, staffing or routing situations, interaction amongst service stations, allowing an additional random effect inside the arrival, service, and routing mechanisms. A convenient and effective approach to address such challenging scenarios is to embed an external process within the primary stochastic system, allowing it to evolve on a potentially different time scale than the time-dependent dynamics governing arrival, service, staffing, and routing mechanisms. Motivated by this framework, in this paper, we study a stochastic network (the primary process) whose arrival, service, and abandonment mechanisms are modulated by an external Markov process (the underlying process). These two interacting processes evolve on two different time scales, enabling the model to effectively capture both stochastic fluctuations (for instance environmental uncertainty) and deterministic patterns (for instance time-of-day effects).

We consider a stochastic network with L service stations (nodes) with $\mathbb{L} = \{1, 2, \dots, L\}$, which is governed by a continuous-time Markov process $\{\mathcal{K}(t), t \geq 0\}$ with state space $\mathbb{S}' = \{1, 2, \dots, d\}$ and invariant distribution $\boldsymbol{\pi} = \{\pi_1, \pi_2, \dots, \pi_d\}$. In the present study, the external Markov process \mathcal{K} is considered to be a *random environment*, whose evolution strongly influence the time-varying arrival, service and abandonment mechanisms of the network, i.e., the rate functions of these mechanisms are defined as $f(t, \mathcal{K}(t))$, for some Borel-measurable function $f : \mathbb{R}_+ \times \mathbb{S}' \mapsto \mathbb{R}$. The dynamics of the routing probabilities $(\phi_{ij}(\cdot), \psi_{ij}(\cdot), i, j \in \mathbb{L})$ of the network is also governed by \mathcal{K} . Each node in the network is operated by a certain number of time-varying parallel servers $K(t)$, independent of \mathcal{K} activities. In the time-varying framework, staffing $K(t)$ defines the number of active time-varying servers in each node, which may decrease or increase with time while all servers are busy. The time-varying queueing network modulated by a random environment is often referred to as *generalized Jackson network (GJN)* (see Mandelbaum et al., 1998; Budhiraja and Liu, 2011). Within the framework of time-varying and randomly modulated environments, it is both natural and essential to study the long-run behaviour of such complex stochastic network systems, particularly when each node operates with a large number of servers. Our aim is to derive a tractable and computationally feasible approximation of the system length process $\{\mathbf{Q}(t) = (Q_i(t))_{i=1}^L, t \geq 0\}$ for the network, by employing heavy-traffic theory under an appropriate asymptotic regime.

1.2. Main Results and Contributions: The present study addresses several fundamental theoretical and methodological challenges in analyzing queueing networks subject to both time-varying parameters and external random modulation. The incorporation of an external stochastic process into a time-dependent framework introduces both predictable and random sources of variability, rendering classical many-server heavy-traffic regimes inapplicable. Notably, the presence of external modulation breaks the stationarity of the system and invalidates key independence assumptions typically imposed on arrival, service, and abandonment processes. Furthermore, within a time-varying framework, the system may transition between subcritical (underloaded), critically loaded, and supercritical (overloaded) phases within a single sample path, causing the offered load to vary in a non-trivial and non-stationary manner—thus complicating modelling and performance approximations. These difficulties are further worsened when the external environment evolves on a different time scale than queueing dynamics, where standard techniques fail to yield uniform approximations or accurate heavy-traffic limits. In this study, we address the aforementioned challenges by developing and applying advanced methodological tools tailored to two-time scale many-server systems

operating in time-varying environments. These techniques not only advance the theoretical understanding but also offer broad applicability to a wide range of two-time scale stochastic dynamic systems. The main contributions of the paper are outlined below.

- (1) We propose a novel *two-time-scale time-varying many-server heavy-traffic regime* to study heavy-traffic theory, offering comprehensive insights into the limiting diffusion structure of stochastic systems evolving on multiple time scales.
- (2) A key innovation is the introduction of a transition scaling parameter $\beta > 0$ in the proposed time-varying many-server heavy-traffic regime resulting in two-time scale separation, which governs the speed of the random environment relative to the primary queueing system, enabling a unified treatment of multiple time-scale interactions.
- (3) The proposed regime not only encompasses the classical Halfin-Whitt heavy-traffic regime (critically loaded) under a constant framework, but accommodates both subcritical and supercritical phases within the same framework, depending on the time-varying offered load.
- (4) We propose an advanced and refined methodology based on the *non-homogeneous stochastic averaging principle*, applicable to a broad class of interacting and coupled non-homogeneous stochastic dynamical systems.
- (5) The multi-scale diffusion approximations (Theorems 3.2 and 3.4) presented in this article offer significant insights and a tractable representation of the asymptotic behavior of the system length process in generalized Jackson networks (GJN) across all possible dynamic loading regimes. These results substantially extend the classical diffusion theory for time-homogeneous stochastic dynamical systems to a more general setting involving two-time-scale interactions.
- (6) The unified diffusion framework for multi-scale modulated systems exhibits a *trichotomous structure* in the limiting diffusion process, depending on whether $\beta < 1$, $\beta = 1$ or $\beta > 1$. This general characterization yields a comprehensive approximation applicable to a wide range of real-world systems influenced by both stochastic fluctuations and deterministic variability.

Time-varying many-server heavy-traffic regime at two-time scale framework: In the present work, to derive diffusion approximations for the generalized Jackson network (GJN), we consider a sequence of such systems indexed by a scaling parameter $N \in \mathbb{N}$. Under the proposed two-time-scale, time-varying, many-server heavy-traffic regime, the external Markov process $\{\mathcal{K}^N(t), t \geq 0\}$ transitions among states at a rate proportional to $\mathcal{O}(N^\beta)$ for some $\beta > 0$. In the primary system, while the arrival rates and staffing levels at each node are increased at order $\mathcal{O}(N)$, the service and abandonment rates remain of order $\mathcal{O}(1)$. This results in a two-time-scale structure within the queueing dynamics. Specifically, the range of the transition parameter β with $\beta = 1$, or $\beta > 1$, or $\beta < 1$ determines whether the primary GJN system evolves at the same speed as \mathcal{K}^N , significantly slower, or faster, respectively. We propose that if the arrival and staffing functions are defined by $f^N(t, \mathcal{K}^N(t)) \approx Nf(t, \mathcal{K}^N(t)) + N^\delta \hat{f}(t, \mathcal{K}^N(t))$ with $\delta = \max\{1 - \beta/2, 1/2\}$, for some locally integrable function $f(\cdot, \cdot), \hat{f}(\cdot, \cdot)$, then the system may experience subcritical, critical, and supercritical loading phases along its sample paths—depending on the fluid characteristics of the arrival and staffing functions (see Section 3). To ensure stability across all loading phases, it is essential that the centered and scaled traffic load, $N^{1-\delta}(\rho^N(t) - \rho_0(t))$, determined by the system traffic load $\rho^N(t)$ (i.e., the ratio of arrival rate to total service capacity) and its fluid-scale counterpart $\rho_0(t)$ converges to a finite limit as $N \rightarrow \infty$.

Averaged Halfin-Whitt regime: By the proposed heavy-traffic scaling framework, it is noteworthy to mention that this novel, advanced and largely generalized many-server heavy-traffic regime, which is significantly different from the conventional heavy-traffic regimes, complements the well existing regimes including the Halfin-Whitt (QED) regime (Halfin and Whitt, 1981; Pang et al., 2007), and regimes involving time-varying dynamics (Mandelbaum et al., 1998; Liu and Whitt, 2012; Pang and Yao, 2013; Puhalskii, 2013; Whitt, 2018). Within this framework, the system maintains

the stability of the entire GJN system at every state of the random environment. Furthermore, it is capable of maintaining a perfect balance between high utilization and quality of service no matter what loading phase the system is experiencing. In addition, within a constant framework, when the traffic load (or staffing) of the system does not change over time, i.e., the system consistently operates at a particular loading regime (subcritical, critical or supercritical), our proposed regime has a significant advantage to operate the system into a phase where it is not required to be critically loaded at each stage of the external environment, instead, at each stage, the system has the relaxation behaving from subcritical to the supercritical regimes. As a result, we derive a further generalized heavy-traffic regime, referred to as *averaged Halfin-Whitt regime*, with respect to the invariant measure π where the system traffic load $\rho^N(t)$ converges to its critical value 1 as $N \rightarrow \infty$ but leaves a residual fluctuation of order $\mathcal{O}(N^{1-\delta})$ (see Remark 3.1). This framework extends the classical Halfin–Whitt regime by accommodating a broader range of the diffusion parameter, with $\delta = \max\{1 - \beta/2, 1/2\}$ as opposed to the fixed value $\delta = 1/2$. Consequently, it enables the system to simultaneously achieve high efficiency and good service quality across a wider spectrum of operational conditions.

Non-homogeneous stochastic averaging principle: To develop the diffusion theory of stochastic dynamical system operating under two different time scale framework, it is essential to characterize the asymptotic behaviour of the external random process within interactive environment. In this work, we present a generalized framework—extending beyond the scope of the current model, to study such dynamical system by employing the *non-homogeneous averaging principle* (see Papanicolaou et al., 1977; Skorokhod, 1989; Kurtz, 1992). This approach leads to Gaussian approximations that yield significant insights into the asymptotic behaviour of the coupled system.

Refined martingale methodology for dependent dynamics: Classical martingale approach (see Pang et al., 2007 and Puhalskii, 2013) in establishing the stochastic approximations, generally applicable for stationary queues or system without any effect of external random dynamics, break down in our present heavy-traffic analysis. It leads to the necessity of developing a new advanced methodology that can be applicable to generalized stochastic dynamical systems within a two-time scale framework. This methodology includes a two-part analysis: firstly, we characterize the random rate processes $\{f^N(t, \mathcal{K}^N(t)), t \geq 0\}$ interacted with the external random environment as coupled process and develop the Gaussian approximations applying two-time scale analogy (Section 5). Secondly, we construct an extended martingale representation for the system length process and tackle the dependency among the arrival, service, and abandonment processes (Sections 7, 8).

Trichotomy nature in diffusion limiting processes: This article demonstrates that, despite the analytical complexity introduced by large staffing levels and the two-time-scale dynamics of the GJN, the proposed stochastic approximations yield a tractable, analytically manageable, and computationally efficient representation. In fact, the fluid scaled system length process in virtue of Theorem 3.2 converges deterministically to a unique solution of an ordinary differential equation, which is built on the averaged parameters with respect to π , independent of the effect of transition rate parameter β . Furthermore, depending on the fluctuation of the fluid characterization resulting in any loading phases (subcritical, critical, or supercritical), it is proved in Theorem 3.4 that the centered and normalized system length process converges in distribution to a trichotomized multi-dimensional Itô diffusion process depending on whether $\beta < 1$, $\beta = 1$, or $\beta > 1$. More precisely, the diffusion limit process can be characterized as a *multi-dimensional Ornstein-Uhlenbeck process* with time-varying drift and diffusion coefficients, which satisfies the Markov property.

1.3. *Related research:* Various kinds of asymptotic approximations of the stochastic networks (GJN and Jackson network) have been studied earlier at a large scale, beginning with the notable work by Reiman (1984) establishing the diffusion approximations inside a suitable heavy-traffic regime. Introducing time-dependency within arrival and service (abandonment) mechanisms, the diffusion

approximations are well established by [Mandelbaum et al. \(1998\)](#), [Puhalskii \(2013\)](#) and [Sun and Whitt \(2018\)](#) under Halfin-Whitt regime. Later, [Pang and Yao \(2013\)](#) derived the heavy-traffic approximations of a multi-server queueing network when the system experiences critically loaded and overloaded. A recent survey by [Whitt \(2018\)](#) extensively discusses the recent developments of time-dependent stochastic systems in a nice manner. In this regard, we can say that the time-varying setup within our novel many-server heavy-traffic regime broadly generalizes the diffusion theory of the aforementioned works. On the other hand, a noteworthy diffusion theory work related to GJN operating in a random environment has been studied by [Budhiraja and Liu \(2011\)](#), where the asymptotic scaling regime is substantially different from our present work. Recently, [Arapostathis et al. \(2019\)](#) and [Arapostathis et al. \(2021\)](#) studied the diffusion control problem for Markov-modulated many-server system within the constant framework, which can be partially recovered from our diffusion limit theorems. Not only limited to these results, the diffusion approximations presented in this article yield the diffusion theory for infinite-server queueing system operating under a Markovian environment presented in [Anderson et al. \(2016\)](#), [Jansen et al. \(2019\)](#) and [Sen and Selvaraju \(2023\)](#).

The rest of the paper is organized as follows: Section 2 describes the GJN model together with the definition of the external random environment. The main Theorems 3.2 and 3.4 are presented in Section 3 with remarks. The required martingale representation of the vector system length process is derived in Section 4. Section 5 demonstrates the averaging principle to prove the diffusion convergence of the external Markov process. The main proofs including fluid approximation and diffusion approximation are established in Sections 7 and 8, respectively. In the end, we derive the necessary convergence results subsequently in Appendix A.

2. Model description and GJN system dynamics

We consider a sequence of generalized Jackson networks (GJN) with L service stations, which are denoted by $\mathbb{L} = \{1, 2, \dots, L\}$ indexed by N . For each $N \in \mathbb{N}$, each station i has $K_i^N(t)$ number of time-varying parallel non-idle servers at time t providing service on FCFS basis. Every stations has an infinite capacity buffer and has its own queue. A job may join a particular station i from outside or inside routing of the network, and after joining it may wait until to get service or abandon the i th station to join another station j , $j \in \mathbb{L}_i = \mathbb{L} \setminus \{i\}$. At the end of the service at station i , the job may either join the other station j or may leave the entire system. In this article, we study the GJN where the arrival, service and abandonment mechanisms are governed by time t as well as an independently evolving random process $\mathcal{K}^N = \{\mathcal{K}^N(t), t \geq 0\}$. The random process \mathcal{K}^N can be interpreted as an external random environment, which is a continuous-time Markov process with state space $(\mathbb{S}', \mathcal{B}_{\mathbb{S}'})$, where $\mathbb{S}' = \{1, 2, \dots, d\}$. For every $N \in \mathbb{N}$, the Markov process \mathcal{K}^N is characterized by the transition probability matrix $\mathcal{P}^N(t) = (\mathcal{P}_{kl}^N(t), k, l \in \mathbb{S}')$ and initial distribution $\boldsymbol{\pi} = \{\pi_k, k \in \mathbb{S}'\}$. Let $\mathcal{Q}^N = (\mathcal{Q}_{kl}^N, k, l \in \mathbb{S}')$ be the infinitesimal generator matrix associated with $\mathcal{P}^N(t)$ with non-diagonal entries $\mathcal{Q}_{kl}^N \geq 0$ and diagonal entries $\mathcal{Q}_k^N = -\sum_{l \neq k} \mathcal{Q}_{kl}^N < 0$. Assume that \mathcal{K}^N has a unique stationary distribution $\boldsymbol{\pi}$. In addition, the recurrent potential kernel $\mathcal{D}^N = (\mathcal{D}_{kl}^N, k, l \in \mathbb{S}')$ exists for the Markov process \mathcal{K}^N with $\mathcal{D}_{kl}^N = \int_0^\infty (\mathcal{P}_{kl}^N(t) - \pi_l) dt$. The dynamics of the scaled process \mathcal{K}^N has been described in detail in later part of this section.

In the N th queueing network, an exogenous job arrives at the i station with time-varying arrival rate $\lambda_{ki}^N(t)$ when the external random environment $\mathcal{K}^N(t)$ stays at the state k at a particular time t , $k \in \mathbb{S}'$. Upon arriving at station i , each job joins the queue in front of the station and may abandon the queue after an exponentially distributed time with time-varying parameter $\gamma_{ki}^N(t)$ when $\mathcal{K}^N(t)$ stays at the state k at time t , or it may wait for at the service i th station and requires an exponentially service time with the parameter $\mu_{ki}^N(t)$ whenever $\mathcal{K}^N(t)$ stays at k . Within time-varying staffing framework, the staffing count $K_i^N(\cdot)$, independent of \mathcal{K}^N dynamics, may decrease while the servers are busy, hence it is necessary to handle the assignment of the job being served

(the most recent customer joining for the service). Since, we prefer this job to be served by the rate function $\mu_{ki}(\cdot)$ for some $k \in \mathbb{S}'$ at same node i , we push the job into the head of queue and assign a complete new service once it reaches to server again (see Puhalskii, 2013). As the service time is assumed to be exponentially distributed, the remaining service time will have the same distribution as the new service time. As per the server to be allowed to take off, we assume that the server who completed its service most recently is allowed to be released. Let $\{E_i^N(t), t \geq 0\}$ be the exogenous cumulative arrival process, where $E_i^N(t)$ denotes the number of arrivals by time t at station i joining from outside. For every $N \in \mathbb{N}$, let $\{Q_i^N(t), t \geq 0\}$ be the system length process at the station i , where $Q_i^N(t)$ represents the number of jobs at this station at time t in the N th network, considering all the jobs in queue and service after routing from another service station $j (j \in \mathbb{L}_i)$. Inside the network, the routing mechanism is determined by the system state. More accurately, in the N th network, a particular job may join the station j in order to avoid significant waiting issue with probability $\psi_{ij}^N(Q_i^N(t-))$ after abandoning the station i or it may exit the system with $\psi_{i0}^N(Q_i^N(t-))$ probability at time t . Similarly, after completion of service at station i , a job may join the station j , ($j \in \mathbb{L}_i$), with probability $\phi_{ij}^N(Q_i^N(t-))$ or may leave the system with $\phi_{i0}^N(Q_i^N(t-))$ probability at time t . Therefore, we are given the $L \times L$ substochastic matrices $\boldsymbol{\psi}^N(\cdot) = (\psi_{ij}^N(\cdot), i, j \in \mathbb{L})$ and $\boldsymbol{\phi}^N(\cdot) = (\phi_{ij}^N(\cdot), i, j \in \mathbb{L})$ with diagonal entries $\psi_{ii}^N(\cdot) = \phi_{ii}^N(\cdot) = 0, i \in \mathbb{L}$. By definition, $\sum_{j=1}^L \psi_{ij}^N(\cdot) + \psi_{i0}^N(\cdot) = 1$ and $\sum_{j=1}^L \phi_{ij}^N(\cdot) + \phi_{i0}^N(\cdot) = 1$, for each $i \in \mathbb{L}$. In addition, for any $l \in \mathbb{N}$, we define the following cumulative processes associated with the nodes $i, j \in \mathbb{L}$ as follows:

- $Y_{ij}^{N,A,l}(t)$ represents the cumulative number of jobs who abandon the queue at station i and join the station j by time t .
- $Y_{ij}^{N,S,l}(t)$ represents the cumulative number of jobs routed to j upon the service completion at station i by time t .
- $Y_{i0}^{N,A,l}(t)$ represents the cumulative number of jobs who depart the entire system by time t at the end of abandonment from the station i .
- $Y_{i0}^{N,S,l}(t)$ represents the cumulative number of jobs exiting the entire system by time t after service at station i is completed.

We assume that all counting processes $E_i^N, Y_{ij}^{N,A,l}, Y_{ij}^{N,S,l}, i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\}$ and the external jump process \mathcal{K}^N have no common jumps \mathbb{P} -a.s, i.e., for every $N \in \mathbb{N}$,

$$\begin{aligned} \Delta E_{i_1}^N(t) \Delta E_{i_2}^N(t) &= 0, i_1, i_2 (\neq i_1) \in \mathbb{L}, \Delta E_{i_1}^N(t) \Delta Y_{i_2, j_2}^{N,I,l}, i_1, i_2 \in \mathbb{L}, j_2 \in \mathbb{L} \cup \{0\}, I = A, S, \\ \Delta Y_{i_1, j_1}^{N,A,l}(t) \Delta Y_{i_2, j_2}^{N,S,l}(t) &= 0, (i_1, j_1), (i_2, j_2) \in \mathbb{L} \times \mathbb{L} \cup \{0\}, \\ \Delta Y_{i_1, j_1}^{N,I,l}(t) \Delta Y_{i_2, j_2}^{N,I,l}(t) &= 0, (i_1, j_1), (i_2, j_2) (\neq (i_1, j_1)) \in \mathbb{L} \times \mathbb{L} \cup \{0\}, I = A, S, \\ \Delta E_i^N \Delta \mathcal{K}^N(t) &= 0, \Delta Y_{ij}^{N,I,l}(t) \Delta \mathcal{K}^N(t) = 0, i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\}, I = A, S. \end{aligned}$$

Therefore, the system dynamics at station i of the network can be represented as follows:

$$\begin{aligned} Q_i^N(t) &= Q_i^N(0) + E_i^N(t) + \sum_{j=1}^{\mathbb{L}_i} \sum_{l=1}^{\infty} \int_0^t \mathbf{1}(Q_j^N(s-) \geq K_j^N(s-) + l) \psi_{ji}^N(Q_j^N(s-)) dY_{ji}^{N,A,l}(s) \\ &\quad + \sum_{j=1}^{\mathbb{L}_i} \sum_{l=1}^{\infty} \int_0^t \mathbf{1}(Q_j^N(s-) \wedge K_j^N(s-) \geq l) \phi_{ji}^N(Q_j^N(s-)) dY_{ji}^{N,S,l}(s) \\ &\quad - \sum_{j=1}^{\mathbb{L}_i} \sum_{l=1}^{\infty} \int_0^t \mathbf{1}(Q_i^N(s-) \geq K_i^N(s-) + l) \psi_{ij}^N(Q_i^N(s-)) dY_{ij}^{N,A,l}(s) \end{aligned}$$

$$\begin{aligned}
 & - \sum_{j=1}^{\mathbb{L}_i} \sum_{l=1}^{\infty} \int_0^t \mathbf{1}(Q_i^N(s-) \wedge K_i^N(s-) \geq l) \phi_{ij}^N(Q_i^N(s-)) dY_{ij}^{N,S,l}(s) \\
 & - \sum_{l=1}^{\infty} \int_0^t \mathbf{1}(Q_i^N(s-) \geq K_i^N(s-) + l) \psi_{i0}^N(Q_i^N(s-)) dY_{i0}^{N,A,l}(s) \\
 & - \sum_{l=1}^{\infty} \int_0^t \mathbf{1}(Q_i^N(s-) \wedge K_i^N(s-) \geq l) \phi_{i0}^N(Q_i^N(s-)) dY_{i0}^{N,S,l}(s),
 \end{aligned} \tag{2.1}$$

where, for each $i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\}$, the counting processes $\{E_i^N(t), t \geq 0\}$, $\{Y_{ij}^{N,A,l}(t), t \geq 0\}$, $\{Y_{ij}^{N,S,l}(t), t \geq 0\}$ have trajectories in $\mathbb{D}([0, \infty), \mathbb{R})$, $l \in \mathbb{N}$. We introduce the L -dimensional random vectors $\mathbf{E}^N(t)$, $\mathbf{Y}_0^{N,A,l}(t)$ and $\mathbf{Y}_0^{N,S,l}(t)$ and $\mathbf{Q}^N(t)$ and the corresponding counting processes in $\mathbb{D}([0, \infty), \mathbb{R}^L)$, where each component $\mathbf{x} = (x_i)_{i=1}^L$ denotes the L -dimensional vector. Similarly, we represent the $L \times L$ -dimensional random matrices $\mathbf{Y}^{N,A,l}(t)$ and $\mathbf{Y}^{N,S,l}(t)$ and the corresponding random processes in $\mathbb{D}([0, \infty), \mathbb{M}^{L \times L})$, for each $l \in \mathbb{N}$, where $\mathbf{y} = (y_{ij})_{i,j=1}^L$ denotes the $L \times L$ -dimensional matrix.

In order to better understand the aforementioned counting processes, for every $N \in \mathbb{N}$, we define the d -dimensional indicator process $\{\mathbf{H}^N(t) = (H_k^N(t))_{k=1}^d, t \geq 0\}$ in $\mathbb{D}([0, \infty), \mathbb{R}^d)$, where $H_k^N(t) = 1$ if $\mathcal{K}^N(t) = k$, and zero otherwise. The definition of \mathbf{H}^N yields that \mathbf{H}^N is a jump Markov process with measurable state space $(\mathbb{S}, \mathcal{B}_{\mathbb{S}})$, where $\mathbb{S} = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_d\}$. The associated transition probability is defined by $\mathcal{P}^N(t, \mathbf{e}_k, \mathbf{e}_l) = \mathbb{P}(\mathbf{H}^N(t) = \mathbf{e}_l | \mathbf{H}^N(0) = \mathbf{e}_k)$, which is same as $\mathcal{P}_{kl}^N(t)$, $\mathbf{e}_k, \mathbf{e}_l \in \mathbb{S}$. The operator \mathcal{Q}^N generating the Markov process \mathbf{H}^N is given by $\mathcal{Q}^N g(\mathbf{h}) = \sum_{\mathbf{z} \in \mathbb{S}} \mathcal{Q}^N(\mathbf{h}, \mathbf{z})(g(\mathbf{z}) - g(\mathbf{h}))$, $\mathbf{h}, \mathbf{z} \in \mathbb{S}$, where g is an $\mathcal{B}_{\mathbb{S}}$ -measurable bounded function. It directly follows from the definition of \mathcal{K}^N that the new random process \mathbf{H}^N is stationary, and thus, for every $t \geq 0$, $\mathbf{H}^N(t)$ has the same distribution as $\mathbf{H}^N(0)$, which is distributed as $\boldsymbol{\pi} = \{\boldsymbol{\pi}(\mathbf{h}), \mathbf{h} \in \mathbb{S}\}$. It follows that $\boldsymbol{\pi} \mathcal{P}^N(t) = \boldsymbol{\pi}$, for any $t \geq 0$, and hence $\boldsymbol{\pi}$ to be the invariant distribution of \mathbf{H}^N , for all $N \in \mathbb{N}$.

With respect to the newly defined random environment \mathbf{H}^N , for $i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\}$, we define the counting processes mentioned in the system evolution (2.1) as follows:

$$\begin{aligned}
 E_i^N(t) &= P_i^E \left(\sum_{k=1}^d \int_0^t \lambda_{ki}^N(s) H_k^N(s) ds \right), Y_{ij}^{N,A,l}(t) = P_{ij}^{A,l} \left(\sum_{k=1}^d \int_0^t \gamma_{ki}^N(s) H_k^N(s) ds \right), \\
 Y_{ij}^{N,S,l}(t) &= P_{ij}^{S,l} \left(\sum_{k=1}^d \int_0^t \mu_{ki}^N(s) H_k^N(s) ds \right),
 \end{aligned} \tag{2.2}$$

where $P_i^E, P_{ij}^{A,l}, P_{ij}^{S,l}$, $i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\}$ are mutually independent unit-rate Poisson processes with non-negative, non-decreasing trajectories in $\mathbb{D}([0, \infty), \mathbb{R})$ associated with arrival, abandonment and service processes respectively.

3. Heavy-traffic approximations and main results

In this section, we develop the heavy-traffic theory of the GJN system which primarily builds on the functional law of large numbers (FLLN) (Theorem 3.2) and functional central limit theorem (FCLT) (Theorem 3.4). To build the heavy-traffic theory, we introduce a novel asymptotic regime at two-time scale analogy, referred to as two-time-scale time-varying many-server heavy-traffic regime, defined by the scaling presented in Assumptions 3.1, 3.3–3.6.

Assumption 3.1. *The infinitesimal generator matrix of \mathbf{H}^N is accelerated by N^β factor, for some $\beta > 0$, i.e., $\mathcal{Q}^N = \mathcal{O}(N^\beta)$, which is equivalent to say that the increment of jump rates are proportional to N^β .*

Before proceeding to the main results, we assume that the \mathbb{R}^d -valued deterministic rate functions $\boldsymbol{\lambda}_i^N(\cdot) = (\lambda_{ki}^N(\cdot), k \in \mathbb{S}')$, $\boldsymbol{\gamma}_i^N(\cdot) = (\gamma_{ki}^N(\cdot), k \in \mathbb{S}')$, $\boldsymbol{\mu}_i^N(\cdot) = (\mu_{ki}^N(\cdot), k \in \mathbb{S}')$, $i \in \mathbb{L}$ satisfy the following assumptions in the heavy-traffic regime, where (\cdot) stands for the time dependency of the rate functions.

Assumption 3.2. *Assume that every rate functions belong to $\mathbb{C}_c^1([0, \infty), \mathbb{R}^d)$ for each $i \in \mathbb{L}$.*

Assumption 3.3. *For every $i \in \mathbb{L}$, there exists \mathbb{R}^d -valued functions $\boldsymbol{\lambda}_i(\cdot)$, $\boldsymbol{\gamma}_i(\cdot)$, $\boldsymbol{\mu}_i(\cdot)$ such that for any $T > 0$,*

$$\left\| \frac{1}{N} \boldsymbol{\lambda}_i^N(\cdot) - \boldsymbol{\lambda}_i(\cdot) \right\|_T \rightarrow 0, \quad \left\| \boldsymbol{\gamma}_i^N(\cdot) - \boldsymbol{\gamma}_i(\cdot) \right\|_T \rightarrow 0, \quad \left\| \boldsymbol{\mu}_i^N(\cdot) - \boldsymbol{\mu}_i(\cdot) \right\|_T \rightarrow 0, \quad \text{as } N \rightarrow \infty.$$

Assumption 3.4. *For every $i \in \mathbb{L}$, given the deterministic fluid approximations in Assumption 3.3, we define*

$$\begin{aligned} \widehat{\boldsymbol{\lambda}}_i^N(\cdot) &:= N^{1-\delta} \left(\frac{1}{N} \boldsymbol{\lambda}_i^N(\cdot) - \boldsymbol{\lambda}_i(\cdot) \right), \\ \widehat{\boldsymbol{\gamma}}_i^N(\cdot) &:= N^{1-\delta} (\boldsymbol{\gamma}_i^N(\cdot) - \boldsymbol{\gamma}_i(\cdot)), \\ \widehat{\boldsymbol{\mu}}_i^N(\cdot) &:= N^{1-\delta} (\boldsymbol{\mu}_i^N(\cdot) - \boldsymbol{\mu}_i(\cdot)). \end{aligned}$$

Then there exists \mathbb{R}^d -valued functions $\widehat{\boldsymbol{\lambda}}_i(\cdot)$, $\widehat{\boldsymbol{\gamma}}_i(\cdot)$, $\widehat{\boldsymbol{\mu}}_i(\cdot)$, such that for any $T > 0$, and $\delta = \max\{1 - \beta/2, 1/2\}$,

$$\left\| \widehat{\boldsymbol{\lambda}}_i^N(\cdot) - \widehat{\boldsymbol{\lambda}}_i(\cdot) \right\|_T \rightarrow 0, \quad \left\| \widehat{\boldsymbol{\gamma}}_i^N(\cdot) - \widehat{\boldsymbol{\gamma}}_i(\cdot) \right\|_T \rightarrow 0, \quad \left\| \widehat{\boldsymbol{\mu}}_i^N(\cdot) - \widehat{\boldsymbol{\mu}}_i(\cdot) \right\|_T \rightarrow 0, \quad \text{as } N \rightarrow \infty.$$

Given \mathbb{R}^d -valued functions $\mathbf{f}(\cdot) = (f_k(\cdot))_{k=1}^d$, we define $\overline{\mathbf{f}}^\pi(\cdot) := \sum_{k=1}^d f_k(\cdot) \pi_k$, associated with the invariant distribution π . By this definition, for every station $i \in \mathbb{L}$, given the \mathbb{R}^d -valued rate functions $\boldsymbol{\lambda}_i^N(\cdot)$, $\boldsymbol{\gamma}_i^N(\cdot)$, $\boldsymbol{\mu}_i^N(\cdot)$, we introduce the *averaged rate functions* $\overline{\boldsymbol{\lambda}}_i^{N,\pi}(\cdot)$, $\overline{\boldsymbol{\gamma}}_i^{N,\pi}(\cdot)$ and $\overline{\boldsymbol{\mu}}_i^{N,\pi}(\cdot)$. Assumption 3.3 results the averaged fluid rate functions $\overline{\boldsymbol{\lambda}}_i^\pi(\cdot)$, $\overline{\boldsymbol{\gamma}}_i^\pi(\cdot)$ and $\overline{\boldsymbol{\mu}}_i^\pi(\cdot)$. Similarly, for the \mathbb{R}^d -valued diffusion rate functions $\widehat{\boldsymbol{\lambda}}_i^N(\cdot)$, $\widehat{\boldsymbol{\gamma}}_i^N(\cdot)$, $\widehat{\boldsymbol{\mu}}_i^N(\cdot)$, we define $\widehat{\boldsymbol{\lambda}}_i^{N,\pi}(\cdot)$, $\widehat{\boldsymbol{\gamma}}_i^{N,\pi}(\cdot)$ and $\widehat{\boldsymbol{\mu}}_i^{N,\pi}(\cdot)$ and Assumption 3.4 results the averaged diffusion rate functions $\widehat{\boldsymbol{\lambda}}_i^\pi(\cdot)$, $\widehat{\boldsymbol{\gamma}}_i^\pi(\cdot)$ and $\widehat{\boldsymbol{\mu}}_i^\pi(\cdot)$. Denote the corresponding \mathbb{R}^L -valued rate functions as $\overline{\boldsymbol{\lambda}}^{N,\pi}(\cdot)$, $\overline{\boldsymbol{\gamma}}^{N,\pi}(\cdot)$, $\overline{\boldsymbol{\mu}}^{N,\pi}(\cdot)$, $\overline{\boldsymbol{\lambda}}^\pi(\cdot)$, $\overline{\boldsymbol{\gamma}}^\pi(\cdot)$, $\overline{\boldsymbol{\mu}}^\pi(\cdot)$, $\widehat{\boldsymbol{\lambda}}^{N,\pi}(\cdot)$, $\widehat{\boldsymbol{\gamma}}^{N,\pi}(\cdot)$, $\widehat{\boldsymbol{\mu}}^{N,\pi}(\cdot)$, $\widehat{\boldsymbol{\lambda}}^\pi(\cdot)$, $\widehat{\boldsymbol{\gamma}}^\pi(\cdot)$, $\widehat{\boldsymbol{\mu}}^\pi(\cdot)$.

Additionally, assume that the deterministic time-dependent servers $\mathbf{K}^N(\cdot) = (K_i^N(\cdot))_{i=1}^L$ and the state dependent routing probability substochastic matrices $\boldsymbol{\psi}^N(\cdot) = (\psi_{ij}^N(\cdot))_{i,j=1}^L$ and $\boldsymbol{\phi}^N(\cdot) = (\phi_{ij}^N(\cdot))_{i,j=1}^L$ and vectors $\boldsymbol{\psi}_0^N(\cdot) = (\psi_{i0}^N(\cdot))_{i=1}^L$, $\boldsymbol{\phi}_0^N(\cdot) = (\phi_{i0}^N(\cdot))_{i=1}^L$ are \mathbb{R}^L -valued, $\mathbb{M}^{L \times L}$ -valued and \mathbb{R}^L -valued Lebesgue measurable functions respectively, which are scaled accordingly in the following way inside the heavy-traffic regime.

Assumption 3.5. *Define $\overline{\mathbf{K}}^N(\cdot) := \frac{\mathbf{K}^N(\cdot)}{N}$. There exists \mathbb{R}^L -valued, non-negative Lebesgue integrable function $\overline{\mathbf{K}}(\cdot)$ such that $\left\| \overline{\mathbf{K}}^N(\cdot) - \overline{\mathbf{K}}(\cdot) \right\|_T \rightarrow 0$, as $N \rightarrow \infty$. For the routing matrices $\boldsymbol{\psi}^N(\cdot)$, $\boldsymbol{\phi}^N(\cdot)$, define $\overline{\boldsymbol{\psi}}^N(\cdot) := \boldsymbol{\psi}^N(N)$, $\overline{\boldsymbol{\phi}}^N(\cdot) := \boldsymbol{\phi}^N(N)$. There exist $\mathbb{M}^{L \times L}$ -valued Lipschitz functions $\overline{\boldsymbol{\psi}}(\cdot)$ and $\overline{\boldsymbol{\phi}}(\cdot)$ with spectral radius strictly less than 1 such that*

$$\left\| \overline{\boldsymbol{\psi}}^N(\cdot) - \overline{\boldsymbol{\psi}}(\cdot) \right\|_T \rightarrow 0, \quad \left\| \overline{\boldsymbol{\phi}}^N(\cdot) - \overline{\boldsymbol{\phi}}(\cdot) \right\|_T \rightarrow 0, \quad \text{as } N \rightarrow \infty.$$

Similarly, for the \mathbb{R}^L -valued functions $\boldsymbol{\psi}_0^N(\cdot)$ and $\boldsymbol{\phi}_0^N(\cdot)$, there exists $\overline{\boldsymbol{\psi}}_0(\cdot)$ and $\overline{\boldsymbol{\phi}}_0(\cdot)$ such that

$$\left\| \overline{\boldsymbol{\psi}}_0^N(\cdot) - \overline{\boldsymbol{\psi}}_0(\cdot) \right\|_T \rightarrow 0, \quad \left\| \overline{\boldsymbol{\phi}}_0^N(\cdot) - \overline{\boldsymbol{\phi}}_0(\cdot) \right\|_T \rightarrow 0, \quad \text{as } N \rightarrow \infty.$$

In addition, let $\bar{\psi}(\cdot), \bar{\psi}_0(\cdot), \bar{\phi}(\cdot), \bar{\psi}_0(\cdot)$ be differentiable, with derivatives that are bounded on every compact interval.

NOTE: The condition that the spectral radii of the routing matrices $\psi(\cdot)$ and $\phi(\cdot)$ are strictly less than 1 is necessary to ensure the stability of the stochastic network. This assumption guarantees the invertibility of $(\mathbf{I} - \bar{\psi}(\cdot)^T)$ and $(\mathbf{I} - \bar{\phi}(\cdot)^T)$, which is crucial for well-defined network dynamics. Such a condition is standard in the literature on Jackson queueing networks and ensures the existence and uniqueness of the associated diffusion approximations.

Assumption 3.6. Fix $\delta = \max\{1 - \beta/2, 1/2\}$. For the diffusion deterministic approximation, define $\widehat{\mathbf{K}}^N(\cdot) := N^{1-\delta}(\bar{\mathbf{K}}^N(\cdot) - \bar{\mathbf{K}}(\cdot))$. Given fluid servers $\bar{\mathbf{K}}(\cdot)$ in Assumption 3.5, there exists an \mathbb{R}^L -valued Lebesgue measurable function $\widehat{\mathbf{K}}(\cdot)$ so that for all $T > 0$,

$$\|\widehat{\mathbf{K}}^N(\cdot) - \widehat{\mathbf{K}}(\cdot)\|_T \rightarrow 0, \quad \text{as } N \rightarrow \infty.$$

Define $\widehat{\psi}^N(\cdot) := N^{1-\delta}(\bar{\psi}^N(\cdot) - \bar{\psi}(\cdot))$ and $\widehat{\psi}_0^N(\cdot) := N^{1-\delta}(\bar{\psi}_0^N(\cdot) - \bar{\psi}_0(\cdot))$. Given the deterministic fluid approximations $\bar{\psi}(\cdot)$ and $\bar{\phi}(\cdot)$ with spectral radius less than 1 in Assumption 3.5, there exist $\mathbb{M}^{L \times L}$ -valued Lipschitz functions $\widehat{\psi}(\cdot)$ and $\widehat{\phi}(\cdot)$ such that for any $T > 0$,

$$\|\widehat{\psi}^N(\cdot) - \widehat{\psi}(\cdot)\|_T \rightarrow 0, \quad \|\widehat{\psi}_0^N(\cdot) - \widehat{\psi}_0(\cdot)\|_T \rightarrow 0, \quad \text{as } N \rightarrow \infty.$$

In similar manner, given $\bar{\psi}_0(\cdot), \bar{\phi}_0(\cdot) \in \mathbb{R}^L$ from Assumption 3.5, there exists $\widehat{\psi}_0(\cdot)$ and $\widehat{\phi}_0(\cdot)$ such that

$$\|\widehat{\psi}_0^N(\cdot) - \widehat{\psi}_0(\cdot)\|_T \rightarrow 0, \quad \|\widehat{\phi}_0^N(\cdot) - \widehat{\phi}_0(\cdot)\|_T \rightarrow 0, \quad \text{as } N \rightarrow \infty.$$

Remark 3.1 (Averaged Halfin-Whitt regime). Assumptions 3.4, 3.6 generate the Halfin-Whitt asymptotic regime for each station $i \in \mathbb{L}$ of the GJN system inside the time-varying framework. In addition, if the \mathbb{R}^L -valued deterministic functions $\lambda^N(\cdot), \gamma^N(\cdot), \mu^N(\cdot)$ and $\mathbf{K}^N(\cdot)$ are assumed to be constants, and the routing probabilities $\psi(\cdot)$ and $\phi(\cdot)$ are state-independent, the averaged Halfin-Whitt regime arises in the system whenever $\bar{\lambda}^{N,\pi} = (\mathbf{I} - \bar{\phi}^T) \mathbf{diag}(\mathbf{K}^N) \bar{\mu}^\pi$ for large N , with respect to the invariant distribution π of \mathbf{H}^N . Using Assumptions 3.3, 3.5, we obtain that

$$\bar{\lambda}^\pi = (\mathbf{I} - \bar{\phi}^T) \mathbf{diag}(\bar{\mathbf{K}}) \bar{\mu}^\pi, \tag{3.1}$$

which is the required condition for the system to be critically loaded. The condition (3.1) yields that the total input at each station $i \in \mathbb{L}$ coincides with the total service capacity, i.e.,

$$\bar{\lambda}_i^{N,\pi} + \sum_{i=1}^{\mathbb{L}_i} \bar{\phi}_{ji} K_j^N \bar{\mu}_j^\pi = \sum_{j=1}^{\mathbb{L}_i} \bar{\phi}_{ij} K_i^N \bar{\mu}_i^\pi + \phi_{i0} K_i^N \bar{\mu}_i^\pi, \quad \text{for large } N.$$

If $\alpha^{N,\pi} = N^{-\delta}(\bar{\lambda}^{N,\pi} - (\mathbf{I} - \bar{\phi}^T) \mathbf{diag}(\mathbf{K}^N) \bar{\mu}^\pi)$, it follows from Assumptions 3.4 and 3.6 that $\lim_{N \rightarrow \infty} \alpha^{N,\pi} = \alpha^\pi$, where $\alpha^\pi = \widehat{\lambda}^\pi - (\mathbf{I} - \bar{\phi}^T) \mathbf{diag}(\bar{\mathbf{K}}) \widehat{\mu}^\pi$ under averaged Halfin-Whitt regime.

For each $N \in \mathbb{N}$, given a process $\{\mathcal{X}^N(t), t \geq 0\}$ with trajectories in $\mathbb{D}([0, \infty), \mathcal{E})$, we define the fluid scaled processes as $\{\bar{\mathcal{X}}^N(t) := \frac{\mathcal{X}^N(t)}{N}, t \geq 0\}$ and diffusion scaled process as $\{\widehat{\mathcal{X}}^N(t) := N^{1-\delta}(\bar{\mathcal{X}}^N(t) - \bar{\mathcal{X}}(t)), t \geq 0\}$. In the remaining of the article, we strict to the notations defined above to denote fluid and diffusion scaled process accordingly. The next two results Theorem 3.2 and 3.4 are the main contributions of the article, which shows that the fluid scaled system length process $\{\bar{\mathbf{Q}}^N(t) = (\bar{Q}_i^N(t))_{i=1}^L, t \geq 0\}$ converges to a deterministic system averaged by the invariant distribution π , and the diffusion scaled system length process $\{\widehat{\mathbf{Q}}^N(t) = (\widehat{Q}_i^N(t))_{i=1}^L, t \geq 0\}$ converges in distribution to a certain diffusion process as $N \rightarrow \infty$ for every value of diffusion parameter $\beta < 1$, or 1, or > 1 .

Theorem 3.2 (Fluid approximation). *Suppose that the initial fluid scaled system length process $\bar{Q}^N(0)$ converges to $\bar{Q}(0)$ in \mathbb{R}^L , as $N \rightarrow \infty$. Under Assumption 3.2–3.3, and Assumption 3.5, $\{\bar{Q}^N(t), t \geq 0\}$ converges to $\{\bar{Q}(t) = (\bar{Q}_i(t))_{i=1}^L, t \geq 0\}$ in probability uniformly on compact sets as $N \rightarrow \infty$, where $\bar{Q}(t)$ uniquely satisfies the following deterministic integral equation*

$$\begin{aligned} \bar{Q}(t) = & \bar{Q}(0) + \int_0^t \bar{\lambda}^\pi(s) ds - \int_0^t \left(\mathbf{I} - \bar{\psi}(\bar{Q}(s))^T \right) \text{diag}(\bar{Q}(s) - \bar{K}(s))^+ \bar{\gamma}^\pi(s) ds \\ & - \int_0^t \left(\mathbf{I} - \bar{\phi}(\bar{Q}(s))^T \right) \text{diag}(\bar{Q}(s) \wedge \bar{K}(s)) \bar{\mu}^\pi(s) ds, \end{aligned} \tag{3.2}$$

where \mathbf{I} is an $L \times L$ -dimensional identity matrix.

Remark 3.3. Theorem 3.2 shows that the fluid approximation \bar{Q} has deterministic trajectories in $\mathbb{D}([0, \infty), \mathbb{R}^L)$ for given deterministic approximations resulting from Assumptions 3.3, 3.5, which is averaged by the invariant distribution π of the external environment. In particular,

- If $\sup_{0 \leq t \leq T} \|\bar{Q}(t) - \bar{K}(t)\| < 0$, the system is said to be *subcritical*.
- If $\inf_{0 \leq t \leq T} \|\bar{Q}(t) - \bar{K}(t)\| > 0$, the system is said to be *supercritical*.
- If $\sup_{0 \leq t \leq T} \|\bar{Q}(t) - \bar{K}(t)\| = 0$, the system is said to be *critical*.

In particular, the system becomes critically loaded if

$$\sup_{0 \leq t \leq T} \left\| \int_0^t \bar{\lambda}^\pi(s) ds - \int_0^t \left(\mathbf{I} - \bar{\phi}(\bar{Q}(s))^T \right) \text{diag}(\bar{K}(s)) \bar{\mu}^\pi(s) ds \right\| = 0, \tag{3.3}$$

which represents the Halfin-Whitt condition of the system.

In particular, if $\bar{\lambda}^\pi(\cdot)$, $\bar{\mu}^\pi(\cdot)$, $\bar{K}(\cdot)$ and $\bar{\phi}(\cdot)$ are constants, then (3.3) yields

$$\bar{\lambda}^\pi - \left(\mathbf{I} - \bar{\phi}^T \right) \text{diag}(\bar{K}) \bar{\mu}^\pi = 0,$$

which actually coincides with the averaged Halfin-Whitt regime condition (3.1), see Remark 3.1.

Theorem 3.4 (Diffusion approximation). *Suppose that $\hat{Q}^N(0)$ converges in distribution to $\hat{Q}(0)$ in \mathbb{R}^L , as $N \rightarrow \infty$, and Assumption 3.2–3.6 are satisfied. In addition, the fluid approximation \bar{Q} of Theorem 3.2 is subcritical, critical, or supercritical. Then $\{\hat{Q}^N(t), t \geq 0\}$ converges in distribution to $\{\hat{Q}(t) = (\hat{Q}_i(t))_{i=1}^L, t \geq 0\}$ in $\mathbb{D}([0, \infty), \mathbb{R}^L)$ as $N \rightarrow \infty$, where the limit $\hat{Q}(t)$ uniquely satisfies the stochastic integral equation as follows.*

$$\begin{aligned} \hat{Q}(t) = & \hat{Q}(0) + \mathbf{1}(\beta \leq 1) \left(\mathbf{W}_\lambda(t) - \int_0^t \left(\mathbf{I} - \bar{\psi}(\bar{Q}(s))^T \right) \text{diag}(\bar{Q}(s) - \bar{K}(s))^+ d\mathbf{W}_\gamma(s) \right. \\ & \left. - \int_0^t \left(\mathbf{I} - \bar{\phi}(\bar{Q}(s))^T \right) \text{diag}(\bar{Q}(s) \wedge \bar{K}(s)) d\mathbf{W}_\mu(s) \right) + \int_0^t \hat{\lambda}^\pi(s) ds \\ & + \int_0^t \left(\hat{\psi}(\bar{Q}(s)) \right)^T \text{diag}(\bar{Q}(s) - \bar{K}(s))^+ \bar{\gamma}^\pi(s) ds \\ & - \int_0^t \left(\mathbf{I} - \bar{\psi}(\bar{Q}(s))^T \right) \text{diag}(\bar{Q}(s) - \bar{K}(s))^+ \hat{\gamma}^\pi(s) ds \\ & + \int_0^t \left(\hat{\phi}(\bar{Q}(s)) \right)^T \text{diag}(\bar{Q}(s) \wedge \bar{K}(s)) \bar{\mu}^\pi(s) ds \\ & - \int_0^t \left(\mathbf{I} - \bar{\phi}(\bar{Q}(s))^T \right) \text{diag}(\bar{Q}(s) \wedge \bar{K}(s)) \hat{\mu}^\pi(s) ds + \mathbf{1}(\beta \geq 1) \mathbf{B}(t) \end{aligned}$$

$$\begin{aligned}
 & + \int_0^t \langle \nabla \bar{\psi}(\bar{Q}(s-)), \hat{Q}(s) \rangle^T \text{diag}(\bar{Q}(s) - \bar{K}(s))^+ \bar{\gamma}^\pi(s) \, ds, \\
 & + \int_0^t \langle \nabla \bar{\phi}(\bar{Q}(s-)), \hat{Q}(s) \rangle^T \text{diag}(\bar{Q}(s) \wedge \bar{K}(s)) \bar{\mu}^\pi(s) \, ds, \\
 & - \int_0^t (\mathbf{I} - \bar{\psi}(\bar{Q}(s-)))^T \text{diag}(h_1(\hat{Q}(s))) \bar{\gamma}^\pi(s) \, ds \\
 & - \int_0^t (\mathbf{I} - \bar{\phi}(\bar{Q}(s-)))^T \text{diag}(h_2(\hat{Q}(s))) \bar{\mu}^\pi(s) \, ds,
 \end{aligned} \tag{3.4}$$

with $h_1(\hat{Q}(t)) = (\hat{Q}(t) - \hat{K}(t))\mathbf{1}(\bar{Q}(t) > \bar{K}(t)) + (\hat{Q}(t) - \hat{K}(t))^+\mathbf{1}(\bar{Q}(t) = \bar{K}(t))$, $h_2(\hat{Q}(t)) = \hat{Q}(t)\mathbf{1}(\bar{Q}(t) < \bar{K}(t)) + \hat{Q}(t) \wedge \hat{K}(t)\mathbf{1}(\bar{Q}(t) = \bar{K}(t)) + \hat{K}(t)\mathbf{1}(\bar{Q}(t) > \bar{K}(t))$, where $\{\mathbf{W}_\lambda(t), t \geq 0\}$, $\{\mathbf{W}_\gamma(t), t \geq 0\}$, $\{\mathbf{W}_\mu(t), t \geq 0\}$ are correlated Brownian motions in $\mathbb{D}([0, \infty), \mathbb{R}^L)$, and the associated covariance function is given by $C_{fg}(t) = (C_{f_i g_j}(t), 1 \leq i, j \leq L)$, $f_i(t) = \lambda_i(t), \gamma_i(t), \mu_i(t)$, $g_j(t) = \lambda_j(t), \gamma_j(t), \mu_j(t)$

$$C_{f_i g_j}(t) = \sum_{k=1}^d \sum_{l=1}^d \pi_k \mathcal{D}_{kl} \int_0^t (f_{ik}(s)g_{jl}(s) + f_{jl}(s)g_{ik}(s)) \, ds, \quad 1 \leq i, j \leq L, \tag{3.5}$$

and $\{\mathbf{B}(t), t \geq 0\}$ is an independent drift-less Brownian motion in $\mathbb{D}([0, \infty), \mathbb{R}^L)$ with

$$\begin{aligned}
 \frac{d}{dt} \mathbb{E}(\mathbf{B}(t)\mathbf{B}(t)^T) & = \text{diag}(\bar{\lambda}^\pi(t) + (\mathbf{I} + \bar{\psi}(\bar{Q}(t)))^T \text{diag}(\bar{Q}(t) - \bar{K}(t))^+ \bar{\gamma}^\pi(t) + (\mathbf{I} + \bar{\phi}(\bar{Q}(t)))^T \\
 & \quad \times \text{diag}(\bar{Q}(t) \wedge \bar{K}(t)) \bar{\mu}^\pi(t)) - \bar{\psi}(\bar{Q}(t))^T \text{diag}(\bar{Q}(t) - \bar{K}(t))^+ \text{diag}(\bar{\gamma}^\pi(t)) \\
 & \quad - \text{diag}(\bar{\gamma}^\pi(t)) \text{diag}(\bar{Q}(t) - \bar{K}(t))^+ \bar{\psi}(\bar{Q}(t)) - \bar{\phi}(\bar{Q}(t))^T \text{diag}(\bar{Q}(t) \wedge \bar{K}(t)) \\
 & \quad \times \text{diag}(\bar{\mu}^\pi(t)) - \text{diag}(\bar{\mu}^\pi(t)) \text{diag}(\bar{Q}(t) \wedge \bar{K}(t)) \bar{\phi}(\bar{Q}(t)).
 \end{aligned}$$

Remark 3.5. Under Assumptions 3.5–3.6, Theorem 3.4 shows that the diffusion scaled system length process \hat{Q}^N converges in distribution to a \mathbb{R}^L -valued Itô diffusion process \hat{Q} regardless of whether the system is subcritical, critical or supercritical. In the diffusion component, the time-varying diffusion coefficients are determined by the deterministic approximations of the rate functions as well as the state-dependent routing probabilities and system length process. In contrast, the time-varying drift coefficients are additionally dependent on the functional \hat{Q} . As a consequence, the limiting process \hat{Q} can be characterized as the multi-dimensional Ornstein–Uhlenbeck process with time-varying drift and diffusion coefficients. Therefore, using a similar analogy to evaluate the performance measures (i.e., mean, covariance, etc.) of the Ornstein–Uhlenbeck process, one can derive the performance measures of interest.

Remark 3.6. The diffusion scaled parameter $\delta = \max\{1 - \beta/2, 1/2\}$ plays a crucial role in deriving the particular form of the limiting process \hat{Q} in (3.4). For example, if $\beta < 1$, which is equivalent to saying that the transition speed of the external Markovian environment is slow compared to the arrival speed of the exogenous arrivals at each station $i \in \mathbb{L}$, the dynamics of the resulting system \hat{Q} are determined by the \mathbb{R}^L -valued rate functions $(\lambda_{ki}(\cdot))_{i=1}^L, (\gamma_{ki}(\cdot))_{i=1}^L, (\mu_{ki}(\cdot))_{i=1}^L$ when the external process stays at a particular state $k \in \mathbb{S}'$. On the contrary, when $\beta > 1$, the arrivals join from outside in each station i at a much lower speed than the transition speed of of the external process, and hence \hat{Q} leads to a *averaged* system with averaged \mathbb{R}^L -valued rate functions $\bar{\lambda}^\pi(\cdot), \bar{\gamma}^\pi(\cdot), \bar{\mu}^\pi(\cdot)$ with respect to the invariant distribution π , which eventually behaves as the general GJN system without acting under random environment. But, in particular, if $\beta = 1$, the speed of transitions of the external process and the exogenous arrivals turn out to be of $\mathcal{O}(N)$, and as a result, the system successfully captures the stochastic as well as the predictable behaviour simultaneously.

Remark 3.7. If the parameters $\lambda^N(\cdot)$, $\mu^N(\cdot)$, $\gamma^N(\cdot)$, $\mathbf{K}^N(\cdot)$, $\psi^N(\cdot)$ and $\phi^N(\cdot)$ are independent of time evolution and the system is operating under critically loaded Halfin-Whitt regime, i.e. the condition (3.3) holds (see Remark 3.1), then the diffusion limit $\widehat{\mathbf{Q}}(t)$ has the following form

$$\begin{aligned} \widehat{\mathbf{Q}}(t) = & \widehat{\mathbf{Q}}(0) + \mathbf{1}(\beta \leq 1) \left(\mathbf{W}_\lambda(t) - \left(\mathbf{I} - \overline{\phi}^T \right) \text{diag} \overline{\mathbf{K}} \mathbf{W}_\mu(t) \right) + \alpha^\pi t + \widehat{\phi}^T \text{diag} \overline{\mathbf{K}} \overline{\mu}^\pi t + \mathbf{1}(\beta \geq 1) \mathbf{B}(t) \\ & - \int_0^t \left(\mathbf{I} - \overline{\psi}^T \right) \text{diag} \left(\widehat{\mathbf{Q}}(s) - \widehat{\mathbf{K}} \right)^+ \overline{\gamma}^\pi ds - \int_0^t \left(\mathbf{I} - \overline{\phi}^T \right) \text{diag} \left(\widehat{\mathbf{Q}}(s) \wedge \widehat{\mathbf{K}} \right) \overline{\mu}^\pi ds, \end{aligned}$$

where $\alpha^\pi = \widehat{\lambda}^\pi - \left(\mathbf{I} - \phi^T \right) \text{diag} \overline{\mathbf{K}} \widehat{\mu}^\pi$, and

$$\frac{d}{dt} \mathbb{E} \left(\mathbf{B}(t) \mathbf{B}(t)^T \right) = \text{diag} \left(\overline{\lambda}^\pi + \left(\mathbf{I} + \overline{\phi}^T \right) \text{diag} \overline{\mathbf{K}} \overline{\mu}^\pi \right) - \overline{\phi}^T \text{diag} \overline{\mathbf{K}} \text{diag} \overline{\mu}^\pi - \text{diag} \left(\overline{\mu}^\pi \right) \text{diag} \overline{\mathbf{K}} \overline{\phi}.$$

4. Martingale representation

The dynamical representation (2.1) of the system length process at station i of the N th system admits a martingale representation incorporating the martingales associated with the counting processes \mathbf{E}^N , $\mathbf{Y}^{N,A,l}$, $\mathbf{Y}^{N,S,l}$, $\mathbf{Y}_0^{N,A,l}$, $\mathbf{Y}_0^{N,S,l}$. The foundations of heavy-traffic approximations (Theorems 3.2, 3.4) of the system length process \mathbf{Q}^N are built on the martingales associated with the above counting processes, and their convergence in the respective Skorokhod space.

For every $N \in \mathbb{N}$, we introduce the sigma algebra $\mathcal{H}^N(t) = \sigma(\mathbf{H}^N(s), s \leq t)$, and the right-continuous filtration $\mathcal{H}^N = \{\mathcal{H}^N(t), t \geq 0\}$ associated with the external Markov process \mathbf{H}^N . We define the complete sigma algebra

$$\mathcal{F}^N(t) = \sigma(\mathbf{Q}^N(0), \mathbf{E}^N(s), \mathbf{Y}^{N,A,l}(s), \mathbf{Y}^{N,S,l}(s), \mathbf{Y}_0^{N,A,l}(s), \mathbf{Y}_0^{N,S,l}(s), l \in \mathbb{N}, s \leq t) \vee \mathcal{N},$$

where \mathcal{N} represents the family of \mathcal{P} -null sets. By definition, $\mathcal{H}^N(t) \subseteq \mathcal{F}^N(t)$, for all $t \geq 0$. For $i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\}$, we introduce

$$\begin{aligned} Z_{\lambda_i^N}^N(t) &= \sum_{k=1}^d \int_0^t \lambda_{ki}^N(s) H_k^N(s) ds, \\ Z_{\gamma_i^N}^N(t) &= \sum_{k=1}^d \int_0^t \gamma_{ki}^N(s) H_k^N(s) ds, \\ Z_{\mu_i^N}^N(t) &= \sum_{k=1}^d \int_0^t \mu_{ki}^N(s) H_k^N(s) ds, \end{aligned} \tag{4.1}$$

and the corresponding \mathbb{R}^L -valued processes are denoted by $\{\mathbf{Z}_{\lambda^N}^N(t), t \geq 0\}$, $\{\mathbf{Z}_{\gamma^N}^N(t), t \geq 0\}$, $\{\mathbf{Z}_{\mu^N}^N(t), t \geq 0\}$. It follows from the definition that, for every $N \in \mathbb{N}$, each process $\mathbf{Z}_{\lambda^N}^N$, $\mathbf{Z}_{\gamma^N}^N$, $\mathbf{Z}_{\mu^N}^N$ is $\mathcal{H}^N(t)$ -measurable, and thus \mathcal{F}^N -adapted predictable processes. Also these Lebesgue-Stieltjes integral processes have non-decreasing trajectories in $\mathbb{C}([0, \infty), \mathbb{R}^L)$. According to the construction, $E_i^N(t) = P_i^E \circ Z_{\lambda_i^N}^N(t)$, $Y_{ij}^{N,A,l}(t) = P_{ij}^{A,l} \circ Z_{\gamma_i^N}^N$, $Y_{ij}^{N,S,l}(t) = P_{ij}^{S,l} \circ Z_{\mu_i^N}^N(t)$, $i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\}$. It is easy to verify that $E_i^N - Z_{\lambda_i^N}^N$, $Y_{ij}^{N,A,l} - Z_{\gamma_i^N}^N$, $Y_{ij}^{N,S,l} - Z_{\mu_i^N}^N$ are \mathcal{F}^N -local martingales with compensators $Z_{\lambda_i^N}^N$, $Z_{\gamma_i^N}^N$, and $Z_{\mu_i^N}^N$ respectively. For example, if $\tau^N(n) := \inf\{t \geq 0 : Z_{\lambda_i^N}^N(t) \geq n\}$, then $\tau^N(n)$ is an \mathcal{F}^N -stopping time for every $n \in \mathbb{N}$, and $\tau^N(n) \uparrow \infty$ as $n \rightarrow \infty$ \mathbb{P} -a.s. By conditional expectation property, we can show that $\{E_i^N(t \wedge \tau^N(n)) - Z_{\lambda_i^N}^N(t \wedge \tau^N(n))\}$ is an \mathcal{F}^N -martingale, for each $i \in \mathbb{L}$. Consequently, Theorem 1.6.3 of Liptser and Shiriyayev (1989) yields that $Z_{\lambda^N}^N$, $Z_{\gamma^N}^N$ and $Z_{\mu^N}^N$ are \mathcal{F}^N -predictable compensators of the counting processes E_i^N , $Y_{ij}^{N,A,l}$ and $Y_{ij}^{N,S,l}$ respectively, for each $l \in \mathbb{N}, i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\}$.

We define the processes

$$\begin{aligned} \{ \mathbf{M}^{N,E}(t) = (M_i^{N,E}(t))_{i=1}^L, t \geq 0 \}, \quad \{ \mathbf{M}_0^{N,A}(t) = (M_{i0}^{N,A}(t))_{i=1}^L, t \geq 0 \}, \\ \{ \mathbf{M}_0^{N,S}(t) = (M_{i0}^{N,S}(t))_{i=1}^L, t \geq 0 \} \end{aligned}$$

with trajectories in $\mathbb{D}([0, \infty), \mathbb{R}^L)$, and

$$\{ \mathbf{M}^{N,A}(t) = (M_{ij}^{N,A}(t))_{i,j=1}^L, t \geq 0 \}, \quad \{ \mathbf{M}^{N,S}(t) = (M_{ij}^{N,S}(t))_{i,j=1}^L, t \geq 0 \}$$

with trajectories in $\mathbb{D}([0, \infty), \mathbb{M}^{L \times L})$ with respect to the compensated processes

$$E_i^N - Z_{\lambda_i^N}^N, \quad Y_{ij}^{N,A,l} - Z_{\gamma_i^N}^N, \quad Y_{ij}^{N,S,l} - Z_{\mu_i^N}^N, \quad \text{for } i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\}, l \in \mathbb{N},$$

as follows:

$$\begin{aligned} M_i^{N,E}(t) &= E_i^N(t) - Z_{\lambda_i^N}^N(t), \\ M_{ij}^{N,A}(t) &= \sum_{l=1}^{\infty} \int_0^t \mathbf{1}(Q_i^N(s-) \geq K_i^N(s-) + l) \psi_{ij}^N(Q_i^N(s-)) d(Y_{ij}^{N,A,l}(s) - Z_{\gamma_i^N}^N(s)), \\ M_{ij}^{N,S}(t) &= \sum_{l=1}^{\infty} \int_0^t \mathbf{1}(Q_i^N(s-) \wedge K_i^N(s-) \geq l) \phi_{ij}^N(Q_i^N(s-)) d(Y_{ij}^{N,S,l}(s) - Z_{\mu_i^N}^N(s)), \end{aligned} \quad (4.2)$$

for $i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\}$. In the next lemma, we show that the processes defined in (4.2) are square-integrable local martingales with certain predictable variation processes. The convergence results (Lemma 7.1 and 8.1) related to the martingales (4.2) are presented in later sections. These results are straightforward consequences of the standard martingale theory associated with counting processes, so we discard the proofs.

Lemma 4.1. *The \mathbb{R}^L -valued processes $\mathbf{M}^{N,E}$, $\mathbf{M}_0^{N,A}$ and $\mathbf{M}_0^{N,S}$ are \mathcal{F}^N -local square-integrable martingales with respective $\mathbb{M}^{L \times L}$ -valued predictable variation processes $\{\mathbf{diag}(\langle \mathbf{M}^{N,E} \rangle(t)), t \geq 0\}$, $\{\mathbf{diag}(\langle \mathbf{M}_0^{N,A} \rangle(t)), t \geq 0\}$, $\{\mathbf{diag}(\langle \mathbf{M}_0^{N,S} \rangle(t)), t \geq 0\}$, and $\mathbb{M}^{L \times L}$ -valued processes $\mathbf{M}^{N,A}$ and $\mathbf{M}^{N,S}$ are \mathcal{F}^N -local square-integrable martingales with $\mathbb{M}^{L \times L}$ -valued predictable variation processes $\{\langle \mathbf{M}^{N,A} \rangle(t), t \geq 0\}$ and $\{\langle \mathbf{M}^{N,S} \rangle(t), t \geq 0\}$, where $\langle \mathbf{M}^{N,E} \rangle(t) = (\langle M_i^{N,E} \rangle(t))_{i=1}^L$, $\langle \mathbf{M}_0^{N,A} \rangle(t) = (\langle M_{i0}^{N,A} \rangle(t))_{i=1}^L$, $\langle \mathbf{M}_0^{N,S} \rangle(t) = (\langle M_{i0}^{N,S} \rangle(t))_{i=1}^L$, and $\langle \mathbf{M}^{N,A} \rangle(t) = (\langle M_{ij}^{N,A} \rangle(t))_{i,j=1}^L$, $\langle \mathbf{M}^{N,S} \rangle(t) = (\langle M_{ij}^{N,S} \rangle(t))_{i,j=1}^L$ with*

$$\begin{aligned} \langle M_i^{N,E} \rangle(t) &= Z_{\lambda_i^N}^N(t), \quad i \in \mathbb{L}, \\ \langle M_{ij}^{N,A} \rangle(t) &= \int_0^t (Q_i^N(s) - K_i^N(s))^+ \psi_{ij}^N(Q_i^N(s-)) dZ_{\gamma_i^N}^N(s), \quad i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\} \\ \langle M_{ij}^{N,S} \rangle(t) &= \int_0^t (Q_i^N(s) \wedge K_i^N(s)) \phi_{ij}^N(Q_i^N(s-)) dZ_{\mu_i^N}^N(s), \quad i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\}. \end{aligned} \quad (4.3)$$

In addition, the martingales $\mathbf{M}^{N,E}$, $\mathbf{M}_0^{N,A}$, $\mathbf{M}_0^{N,S}$, $\mathbf{M}^{N,A}$ and $\mathbf{M}^{N,S}$ are pairwise orthogonal to each other.

In addition, we define the integral processes

$$\begin{aligned} \{ \mathbf{R}^{N,E}(t) = (R_i^{N,E}(t))_{i=1}^L, t \geq 0 \}, \quad \{ \mathbf{R}_0^{N,A}(t) = (R_{i0}^{N,A}(t))_{i=1}^L, t \geq 0 \}, \\ \{ \mathbf{R}_0^{N,S}(t) = (R_{i0}^{N,S}(t))_{i=1}^L, t \geq 0 \} \end{aligned}$$

in $\mathbb{D}([0, \infty), \mathbb{R}^L)$, and

$$\{ \mathbf{R}^{N,A}(t) = (R_{ij}^{N,A}(t))_{i,j=1}^L, t \geq 0 \}, \quad \{ \mathbf{R}^{N,S}(t) = (R_{ij}^{N,S}(t))_{i,j=1}^L, t \geq 0 \}$$

in $\mathbb{D}([0, \infty), \mathbb{M}^{L \times L})$, where each component is defined by

$$\begin{aligned} R_i^{N,E}(t) &= Z_{\lambda_i^N}^N(t), \quad i \in \mathbb{L}, \\ R_{ij}^{N,A}(t) &= \int_0^t (Q_i^N(s) - K_i^N(s))^+ \psi_{ij}^N(Q_i^N(s-)) dZ_{\gamma_i^N}^N(s), \quad i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\}, \\ R_{ij}^{N,S}(t) &= \int_0^t (Q_i^N(s) \wedge K_i^N(s)) \phi_{ij}^N(Q_i^N(s-)) dZ_{\mu_i^N}^N(s), \quad i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\}, \end{aligned} \tag{4.4}$$

By definition (4.4), these Lebesgue-Stieltjes integral processes with respect to the finite-variation process $\mathbf{Z}_{\lambda^N}^N, \mathbf{Z}_{\gamma^N}^N, \mathbf{Z}_{\mu^N}^N$ are \mathcal{F}^N -adapted semimartingales in $\mathbb{D}([0, \infty), \mathbb{R}^L)$.

To the end of this section, we represent the martingale representation of the evolution of the stochastic GJN system defined in (2.1) as follows:

$$\begin{aligned} \mathbf{Q}^N(t) &= \mathbf{Q}^N(0) + \mathbf{M}^{N,E}(t) + (\mathbf{M}^{N,A}(t)^T - \mathbf{M}^{N,A}(t)) \cdot \mathbf{1} + (\mathbf{M}^{N,S}(t)^T - \mathbf{M}^{N,S}(t)) \cdot \mathbf{1} \\ &\quad - \mathbf{M}_0^{N,A}(t) - \mathbf{M}_0^{N,S}(t) + \mathbf{R}^{N,E}(t) + (\mathbf{R}^{N,A}(t)^T - \mathbf{R}^{N,A}(t)) \cdot \mathbf{1} \\ &\quad + (\mathbf{R}^{N,S}(t)^T - \mathbf{R}^{N,S}(t)) \cdot \mathbf{1} - \mathbf{R}_0^{N,A}(t) - \mathbf{R}_0^{N,S}(t). \end{aligned} \tag{4.5}$$

5. Gaussian approximations for the external random process

The main two proofs of the present article Theorems 3.2 and 3.4 are essentially determined by stochastic approximations of the external process \mathbf{H}^N , which are presented in Theorems 5.6 and 5.8 in this section. In particular, we treat the rate mechanisms of the primary network interacted with the external process as an coupled random process $\{(\mathbf{X}^N(t), \mathbf{H}^N(t)), t \geq 0\}$ representing the trajectories (5.1). This dynamics motivate us to adapt the *non-homogeneous stochastic averaging principle* under two-time scale analogy, which is highly influenced by Papanicolaou et al. (1977) and Skorokhod (1989).

5.1. *Averaging principle for external process under central limit theorem regime.* For each $N \in \mathbb{N}$, we consider a joint Markov process $\{(\mathbf{X}^N(t), \mathbf{H}^N(t)), t \geq 0\}$ with measurable state space $(\mathbb{R}^L \times \mathbb{S}, \mathcal{B}_{\mathbb{R}^L} \otimes \mathcal{B}_{\mathbb{S}})$, where $\mathbb{S} = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_d\}$ is a finite subset of \mathbb{R}^d embedded with discrete topology. The first component $\{\mathbf{X}^N(t), t \geq 0\}$ has continuous trajectories in $\mathbb{C}([0, \infty), \mathbb{R}^L)$ \mathbb{P} -a.s.. The other component \mathbf{H}^N is assumed to be a jump Markov process with trajectories in $\mathbb{D}([0, \infty), \mathbb{R}^d)$. The increment of intensity of the jumps of the discrete component \mathbf{H}^N is taken proportional to N^β , for some $\beta > 0$, which is equivalent to scaling the time factor t by N^β , i.e., $\mathbf{H}^N(t) = \mathbf{H}(N^\beta t)$. For each $N \in \mathbb{N}$, the stochastic evolution of the joint process $(\mathbf{X}^N, \mathbf{H}^N)$ is defined by the following differential equation

$$d\mathbf{X}^N(t) = N^{\beta/2} \mathbf{A}^N(t, \mathbf{X}^N(t), \mathbf{H}^N(t)) dt + \mathbf{B}^N(t, \mathbf{X}^N(t), \mathbf{H}^N(t)) dt, \tag{5.1}$$

with initial condition $\mathbf{X}^N(0) = \mathbf{x}, \mathbf{H}^N(0) = \mathbf{h}$. Here, the dynamics are determined through the coefficients $\mathbf{A}^N : [0, \infty) \times \mathbb{R}^L \times \mathbb{S} \rightarrow \mathbb{R}^L, \mathbf{B}^N : [0, \infty) \times \mathbb{R}^L \times \mathbb{S} \rightarrow \mathbb{R}^L$. Given the coefficients $\mathbf{A}^N(t, \mathbf{x}, \mathbf{h}), \mathbf{B}^N(t, \mathbf{x}, \mathbf{h})$, the continuous component \mathbf{X}^N follows the trajectories of the vector field $N^{\beta/2} \mathbf{A}^N(t, \mathbf{x}, \mathbf{h}) + \mathbf{B}^N(t, \mathbf{x}, \mathbf{h})$ in between each transitions of the jump process \mathbf{H}^N . Our aim is to show that this particular stochastic system will reach to equilibrium by speeding up the transition rates of the jump process under suitable conditions. In particular, Theorem 5.1 proves that, based on the longer run behaviour of the jump process \mathbf{H}^N under certain asymptotic regime, the asymptotic distribution of \mathbf{X}^N can be essentially determined.

Consider the joint Markov process $\{(\mathbf{X}(t), \mathbf{H}(t)), t \geq 0\}$ with trajectories in $\mathbb{C}([0, \infty), \mathbb{R}^L) \times \mathbb{D}([0, \infty), \mathbb{R}^d)$. For each $\mathbf{x} \in \mathbb{R}^L$, let $q : \mathbb{R}^L \times \mathbb{S} \rightarrow [0, \infty)$ be the intensity rate function, and $\mathcal{R} : \mathbb{R}^L \times \mathbb{S} \times \mathbb{S} \rightarrow [0, 1]$ be the transition probability kernel satisfying $\sum_{z \in \mathbb{S}} \mathcal{R}(\mathbf{x}, \mathbf{h}, z) = 1$ and

$\mathcal{R}(\mathbf{x}, \mathbf{h}, \mathbf{h}) = 0$. Let $\{\mathcal{T}_n, n \in \mathbb{N} \cup \{0\}\}$ be the sequence of the jump times of \mathbf{H} . Besides, the \mathbb{R}^L -valued diffusion component $\{\mathbf{X}(t), t \geq 0\}$ has continuous trajectories \mathbb{P} -a.s., and thus the transition probability kernel can be expressed as

$$\mathcal{R}(\mathbf{x}, \mathbf{h}, \mathcal{A}) = \mathbb{P}(\mathbf{H}(\mathcal{T}_1) \in \mathcal{A} | \mathbf{X}(\mathcal{T}_1-) = \mathbf{x}, \mathbf{H}(\mathcal{T}_1-) = \mathbf{h}), \mathbf{x} \in \mathbb{R}^L, \mathbf{h} \in \mathbb{S}, \mathcal{A} \in \mathcal{B}_{\mathbb{S}}.$$

We will occasionally prefer $q_{\mathbf{x}}(\cdot), \mathcal{R}_{\mathbf{x}}(\mathbf{h}, \cdot)$ in stead of $q(\mathbf{x}, \cdot), \mathcal{R}(\mathbf{x}, \mathbf{h}, \cdot)$ in the remaining part of this section. For every $\mathbf{x} \in \mathbb{R}^L$, let $\mathcal{Q}_{\mathbf{x}}$ be the generating operator acting on every bounded $\mathcal{B}_{\mathbb{S}}$ -measurable real-valued function by $\mathcal{Q}_{\mathbf{x}}g(\mathbf{h}) = q_{\mathbf{x}}(\mathbf{h}) \sum_{\mathbf{z} \in \mathbb{S}} g(\mathbf{z})\mathcal{R}_{\mathbf{x}}(\mathbf{h}, \mathbf{z}) - q_{\mathbf{x}}(\mathbf{h})g(\mathbf{h})$. The operator $\mathcal{Q}_{\mathbf{x}}$ represents the infinitesimal generator of \mathbb{S} -valued jump process \mathbf{H} , for fixed $\mathbf{x} \in \mathbb{R}^L$. Let $\mathcal{P}_{\mathbf{x}}(t, \mathbf{h}, \mathcal{A})$ be the transition probability function of \mathbf{H} for each $\mathbf{x} \in \mathbb{R}^L$. Assume that there exists a unique invariant distribution $\pi_{\mathbf{x}} = \{\pi_{\mathbf{x}}(\mathbf{h}), \mathbf{h} \in \mathbb{S}\}$ for the process \mathbf{H} , i.e., a probability measure depending on $\mathbf{x} \in \mathbb{R}^L$ such that $\pi_{\mathbf{x}}(\mathcal{A}) = \sum_{\mathbf{h}} \pi_{\mathbf{x}}(\mathbf{h})\mathcal{P}_{\mathbf{x}}(t, \mathbf{h}, \mathcal{A})$, $\mathcal{A} \in \mathcal{B}_{\mathbb{S}}, t \geq 0$. In other words, $\sum_{\mathbf{h}} \pi_{\mathbf{x}}(\mathbf{h})\mathcal{Q}_{\mathbf{x}}f(\mathbf{h}) = 0, \mathbf{x} \in \mathbb{R}^L$, for every $\mathcal{B}_{\mathbb{S}}$ -measurable bounded function f . Consequently, the recurrent potential kernel $\mathcal{D}_{\mathbf{x}}$ is defined as

$$\mathcal{D}_{\mathbf{x}}(\mathbf{h}, \mathcal{A}) = \int_0^\infty (\mathcal{P}_{\mathbf{x}}(t, \mathbf{h}, \mathcal{A}) - \pi_{\mathbf{x}}(\mathcal{A})) dt, \mathcal{A} \in \mathcal{B}_{\mathbb{S}}, \tag{5.2}$$

which is well defined for each $\mathbf{x} \in \mathbb{R}^L$. It follows from (5.2) that $\mathcal{P}_{\mathbf{x}}(t, \mathbf{h}, \mathcal{A})$ tends to $\pi_{\mathbf{x}}(\mathcal{A})$ sufficiently rapidly as $t \uparrow \infty$ for all $\mathbf{h} \in \mathbb{S}$, and $\mathcal{A} \in \mathcal{B}_{\mathbb{S}}$.

In order to establish the diffusion approximation of the joint process $(\mathbf{X}^N, \mathbf{H}^N)$ in $\mathbb{C}([0, \infty), \mathbb{R}^L) \times \mathbb{D}([0, \infty), \mathbb{R}^d)$, we require the following assumptions on the characteristic functions of the jump process \mathbf{H}^N as well as on the coefficients $\mathbf{A}^N(t, \mathbf{x}, \mathbf{h}), \mathbf{B}^N(t, \mathbf{x}, \mathbf{h})$ of (5.1).

Assumption 5.1. For every $\mathbf{x} \in \mathbb{R}^L$, the non-negative transition rate function $q_{\mathbf{x}}(\cdot)$ is bounded.

Assumption 5.2. The transition rate function $q_{\mathbf{x}}(\cdot)$, non-trivial probability measure $\mathcal{R}_{\mathbf{x}}(\cdot, \cdot)$, the transition function $\mathcal{P}_{\mathbf{x}}(t, \cdot, \cdot)$, the generating operator $\mathcal{Q}_{\mathbf{x}}$, the recurrent potential kernel $\mathcal{D}_{\mathbf{x}}(\cdot, \cdot)$, and the invariant probability measure $\pi_{\mathbf{x}}$ are assumed to be smoothly depending upon \mathbf{x} .

Assumption 5.3. The coefficients of (5.1), i.e., \mathbb{R}^L -valued functions $\mathbf{A}^N(t, \mathbf{x}, \mathbf{h})$, and $\mathbf{B}^N(t, \mathbf{x}, \mathbf{h})$ are jointly measurable on every $(t, \mathbf{x}, \mathbf{h}) \in [0, \infty) \times \mathbb{R}^L \times \mathbb{S}$.

Assumption 5.4. For every $t \in [0, \infty)$, $\mathbf{h} \in \mathbb{S}$, and $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^L$, there exists a constant K_{Lip} , independent of t, \mathbf{h} , such that

$$\|\mathbf{A}^N(t, \mathbf{x}, \mathbf{h}) - \mathbf{A}^N(t, \mathbf{x}', \mathbf{h})\| + \|\mathbf{B}^N(t, \mathbf{x}, \mathbf{h}) - \mathbf{B}^N(t, \mathbf{x}', \mathbf{h})\| \leq K_{Lip}\|\mathbf{x} - \mathbf{x}'\|.$$

Assumption 5.5. For every $(t, \mathbf{x}, \mathbf{h}) \in [0, \infty) \times \mathbb{R}^L \times \mathbb{S}$, there exists K_{growth} such that

$$\|\mathbf{A}^N(t, \mathbf{x}, \mathbf{h})\| + \|\mathbf{B}^N(t, \mathbf{x}, \mathbf{h})\| \leq K_{growth}(1 + \|\mathbf{x}\|).$$

Assumption 5.6. For every $i, j, k = 1, 2, \dots, L$, there exist some constants K_1, K_2 ,

$$\begin{aligned} \left| \frac{\partial}{\partial x_i} A_k^N(t, \mathbf{x}, \mathbf{h}) \right| &\leq K_1, \quad \left| \frac{\partial^2}{\partial x_j \partial x_j} A_k^N(t, \mathbf{x}, \mathbf{h}) \right| \leq K_1, \quad \left| \frac{\partial}{\partial x_i} B_k^N(t, \mathbf{x}, \mathbf{h}) \right| \leq K_1, \\ \left| \frac{\partial}{\partial t} A_k^N(t, \mathbf{x}, \mathbf{h}) \right| &< K_2, \quad \left| \frac{\partial}{\partial t} B_k^N(t, \mathbf{x}, \mathbf{h}) \right| < K_2. \end{aligned}$$

Assumption 5.7. There exist \mathbb{R}^L -valued Lipschitz continuous functions $\mathbf{A}(t, \mathbf{x}, \mathbf{h})$ and $\mathbf{B}(t, \mathbf{x}, \mathbf{h})$ defined on $[0, \infty] \times \mathbb{R}^L \times \mathbb{S}$ such that

$$\|\mathbf{A}^N(t, \mathbf{x}, \mathbf{h}) - \mathbf{A}(t, \mathbf{x}, \mathbf{h})\| \rightarrow 0, \quad \|\mathbf{B}^N(t, \mathbf{x}, \mathbf{h}) - \mathbf{B}(t, \mathbf{x}, \mathbf{h})\| \rightarrow 0, \quad \text{as } N \rightarrow \infty,$$

uniformly on $(t, \mathbf{x}, \mathbf{h})$ in compact subset of $[0, \infty) \times \mathbb{R}^L \times \mathbb{S}$.

Assumption 5.8. *In addition, we assume the centering condition on the function $\mathbf{A}^N(t, \mathbf{x}, \mathbf{h})$ with respect to the invariant measure $\pi_{\mathbf{x}}$, i.e.,*

$$\sum_{\mathbf{h}} \mathbf{A}^N(t, \mathbf{x}, \mathbf{h}) \pi_{\mathbf{x}}(\mathbf{h}) = \mathbf{0}, \quad t \geq 0, \quad \mathbf{x} \in \mathbb{R}^L.$$

Let $\mathcal{G}^N(t) = \sigma(\mathbf{X}^N(s), \mathbf{H}^N(s), s \leq t)$ be the sigma algebra generated by the joint process $(\mathbf{X}^N, \mathbf{H}^N)$ and $\mathcal{G}^N = \{\mathcal{G}^N(t), t \geq 0\}$ be the associated filtration. The proof of Theorem 5.1 is heavily relied on the martingale approach. In particular, we construct an \mathcal{G}^N -martingale associated with the joint process $(\mathbf{X}^N, \mathbf{H}^N)$ through its generator using the Markov property. For each $N \in \mathbb{N}$, the non-homogeneous bounded operator \mathcal{L}^N generating the joint Markov process $(\mathbf{X}^N, \mathbf{H}^N)$ is of the following form

$$\mathcal{L}_t^N = N^\beta \mathcal{Q}_{\mathbf{x}} + N^{\beta/2} \mathbf{A}^N(t, \mathbf{x}, \mathbf{h}) \cdot \nabla_{\mathbf{x}} + \mathbf{B}^N(t, \mathbf{x}, \mathbf{h}) \cdot \nabla_{\mathbf{x}}, \quad t \geq 0, \tag{5.3}$$

where, for each $\mathbf{x} \in \mathbb{R}^d$, among the following forms, $\nabla_{\mathbf{x}}$ denotes the gradient, i.e.,

$$\mathbf{A}^N(t, \mathbf{x}, \mathbf{h}) \cdot \nabla_{\mathbf{x}} = \sum_{i=1}^L A_i^N(t, \mathbf{x}, \mathbf{h}) \frac{\partial}{\partial x_i}, \quad \mathbf{B}^N(t, \mathbf{x}, \mathbf{h}) \cdot \nabla_{\mathbf{x}} = \sum_{i=1}^L B_i^N(t, \mathbf{x}, \mathbf{h}) \frac{\partial}{\partial x_i}.$$

Under the given hypotheses, there is a unique probability measure $\mathbb{P}_{\mathbf{x}, \mathbf{h}}$ with the initial point (\mathbf{x}, \mathbf{h}) , such that

- (1) $\mathbb{P}_{\mathbf{x}, \mathbf{h}}(\mathbf{X}^N(0) = \mathbf{x}, \mathbf{H}^N(0) = \mathbf{h}) = 1$,
- (2) $f(t, \mathbf{X}^N(t), \mathbf{H}^N(t)) - f(0, \mathbf{x}, \mathbf{h}) - \int_0^t (\frac{\partial}{\partial s} + \mathcal{L}_s^N) f(s, \mathbf{X}^N(s), \mathbf{H}^N(s)) ds$ is an \mathcal{G}^N -local square-integrable martingale with respect to $\mathbb{P}_{\mathbf{x}, \mathbf{h}}$ for every bounded measurable function $f(t, \mathbf{x}, \mathbf{h})$.

Introduce the averaged operator $\bar{\mathcal{L}}_t^{\pi_{\mathbf{x}}}$ depending on t and \mathbf{x} , averaged by the invariant distribution $\pi_{\mathbf{x}}$ of \mathbf{H}^N process

$$\bar{\mathcal{L}}_t^{\pi_{\mathbf{x}}} = \frac{1}{2} \sum_{i=1}^L \sum_{j=1}^L \bar{A}_{ij}^{1, \pi_{\mathbf{x}}}(t, \mathbf{x}) \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^L \bar{A}_i^{2, \pi_{\mathbf{x}}}(t, \mathbf{x}) \frac{\partial}{\partial x_i} + \sum_{i=1}^L \bar{B}_i^{\pi_{\mathbf{x}}}(t, \mathbf{x}) \frac{\partial}{\partial x_i}, \tag{5.4}$$

where the coefficient functions $\bar{A}_{ij}^{1, \pi_{\mathbf{x}}}(t, \mathbf{x})$, $\bar{A}_i^{2, \pi_{\mathbf{x}}}(t, \mathbf{x})$, $\bar{B}_i^{\pi_{\mathbf{x}}}(t, \mathbf{x})$ determined by the limiting functions $\mathbf{A}(t, \mathbf{x}, \mathbf{h})$ and $\mathbf{B}(t, \mathbf{x}, \mathbf{h})$ from Assumptions 5.7 are defined as follows:

$$\begin{aligned} \bar{A}_{ij}^{1, \pi_{\mathbf{x}}}(t, \mathbf{x}) &= \sum_{\mathbf{h}} \pi_{\mathbf{x}}(\mathbf{h}) \sum_z \mathcal{D}_{\mathbf{x}}(\mathbf{h}, \mathbf{z}) (A_i(t, \mathbf{x}, \mathbf{h}) A_j(t, \mathbf{x}, \mathbf{z}) + A_i(t, \mathbf{x}, \mathbf{z}) A_j(t, \mathbf{x}, \mathbf{h})), \\ \bar{A}_i^{2, \pi_{\mathbf{x}}}(t, \mathbf{x}) &= \sum_{\mathbf{h}} \pi_{\mathbf{x}}(\mathbf{h}) \sum_{l=1}^L A_j(t, \mathbf{x}, \mathbf{h}) \frac{\partial}{\partial x_j} \left(\sum_z \mathcal{D}_{\mathbf{x}}(\mathbf{h}, \mathbf{z}) A_i(t, \mathbf{x}, \mathbf{z}) \right), \\ \bar{B}_i^{\pi_{\mathbf{x}}}(t, \mathbf{x}) &= \sum_{\mathbf{h}} B_i(t, \mathbf{x}, \mathbf{h}) \pi_{\mathbf{x}}(\mathbf{h}), \quad i, j \in \mathbb{L}. \end{aligned} \tag{5.5}$$

To achieve the unique limiting process generated by the averaged operator $\bar{\mathcal{L}}_t^{\pi_{\mathbf{x}}}$, we require the Lipschitz property of the averaged coefficients $\bar{\mathbf{A}}^{1, \pi_{\mathbf{x}}}(t, \mathbf{x}, \mathbf{h})$, $\bar{\mathbf{A}}^{2, \pi_{\mathbf{x}}}(t, \mathbf{x}, \mathbf{h})$, $\bar{\mathbf{B}}^{\pi_{\mathbf{x}}}(t, \mathbf{x}, \mathbf{h})$, which simply follows from Assumption 5.4, 5.7.

Theorem 5.1. *Suppose the joint Markov process $(\mathbf{X}^N, \mathbf{H}^N)$ is a solution of (5.1) with initial conditions $\mathbf{X}^N(0) = \mathbf{x}, \mathbf{H}^N(0) = \mathbf{h}$. Let Assumptions 5.1–5.8 hold. Then \mathbf{X}^N converges in distribution to $\{\mathbf{X}(t), t \geq 0\}$ in $\mathbb{C}([0, \infty), \mathbb{R}^L)$ generated by the averaged operator $\bar{\mathcal{L}}_t^{\pi_{\mathbf{x}}}$ given by (5.4) as $N \rightarrow \infty$. Moreover, the diffusion Markov process \mathbf{X} uniquely satisfies the stochastic differential equation*

$$d\mathbf{X}(t) = \bar{\mathbf{A}}_*^{1, \pi_{\mathbf{x}}}(t, \mathbf{X}(t)) d\mathbf{W}(t) + \bar{\mathbf{A}}^{2, \pi_{\mathbf{x}}}(t, \mathbf{X}(t)) dt + \bar{\mathbf{B}}^{\pi_{\mathbf{x}}}(t, \mathbf{X}(t)) dt, \tag{5.6}$$

with initial condition $\mathbf{X}(0) = \mathbf{x}$, where $\bar{\mathbf{A}}^{1, \pi_{\mathbf{x}}}(\mathbf{x}) = \bar{\mathbf{A}}_*^{1, \pi_{\mathbf{x}}}(\mathbf{x}) \bar{\mathbf{A}}_*^{1, \pi_{\mathbf{x}}}(\mathbf{x})^T$, and $\{\mathbf{W}(t), t \geq 0\}$ is an independent L -dimensional standard Brownian motion.

Remark 5.2. Due to scaling up the transition rate of the jump process \mathbf{H}^N by N^β factor, \mathbf{H}^N is often referred to as *rapidly varying component* of the joint process $(\mathbf{X}^N, \mathbf{H}^N)$. On the contrary, after imposing the centering Assumption (5.8) on $\mathbf{A}^N(t, \mathbf{x}, \mathbf{h})$, the rate of change of \mathbf{X}^N process roughly becomes $\mathcal{O}(1)$, and therefore \mathbf{X}^N is often referred to as the *slowly varying component* in the literature (see Budhiraja et al., 2018; Papanicolaou, 1975; Skorokhod, 1989). In (5.1) with \mathbf{H}^N being rapidly-varying component, Theorem 5.1 actually proves that the diffusion approximation of \mathbf{X}^N is determined by the evolution of \mathbf{H}^N .

Remark 5.3. In particular, when $\mathbf{A}^N(t, \mathbf{x}, \mathbf{h}) = 0$, the differential equation (5.1) simply becomes the Itô equation, and the corresponding limiting process with $\bar{\mathbf{A}}^{1, \pi_{\mathbf{x}}}(t, \mathbf{x}) = \bar{\mathbf{A}}^{2, \pi_{\mathbf{x}}}(t, \mathbf{x}) = 0$ is only determined by the averaged function $\bar{\mathbf{B}}^{\pi_{\mathbf{x}}}(t, \mathbf{x})$, which is eventually a consequence of Theorem II.8 of Skorokhod (1989). In such scenario, \mathbf{X}^N converges to a diffusion process $\{\mathbf{X}^0(t), t \geq 0\}$ as $N \rightarrow \infty$, which has deterministic trajectory on $\mathbb{C}([0, \infty), \mathbb{R}^L)$ as unique solution of the ordinary differential equation $d\mathbf{X}^0(t) = \bar{\mathbf{B}}^{\pi_{\mathbf{x}}}(t, \mathbf{X}^0(t)) dt$, $\mathbf{X}^0(0) = \mathbf{x}$.

Remark 5.4. Under certain assumptions, Theorem 5.1 states that the limiting process $\{\mathbf{X}(t), t \geq 0\}$ is a Itô diffusion process satisfying SDE (5.6) uniquely with averaged coefficients $\bar{\mathbf{A}}_*^{1, \pi_{\mathbf{x}}}(t, \mathbf{x})$, $\bar{\mathbf{A}}^{2, \pi_{\mathbf{x}}}(t, \mathbf{x})$, $\bar{\mathbf{B}}^{\pi_{\mathbf{x}}}(t, \mathbf{x})$, even though the original stochastic dynamics (5.1) is governed by \mathbf{H}^N . The averaged coefficients incorporate the asymptotic influence of the jump process on the limiting process \mathbf{X} and effectively capture both predictable and stochastic fluctuations.

Proof of Theorem 5.1: The complete proof of weak convergence of the diffusion component \mathbf{X}^N relies on establishing finite dimensional convergence and \mathbb{C} -tightness of the scaled process \mathbf{X}^N in $\mathbb{C}([0, \infty), \mathbb{R}^L)$. The primary step of proving the finite dimensional convergence is based upon the martingale problem, as the solution of a martingale problem represents its finite dimensional distributions, and the uniqueness of solution of martingale problem leads us to the required limiting process.

Let $g \in \mathbb{C}_c^\infty(\mathbb{R}^L, \mathbb{R})$ be a test function. Given the recurrent potential kernel $\mathcal{D}_{\mathbf{x}}(\cdot, \cdot)$ associated to the jump process \mathbf{H}^N , we define

$$\theta_g^N(t, \mathbf{x}, \mathbf{h}) = \sum_z \mathcal{D}_{\mathbf{x}}(\mathbf{h}, z) \mathbf{A}^N(t, \mathbf{x}, z) \cdot \nabla_{\mathbf{x}} g(\mathbf{x}). \tag{5.7}$$

For given $\mathbf{A}^N(t, \mathbf{x}, \mathbf{h})$, $\mathbf{B}^N(t, \mathbf{x}, \mathbf{h})$, we additionally define

$$\alpha_g^N(t, \mathbf{x}, \mathbf{h}) = \mathbf{A}^N(t, \mathbf{x}, \mathbf{h}) \cdot \nabla_{\mathbf{x}} \theta_g(t, \mathbf{x}, \mathbf{h}) + \mathbf{B}^N(t, \mathbf{x}, \mathbf{h}) \cdot \nabla_{\mathbf{x}} g(\mathbf{x}), \tag{5.8}$$

$$\phi_g^N(t, \mathbf{x}, \mathbf{h}) = \sum_z \mathcal{D}_{\mathbf{x}}(\mathbf{h}, z) \alpha_g^N(t, \mathbf{x}, z). \tag{5.9}$$

We define the averaged generator with respect to invariant measure π

$$\bar{\mathcal{L}}_t^{N, \pi_{\mathbf{x}}} g(\mathbf{x}) = \sum_{\mathbf{h}} \pi_{\mathbf{x}}(\mathbf{h}) \alpha_g^N(t, \mathbf{x}, \mathbf{h}). \tag{5.10}$$

in such way that $\|\bar{\mathcal{L}}_t^{N, \pi_{\mathbf{x}}} g - \bar{\mathcal{L}}_t^{\pi_{\mathbf{x}}} g\| \rightarrow 0$ as $N \rightarrow \infty$ by Assumption 5.7. It follows from Assumptions 5.2–5.6 that for each $N \in \mathbb{N}$ the entities $\theta_g^N(t, \mathbf{x}, \mathbf{h})$, $\alpha_g^N(t, \mathbf{x}, \mathbf{h})$, $\phi_g^N(t, \mathbf{x}, \mathbf{h})$ are bounded.

Given g , choose the sequence of functions $\{g^N, N \in \mathbb{N}\}$ defined by

$$g^N(t, \mathbf{x}, \mathbf{h}) = g(\mathbf{x}) + N^{-\beta/2} \theta_g^N(t, \mathbf{x}, \mathbf{h}) + N^{-\beta} \phi_g^N(t, \mathbf{x}, \mathbf{h}),$$

such that g^N converges to g uniformly on compact sets. In order to show the existence of solutions of the martingale problem for $\bar{\mathcal{L}}_t^{\pi_{\mathbf{x}}}$, we need to identify the weak limit of solutions of the approximating martingale problems for \mathcal{L}_t^N such that g^N converges to g and $\mathcal{L}_t^N g^N$ converges to $\bar{\mathcal{L}}_t^{\pi_{\mathbf{x}}} g$ uniformly on compact sets as $N \rightarrow \infty$ in view of Lemma 5.1 in Ethier and Kurtz (1986). In addition, the uniqueness of solution of the martingale problem for $\bar{\mathcal{L}}_t^{\pi_{\mathbf{x}}}$ follows from Theorem 4.4.1 of Ethier and

Kurtz (1986) as the averaged diffusion operator $\bar{\mathcal{L}}_t^{\pi_x}$ defined on $\mathbb{C}_c^\infty(\mathbb{R}^L, \mathbb{R})$ satisfies the necessary conditions of the same theorem as a consequence of Hille-Yoshida theorem (see Theorem 4.2.2 of Ethier and Kurtz, 1986).

Applying the generator defined in (5.3) on $g^N(t, \mathbf{x}, \mathbf{h})$, we obtain

$$\begin{aligned} \left(\frac{\partial}{\partial t} + \mathcal{L}_t^N\right) g^N(t, \mathbf{x}, \mathbf{h}) = & N^\beta \mathcal{Q}_x g(\mathbf{x}) + N^{\beta/2} \left(\mathcal{Q}_x \theta_g^N(t, \mathbf{x}, \mathbf{h}) + \mathbf{A}^N(t, \mathbf{x}, \mathbf{h}) \cdot \nabla_x g(\mathbf{x}) \right) + \frac{\partial}{\partial t} g(\mathbf{x}) \\ & + \mathcal{Q}_x \phi_g^N(t, \mathbf{x}, \mathbf{h}) + \mathbf{A}^N(t, \mathbf{x}, \mathbf{h}) \cdot \nabla_x \theta_g^N(t, \mathbf{x}, \mathbf{h}) \\ & + \mathbf{B}^N(t, \mathbf{x}, \mathbf{h}) \cdot \nabla_x g(\mathbf{x}) + \mathfrak{T}_g^N(t, \mathbf{x}, \mathbf{h}), \end{aligned} \quad (5.11)$$

where

$$\begin{aligned} \mathfrak{T}_g^N(t, \mathbf{x}, \mathbf{h}) = & N^{-\beta/2} \left(\mathbf{A}^N(t, \mathbf{x}, \mathbf{h}) \cdot \nabla_x \phi_g^N(t, \mathbf{x}, \mathbf{h}) + \mathbf{B}^N(t, \mathbf{x}, \mathbf{h}) \cdot \nabla_x \theta_g^N(t, \mathbf{x}, \mathbf{h}) \right. \\ & \left. + \frac{\partial}{\partial t} \theta_g^N(t, \mathbf{x}, \mathbf{h}) \right) + N^{-\beta} \left(\mathbf{B}^N(t, \mathbf{x}, \mathbf{h}) \cdot \nabla_x \phi_g^N(t, \mathbf{x}, \mathbf{h}) + \frac{\partial}{\partial t} \phi_g^N(t, \mathbf{x}, \mathbf{h}) \right). \end{aligned} \quad (5.12)$$

For every $N \in \mathbb{N}$, the generator \mathcal{Q}_x acting on $\theta_g^N(t, \mathbf{x}, \mathbf{h})$ and $\phi_g^N(t, \mathbf{x}, \mathbf{h})$ results

$$\mathcal{Q}_x \theta_g^N(t, \mathbf{x}, \mathbf{h}) + \mathbf{A}^N(t, \mathbf{x}, \mathbf{h}) \cdot \nabla_x g(\mathbf{x}) = 0, \quad (5.13)$$

$$\mathcal{Q}_x \phi_g^N(t, \mathbf{x}, \mathbf{h}) + \mathbf{A}^N(t, \mathbf{x}, \mathbf{h}) \cdot \nabla_x \theta_g^N(t, \mathbf{x}, \mathbf{h}) + \mathbf{B}^N(t, \mathbf{x}, \mathbf{h}) \cdot \nabla_x g(\mathbf{x}) - \bar{\mathcal{L}}_t^{N, \pi_x} g(\mathbf{x}) = 0. \quad (5.14)$$

By the definition, the last equality (5.14) can be rewritten as

$$\mathcal{Q}_x \phi_g^N(t, \mathbf{x}, \mathbf{h}) + \alpha_g^N(t, \mathbf{x}, \mathbf{h}) - \bar{\mathcal{L}}_t^{N, \pi_x} g(\mathbf{x}) = 0.$$

Besides, Assumption 5.6 yields that $\mathfrak{T}_g^N(t, \mathbf{x}, \mathbf{h})$ associated with the derivatives of coefficients $\mathbf{A}^N(t, \mathbf{x}, \mathbf{h})$, $\mathbf{B}^N(t, \mathbf{x}, \mathbf{h})$ is bounded.

Define

$$M_{g^N}^N(t) = g^N(t, \mathbf{X}^N(t), \mathbf{H}^N(t)) - g^N(0, \mathbf{x}, \mathbf{h}) - \int_0^t \left(\frac{\partial}{\partial s} + \mathcal{L}_s^N \right) g^N(s, \mathbf{X}^N(s), \mathbf{H}^N(s)) ds. \quad (5.15)$$

It is clear that $M_{g^N}^N$ is an \mathcal{G}^N -square-integrable, bounded, zero-mean martingale, and the predictable quadratic variation process is given by

$$\left\langle M_{g^N}^N \right\rangle (t) = \int_0^t \Gamma_{g^N}^N(u, \mathbf{X}^N(u), \mathbf{H}^N(u)) du$$

where $\Gamma_{g^N}^N(t, \mathbf{x}, \mathbf{h}) = \mathcal{L}_t^N(g^N(t, \mathbf{x}, \mathbf{h}))^2 - 2g^N(t, \mathbf{x}, \mathbf{h})(\mathcal{L}_t^N g^N(t, \mathbf{x}, \mathbf{h}))$. After substituting the value of $g^N(t, \mathbf{x}, \mathbf{h})$ in (5.15), we obtain

$$\begin{aligned} M_{g^N}^N(t) = & g(\mathbf{X}^N(t)) - g(\mathbf{x}) + N^{-\beta/2} \left(\theta_g^N(t, \mathbf{X}^N(t), \mathbf{H}^N(t)) - \theta_g^N(0, \mathbf{x}, \mathbf{h}) \right) \\ & + N^{-\beta} \left(\phi_g^N(t, \mathbf{X}^N(t), \mathbf{H}^N(t)) - \phi_g^N(0, \mathbf{x}, \mathbf{h}) \right) \\ & - \int_0^t \left(\mathcal{Q}_x \phi_g^N(s, \mathbf{X}^N(s), \mathbf{H}^N(s)) + \alpha_g^N(s, \mathbf{X}^N(s), \mathbf{H}^N(s)) \right. \\ & \left. + \mathfrak{T}_g^N(s, \mathbf{X}^N(s), \mathbf{H}^N(s)) \right) ds. \end{aligned} \quad (5.16)$$

Next, to show that \mathbf{X}^N is tight in $\mathbb{C}([0, \infty), \mathbb{R}^L)$ in view of Theorem 1.3.2 of Stroock and Varadhan (2006), it suffices to show that for any $T > 0$, and $\epsilon > 0$,

$$\lim_{\delta \rightarrow 0} \limsup_{N \rightarrow \infty} \mathbb{P} \left(\sup_{\substack{0 \leq s \leq t \leq T, \\ |t-s| < \delta}} \|\mathbf{X}^N(t) - \mathbf{X}^N(s)\| > \epsilon \right) = 0.$$

Take $g(\mathbf{x}) = x_i, i = 1, \dots, L$, and introduce the vector valued functions $\boldsymbol{\theta}_x^N(t, \mathbf{x}, \mathbf{h}) = (\theta_{x_i}^N(t, \mathbf{x}, \mathbf{h}))_{i=1}^L$, $\boldsymbol{\alpha}_x^N(t, \mathbf{x}, \mathbf{h}) = (\alpha_{x_i}^N(t, \mathbf{x}, \mathbf{h}))_{i=1}^L$, $\boldsymbol{\phi}_x^N(t, \mathbf{x}, \mathbf{h}) = (\phi_{x_i}^N(t, \mathbf{x}, \mathbf{h}))_{i=1}^L$, and $\boldsymbol{\Upsilon}_x^N(t, \mathbf{x}, \mathbf{h}) = (\Upsilon_{x_i}^N(t, \mathbf{x}, \mathbf{h}))_{i=1}^L$ accordingly. Also, consider the \mathbb{R}^L -valued martingale $\{\mathbf{M}_x^N(t) = (M_{x_i}^N(t))_{i=1}^L, t \geq 0\}$, standing for the martingale \mathbf{M}_g^N corresponding to the vector valued function $\mathbf{g}^N(t, \mathbf{x}, \mathbf{h}) = (g_{x_i}^N(t, \mathbf{x}, \mathbf{h}))_{i=1}^L$. Each component of the associated $\mathbb{M}^{L \times L}$ -valued predictable variation process $\langle \mathbf{M}_x^N \rangle$ is given by

$$\langle M_{x_i}^N, M_{x_j}^N \rangle(t) = \int_0^t \Gamma_{g_{x_i}, g_{x_j}}^N(u, \mathbf{X}^N(u), \mathbf{H}^N(u)) du, \quad 1 \leq i, j \leq L,$$

where $\Gamma_{g_{x_i}, g_{x_j}}^N(t, \mathbf{x}, \mathbf{h}) = \mathcal{L}_t^N(g_{x_i}^N \cdot g_{x_j}^N)(t, \mathbf{x}, \mathbf{h}) - g_{x_i}^N(t, \mathbf{x}, \mathbf{h}) \mathcal{L}_t^N(g_{x_j}^N)(t, \mathbf{x}, \mathbf{h}) - g_{x_j}^N(t, \mathbf{x}, \mathbf{h}) \mathcal{L}_t^N(g_{x_i}^N)(t, \mathbf{x}, \mathbf{h})$.

Despite the unbounded choice of g , the corresponding functions $\theta_g^N(t, \mathbf{x}, \mathbf{h})$, $\alpha_g^N(t, \mathbf{x}, \mathbf{h})$, $\phi_g^N(t, \mathbf{x}, \mathbf{h})$, and $\Upsilon_g^N(t, \mathbf{x}, \mathbf{h})$ remain bounded, as they depend on the derivatives of g . Therefore, using the expression (5.16) we get

$$\begin{aligned} & \mathbb{P} \left(\sup_{\substack{0 \leq s \leq t \leq T, \\ |t-s| < \delta}} \|\mathbf{X}^N(t) - \mathbf{X}^N(s)\| > 4\epsilon \right) \\ & \leq \mathbb{P} \left(\sup_{\substack{0 \leq s \leq t \leq T, \\ |t-s| < \delta}} \|\mathbf{M}_x^N(t) - \mathbf{M}_x^N(s)\| > \epsilon \right) \\ & \quad + \mathbb{P} \left(\sup_{\substack{0 \leq s \leq t \leq T, \\ |t-s| < \delta}} \|N^{-\beta/2}(\boldsymbol{\theta}_x^N(t, \mathbf{X}^N(t), \mathbf{H}^N(t)) - \boldsymbol{\theta}_x^N(s, \mathbf{X}^N(s), \mathbf{H}^N(s)))\| > \epsilon \right) \\ & \quad + \mathbb{P} \left(\sup_{\substack{0 \leq s \leq t \leq T, \\ |t-s| < \delta}} \|N^{-\beta}(\boldsymbol{\phi}_x^N(t, \mathbf{X}^N(t), \mathbf{H}^N(t)) - \boldsymbol{\phi}_x^N(s, \mathbf{X}^N(s), \mathbf{H}^N(s)))\| > \epsilon \right) \\ & \quad + \mathbb{P} \left(\sup_{\substack{0 \leq s \leq t \leq T, \\ |t-s| < \delta}} \left\| \int_s^t \left(\mathcal{Q}_x \boldsymbol{\phi}_x^N(u, \mathbf{X}^N(u), \mathbf{H}^N(u)) + \boldsymbol{\alpha}_x^N(u, \mathbf{X}^N(u), \mathbf{H}^N(u)) \right. \right. \right. \\ & \quad \left. \left. \left. + \boldsymbol{\Upsilon}_x^N(u, \mathbf{X}^N(u), \mathbf{H}^N(u)) \right) du \right\| > \epsilon \right). \end{aligned} \tag{5.17}$$

For the first probability term of (5.17) associated with the martingale \mathbf{M}_x^N , the Lenglart-Rebolledo inequality (see Theorem 1.9.3 in [Liptser and Shiryaev, 1989](#)) yields

$$\begin{aligned} \mathbb{P} \left(\sup_{\substack{0 \leq s \leq t \leq T, \\ |t-s| < \delta}} \|\mathbf{M}_x^N(t) - \mathbf{M}_x^N(s)\| > \epsilon \right) & \leq \sum_{i=1}^L \mathbb{P} \left(\sup_{\substack{0 \leq s \leq t \leq T, \\ |t-s| < \delta}} |M_{x_i}^N(t) - M_{x_i}^N(s)| > \frac{\epsilon}{L} \right) \\ & \leq \frac{\eta L^2}{\epsilon^2} + \sum_{i=1}^L \mathbb{P} \left(\langle M_{x_i}^N \rangle(T + \delta) - \langle M_{x_i}^N \rangle(T) > \eta \right) \\ & \leq \frac{\eta L^2}{\epsilon^2} + \sum_{i=1}^L \mathbb{P} \left(\int_T^{T+\delta} \Gamma_{g_{x_i}}^N(u, \mathbf{X}^N(u), \mathbf{H}^N(u)) du > \eta \right). \end{aligned} \tag{5.18}$$

By Theorem 9.6.1 of [Liptser and Shiryaev \(1989\)](#), it can be shown that the latter term appearing in the right hand side of (5.18) goes to zero as $\delta \rightarrow 0$. Also, the convergence of the remaining entities in (5.17) directly follows from the boundedness of $\boldsymbol{\theta}_x^N(t, \mathbf{x}, \mathbf{h})$, $\boldsymbol{\alpha}_x^N(t, \mathbf{x}, \mathbf{h})$, $\boldsymbol{\phi}_x^N(t, \mathbf{x}, \mathbf{h})$, $\mathcal{Q}_x \boldsymbol{\phi}_x^N(t, \mathbf{x}, \mathbf{h})$, $\boldsymbol{\Upsilon}_x^N(t, \mathbf{x}, \mathbf{h})$, and therefore concludes the tightness criterion of \mathbf{X}^N , and hence concludes the weak convergence.

To identify the limit as a solution of the martingale problem for $\overline{\mathcal{L}}_t^{\pi_x}$, we consider

$$\begin{aligned} &g(\mathbf{X}^N(t)) - g(\mathbf{X}^N(s)) - \int_s^t \overline{\mathcal{L}}_u^{\pi_x} g(\mathbf{X}^N(u)) du \\ &= M_{g^N}^N(t) - M_{g^N}^N(s) - N^{-\beta/2} (\theta_g^N(t, \mathbf{X}^N(t), \mathbf{H}^N(t)) - \theta_g^N(s, \mathbf{X}^N(s), \mathbf{H}^N(s))) \\ &\quad - N^{-\beta} \left(\phi_g^N(t, \mathbf{X}^N(t), \mathbf{H}^N(t)) - \phi_g^N(s, \mathbf{X}^N(s), \mathbf{H}^N(s)) \right) + \int_s^t \left(\mathcal{Q}_x \phi_g^N(u, \mathbf{X}^N(u), \mathbf{H}^N(u)) \right. \\ &\quad \left. + \alpha_g^N(u, \mathbf{X}^N(u), \mathbf{H}^N(u)) + \Upsilon_g^N(u, \mathbf{X}^N(u), \mathbf{H}^N(u)) - \overline{\mathcal{L}}_u^{\pi_x} g(\mathbf{X}^N(u)) \right) du. \end{aligned}$$

Our aim is to show that

$$\limsup_{N \rightarrow \infty} \mathbb{E}_{(\mathbf{x}, \mathbf{h})} \left(g(\mathbf{X}^N(t)) - g(\mathbf{X}^N(s)) - \int_s^t \overline{\mathcal{L}}_u^{\pi_x} g(\mathbf{X}^N(u)) du \mid \mathcal{G}^N(s) \right) = 0,$$

which essentially follows from the boundedness of $\theta_g^N(t, \mathbf{x}, \mathbf{h})$, $\alpha_g^N(t, \mathbf{x}, \mathbf{h})$, $\phi_g^N(t, \mathbf{x}, \mathbf{h})$, $\mathcal{Q}_x \phi_g^N(t, \mathbf{x}, \mathbf{h})$, $\Upsilon_g^N(t, \mathbf{x}, \mathbf{h})$, and definition (5.10), and the condition (5.14), and martingale property of $M_{g^N}^N$ with respect to \mathcal{G}^N . As a consequence of the continuous mapping theorem and tightness, any limiting process of \mathbf{X}^N is a solution of the martingale problem for $\overline{\mathcal{L}}_t^{\pi_x}$ in $\mathbb{C}([0, \infty), \mathbb{R}^L)$ for every $g \in \mathbb{C}_c^\infty(\mathbb{R}^L, \mathbb{R})$. Thus, the uniqueness of solution of the martingale problem concludes the finite-dimensional convergence. Moreover, such uniqueness of solution yields the Markov property of the limiting process $\{\mathbf{X}(t), t \geq 0\}$, the unique solution of the martingale problem for $\overline{\mathcal{L}}_t^{\pi_x}$. \square

Remark 5.5. One can adapt an alternative approach via Theorem 2.1 of Kurtz (1992) to deduce the Gaussian approximation of \mathbf{X}^N or $\widehat{\mathbf{Z}}_{f^N}^N$ (in which we are particularly interested). This approach is primarily relied on the martingale problem for \mathcal{L}_t^N defined in (5.3) showing that $\exists p > 1$ such that for any $T > 0$, $\mathbb{E}(\int_0^T |(\frac{\partial}{\partial s} + \mathcal{L}_s^N)g^N(s, \mathbf{X}^N(s), \mathbf{H}^N(s))|^p ds) < \infty$ and the expected value of the corrector term consisting with $\theta_g^N(t, \mathbf{x}, \mathbf{h})$ and $\phi_g^N(t, \mathbf{x}, \mathbf{h})$ in the martingale formulation (5.16) vanishes to zero, and the tightness of the associated occupation measures. The latter tightness condition can be established in a straightforward manner under the present formulation defined in (5.19) and (5.20), together with Assumptions 3.3, 3.4. Nonetheless, to verify the former conditions, it is necessary to analyze the generator \mathcal{L}_t^N and the martingale M_g^N in (5.15) derived in the above proof, which play the key roles in our adapted martingale approach to establish the tightness and the identification of the limit process.

5.2. *Asymptotic approximations for stochastic integral processes derived by external process.* This particular section deals with certain integral processes determined by the external jump process \mathbf{H}^N and their convergences, which arise in the main proofs of Theorems 3.2 and 3.4. We discuss the sufficient conditions to establish the Gaussian approximations of the stochastic integral processes with respect to semimartingales (5.19), (5.20) by adding the classical concept of P-UT criterion (see Definition A.1 in Appendix A). In the end, Theorem 5.8 provides the necessary requirements to derive the Gaussian approximations of such integral processes. The detailed proofs of this section are presented in Appendix A.

For every $i \in \mathbb{L}$, the Lebesgue-Stieltjes integral processes defined in (4.1) are centered with respect to the invariant measure π in the following manner:

$$\begin{aligned} \overline{Z}_{\frac{1}{N}\lambda_i^N}^N(t) &= \int_0^t \frac{1}{N} (\lambda_i^N(s))^T (\mathbf{H}^N(s) - \pi) ds, \quad \overline{Z}_{\gamma_i^N}^N(t) = \int_0^t (\gamma_i^N(s))^T (\mathbf{H}^N(s) - \pi) ds, \\ \overline{Z}_{\mu_i^N}^N(t) &= \int_0^t (\mu_i^N(s))^T (\mathbf{H}^N(s) - \pi) ds, \quad i = 1, 2, \dots, L, \end{aligned} \tag{5.19}$$

and the associated \mathbb{R}^L -valued centered integral processes are $\{\overline{\mathbf{Z}}_{\frac{1}{N}\lambda^N}^N(t), t \geq 0\}$, $\{\overline{\mathbf{Z}}_{\gamma^N}^N(t), t \geq 0\}$, $\{\overline{\mathbf{Z}}_{\mu^N}^N(t), t \geq 0\}$ with $\overline{\mathbf{Z}}_{\mathbf{f}^N}^N(0) = \mathbf{0}$, $\mathbf{f}^N(\cdot) = (\frac{1}{N}\lambda_i^N(\cdot))_{i=1}^L, (\gamma_i^N(\cdot))_{i=1}^L, (\mu_i^N(\cdot))_{i=1}^L$.

Introducing the parameter β , we define the diffusion scaled processes as

$$\begin{aligned} \widehat{\mathbf{Z}}_{\frac{1}{N}\lambda_i^N}^N(t) &= N^{\beta/2} \int_0^t \frac{1}{N} (\lambda_i^N(s))^T (\mathbf{H}^N(s) - \boldsymbol{\pi}) \, ds, \widehat{\mathbf{Z}}_{\gamma_i^N}^N(t) = N^{\beta/2} \int_0^t (\gamma_i^N(s))^T (\mathbf{H}^N(s) - \boldsymbol{\pi}) \, ds \\ \widehat{\mathbf{Z}}_{\mu_i^N}^N(t) &= N^{\beta/2} \int_0^t (\mu_i^N(s))^T (\mathbf{H}^N(s) - \boldsymbol{\pi}) \, ds, \beta > 0, i = 1, 2, \dots, L, \end{aligned} \tag{5.20}$$

and the corresponding \mathbb{R}^L -valued diffusion scaled processes are $\{\widehat{\mathbf{Z}}_{\frac{1}{N}\lambda^N}^N(t), t \geq 0\}$, $\{\widehat{\mathbf{Z}}_{\gamma^N}^N(t), t \geq 0\}$, $\{\widehat{\mathbf{Z}}_{\mu^N}^N(t), t \geq 0\}$ with $\widehat{\mathbf{Z}}_{\mathbf{f}^N}^N(0) = \mathbf{0}$, $\mathbf{f}^N(\cdot) = (\frac{1}{N}\lambda_i^N(\cdot))_{i=1}^L, (\gamma_i^N(\cdot))_{i=1}^L, (\mu_i^N(\cdot))_{i=1}^L$. For convenience, and with a slight abuse of notation, we denote $\overline{\mathbf{Z}}_{\frac{1}{N}\lambda^N}^N$ and $\widehat{\mathbf{Z}}_{\frac{1}{N}\lambda^N}^N$ by $\overline{\mathbf{Z}}_{\lambda^N}^N$ and $\widehat{\mathbf{Z}}_{\lambda^N}^N$, respectively.

In the following theorem, we explicitly derive the stochastic convergence of the centered and diffusion scaled processes $\overline{\mathbf{Z}}_{\mathbf{f}^N}^N, \widehat{\mathbf{Z}}_{\mathbf{f}^N}^N, \mathbf{f}^N(\cdot) = (\lambda_i^N(\cdot))_{i=1}^L, (\gamma_i^N(\cdot))_{i=1}^L, (\mu_i^N(\cdot))_{i=1}^L$.

Theorem 5.6. *Suppose that, for each $i \in \mathbb{L}$, the deterministic rate functions $\lambda_i^N(\cdot), \gamma_i^N(\cdot), \mu_i^N(\cdot)$ satisfy Assumption 3.2–3.3. Then the centered joint process $(\overline{\mathbf{Z}}_{\lambda^N}^N, \overline{\mathbf{Z}}_{\gamma^N}^N, \overline{\mathbf{Z}}_{\mu^N}^N)$ converges in probability to the trivial process (Ξ, Ξ, Ξ) uniformly on compact sets as $N \rightarrow \infty$, and the diffusion scaled joint process $(\widehat{\mathbf{Z}}_{\lambda^N}^N, \widehat{\mathbf{Z}}_{\gamma^N}^N, \widehat{\mathbf{Z}}_{\mu^N}^N)$ converges in distribution to*

$$(\{\mathbf{W}_\lambda(t), t \geq 0\}, \{\mathbf{W}_\gamma(t), t \geq 0\}, \{\mathbf{W}_\mu(t), t \geq 0\}) \quad \text{in } \mathbb{D}([0, \infty), \mathbb{R}^{3L}),$$

where $\mathbf{W}_\lambda, \mathbf{W}_\gamma, \mathbf{W}_\mu$ are correlated drift-less L -dimensional Brownian motions with covariances $\mathbf{C}_{\mathbf{f}\mathbf{g}}(t) = (C_{\mathbf{f}_i\mathbf{g}_j}(t))_{i,j=1}^L$ defined in (3.5), $\mathbf{f}_i(t) = \lambda_i(t), \gamma_i(t), \mu_i(t), \mathbf{g}_j(t) = \lambda_j(t), \gamma_j(t), \mu_j(t)$.

Remark 5.7. The diffusion approximation of the coupled process $(\widehat{\mathbf{Z}}_{\lambda^N}^N, \widehat{\mathbf{Z}}_{\gamma^N}^N, \widehat{\mathbf{Z}}_{\mu^N}^N)$ to the Gaussian martingale $(\mathbf{W}_\lambda, \mathbf{W}_\gamma, \mathbf{W}_\mu)$ is characterized by the covariance function

$$\begin{aligned} \langle \mathbf{W}_\lambda \rangle(t) &= \mathbf{C}_{\lambda\lambda}(t), \quad \langle \mathbf{W}_\gamma \rangle(t) = \mathbf{C}_{\gamma\gamma}(t), \quad \langle \mathbf{W}_\mu \rangle(t) = \mathbf{C}_{\mu\mu}(t), \\ \langle \mathbf{W}_\lambda, \mathbf{W}_\gamma \rangle(t) &= \mathbf{C}_{\lambda\gamma}(t), \quad \langle \mathbf{W}_\lambda, \mathbf{W}_\mu \rangle(t) = \mathbf{C}_{\lambda\mu}(t), \quad \langle \mathbf{W}_\gamma, \mathbf{W}_\mu \rangle(t) = \mathbf{C}_{\gamma\mu}(t), \end{aligned}$$

given by (3.5).

Next, we are interested in establishing the Skorokhod convergences of an stochastic integral process of the form

$$\mathcal{K}_-^N \cdot \mathbf{Z}_{\mathbf{f}^N}^N(t) = \int_0^t \mathcal{K}^N(s-) \, d\mathbf{Z}_{\mathbf{f}^N}^N(s), \quad \mathbf{Z}_{\mathbf{f}^N}^N(t) = \overline{\mathbf{Z}}_{\mathbf{f}^N}^N(t), \widehat{\mathbf{Z}}_{\mathbf{f}^N}^N(t),$$

with $\mathbf{f}^N(\cdot) = (\lambda_i^N(\cdot))_{i=1}^L, (\gamma_i^N(\cdot))_{i=1}^L, (\mu_i^N(\cdot))_{i=1}^L$. Here $\{\mathcal{K}^N(t), t \geq 0\}$ is an \mathcal{G}^N -predictable locally bounded process, which has trajectories in Skorokhod space $\mathbb{D}([0, \infty), \mathbb{M}^{M \times L})$, and $\{\mathbf{Z}_{\mathbf{f}^N}^N(t), t \geq 0\}$ is an \mathcal{G}^N -adapted semimartingale in $\mathbb{D}([0, \infty), \mathbb{R}^L)$, such that the integral process $\{\mathcal{K}_-^N \cdot \mathbf{Z}_{\mathbf{f}^N}^N(t), t \geq 0\}$ is well-defined. It is noted that the convergence of the semimartingale $\mathbf{Z}_{\mathbf{f}^N}^N$ to the process $\{\mathbf{Z}_{\mathbf{f}}(t), t \geq 0\}$ (say) as $N \rightarrow \infty$ by virtue of Theorem 5.6 is not enough to ensure the Skorokhod convergence of the integral process $\mathcal{K}_-^N \cdot \mathbf{Z}_{\mathbf{f}^N}^N$ to $\mathcal{K}_- \cdot \mathbf{Z}_{\mathbf{f}}$ as $N \rightarrow \infty$ in $\mathbb{D}([0, \infty), \mathbb{R}^M)$. We need to straighten the condition on the semimartingale $\mathbf{Z}_{\mathbf{f}^N}^N$, which we refer to as *predictable uniformly tight (P-UT)* condition (see Appendix A). Despite having non-PUT feature of the sequence $\{\widehat{\mathbf{Z}}_{\mathbf{f}^N}^N, N \in \mathbb{N}\}$, the next theorem establishes the multi-dimensional joint convergence of the Lebesgue-Stieltjes integral process with respect to semimartingale $(\widehat{\mathbf{Z}}_{\lambda^N}^N, \widehat{\mathbf{Z}}_{\gamma^N}^N, \widehat{\mathbf{Z}}_{\mu^N}^N)$.

Theorem 5.8. For each $N \in \mathbb{N}$, let $\mathcal{K}_1^N, \mathcal{K}_2^N$ and \mathcal{K}_3^N be \mathcal{G}^N -adapted predictable processes with each \mathcal{K}_i^N having trajectories in $\mathbb{D}([0, \infty), \mathbb{M}^{M \times L})$, $i = 1, 2, 3$, and $(\widehat{\mathbf{Z}}_{\lambda^N}^N, \widehat{\mathbf{Z}}_{\gamma^N}^N, \widehat{\mathbf{Z}}_{\mu^N}^N)$ be the \mathcal{G}^N -adapted semimartingale in $\mathbb{C}([0, \infty), \mathbb{R}^{3L})$ satisfying the assumptions in Theorem 5.6. If $(\mathcal{K}_1^N, \mathcal{K}_2^N, \mathcal{K}_3^N, \widehat{\mathbf{Z}}_{\lambda^N}^N, \widehat{\mathbf{Z}}_{\gamma^N}^N, \widehat{\mathbf{Z}}_{\mu^N}^N)$ converges in distribution to $(\mathcal{K}_1, \mathcal{K}_2, \mathcal{K}_3, \mathbf{W}_\lambda, \mathbf{W}_\gamma, \mathbf{W}_\mu)$ in the Skorokhod topology on $\mathbb{D}([0, \infty), \mathbb{M}^{3M \times 3L}) \times \mathbb{C}([0, \infty), \mathbb{R}^{3L})$, then $(\mathcal{K}_{1-}^N \cdot \widehat{\mathbf{Z}}_{\lambda^N}^N, \mathcal{K}_{2-}^N \cdot \widehat{\mathbf{Z}}_{\gamma^N}^N, \mathcal{K}_{3-}^N \cdot \widehat{\mathbf{Z}}_{\mu^N}^N)$ converges in distribution to $(\mathcal{K}_{1-} \cdot \mathbf{W}_\lambda, \mathcal{K}_{2-} \cdot \mathbf{W}_\gamma, \mathcal{K}_{3-} \cdot \mathbf{W}_\mu)$ in the Skorokhod topology on $\mathbb{D}([0, \infty), \mathbb{R}^{3L})$ as $N \rightarrow \infty$.

Note that, the centered family $\overline{\mathbf{Z}}_{f^N}^N$ given $f^N(\cdot) = (\lambda_i^N(\cdot))_{i=1}^L, (\gamma_i^N(\cdot))_{i=1}^L, (\mu_i^N(\cdot))_{i=1}^L$ satisfy the required P-UT criterion to derive the Skorokhod convergence of $\mathcal{K}_-^N \cdot \overline{\mathbf{Z}}_{f^N}^N$, and hence the next corollary of Theorem 5.8 proves the convergence of the Lebesgue-Stieltjes stochastic integral family $\mathcal{K}_-^N \cdot \overline{\mathbf{Z}}_{f^N}^N$ with respect to $\overline{\mathbf{Z}}_{f^N}^N$.

Corollary 5.9. Under the assumptions of Theorem 5.8, if $(\mathcal{K}_1^N, \mathcal{K}_2^N, \mathcal{K}_3^N, \overline{\mathbf{Z}}_{\lambda^N}^N, \overline{\mathbf{Z}}_{\gamma^N}^N, \overline{\mathbf{Z}}_{\mu^N}^N)$ converges in probability to $(\mathcal{K}_1, \mathcal{K}_2, \mathcal{K}_3, \Xi, \Xi, \Xi)$ uniformly on compact sets with trajectories in $\mathbb{D}([0, \infty), \mathbb{M}^{3M \times 3L}) \times \mathbb{C}([0, \infty), \mathbb{R}^{3L})$, then $(\mathcal{K}_{1-}^N \cdot \overline{\mathbf{Z}}_{\lambda^N}^N, \mathcal{K}_{2-}^N \cdot \overline{\mathbf{Z}}_{\gamma^N}^N, \mathcal{K}_{3-}^N \cdot \overline{\mathbf{Z}}_{\mu^N}^N)$ converges in probability to $(\mathcal{K}_{1-} \cdot \Xi, \mathcal{K}_{2-} \cdot \Xi, \mathcal{K}_{3-} \cdot \Xi)$ (i.e., the trivial process (Ξ, Ξ, Ξ)) uniformly on compact sets in $\mathbb{D}([0, \infty), \mathbb{R}^{3M})$ as $N \rightarrow \infty$.

6. Continuity of multi-dimensional integral representation

In this section, we state the continuity of multi-dimensional integral representation (6.1) modulated by the external random environment \mathcal{K}^N within the two-time-scale framework, which is highly influenced by Pang et al. (2007). In the following sections, this continuity plays a crucial role in order to apply the continuous mapping theorem to conclude the assertions of Theorems 3.2 and 3.4.

Given $\mathbf{C} \in \mathbb{R}_+^L$, and the process $\mathbf{Y} \in \mathbb{D}([0, \infty), \mathbb{R}^L)$ and the functions $\mathbf{h}_1, \mathbf{h}_2 : \mathbb{R}^L \rightarrow \mathbb{R}^L$, we consider an integral representation governed by the external Markov process \mathbf{H} defined by

$$\begin{aligned} \mathbf{X}(t) = & \mathbf{C} + \mathbf{Y}(t) - \int_0^t \left(\mathbf{I} - \mathbf{P}_1(\mathbf{X}(s))^T \right) \text{diag}(\mathbf{h}_1(\mathbf{X}(s))) d\mathbf{Z}_\gamma(s) \\ & - \int_0^t \left(\mathbf{I} - \mathbf{P}_2(\mathbf{X}(s))^T \right) \text{diag}(\mathbf{h}_2(\mathbf{X}(s))) d\mathbf{Z}_\mu(s), \end{aligned} \quad (6.1)$$

where $\mathbf{P}_1, \mathbf{P}_2 : \mathbb{R}^L \rightarrow [0, 1]^{L \times L}$ are two substochastic matrices satisfying Lipschitz condition, i.e., there exists K_{Lip}^1, K_{Lip}^2 such that $\|\mathbf{P}_1(\mathbf{x}_1) - \mathbf{P}_1(\mathbf{x}_2)\| \leq K_{Lip}^1 \|\mathbf{x}_1 - \mathbf{x}_2\|$, $\|\mathbf{P}_2(\mathbf{x}_1) - \mathbf{P}_2(\mathbf{x}_2)\| \leq K_{Lip}^2 \|\mathbf{x}_1 - \mathbf{x}_2\|$, for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^L$. The integral processes in (6.1) are Lebesgue-Stieltjes processes with respect to semimartingales $\{\mathbf{Z}_\gamma(t), t \geq 0\}$ and $\{\mathbf{Z}_\mu(t), t \geq 0\}$ in $\mathbb{C}([0, \infty), \mathbb{R}^L)$ defined by

$$\mathbf{Z}_\gamma(t) = \int_0^t \mathbf{g}_\gamma(s, \mathbf{H}(s)) ds, \quad \mathbf{Z}_\mu(t) = \int_0^t \mathbf{g}_\mu(s, \mathbf{H}(s)) ds,$$

where $\mathbf{g}_\gamma, \mathbf{g}_\mu : \mathbb{R}_+ \times \mathbb{S} \rightarrow \mathbb{R}^L$ are two Borel measurable functions. In particular, for each $N \in \mathbb{N}$, if $\mathbf{g}_{\gamma^N}^N(t, \mathbf{h}) = (\gamma_i^N(t)^T \mathbf{h})_{i=1}^L$, and $\mathbf{g}_{\mu^N}^N(t, \mathbf{h}) = (\mu_i^N(t)^T \mathbf{h})_{i=1}^L$, $t \geq 0$, then it follows from Theorem 5.6 that the \mathbb{R}^L -valued scaled integral processes $\mathbf{Z}_{\gamma^N}^N$ and $\mathbf{Z}_{\mu^N}^N$ converge in probability to the deterministic processes $\int_0^\cdot \overline{\gamma}^\pi(s) ds$ and $\int_0^\cdot \overline{\mu}^\pi(s) ds$, respectively, uniformly on compact set as $N \rightarrow \infty$. If $\mathbf{g}_\gamma(t, \mathbf{h}) = (\gamma_i(t)^T \mathbf{h})_{i=1}^L$, $\mathbf{g}_\mu(t, \mathbf{h}) = (\mu_i(t)^T \mathbf{h})_{i=1}^L$, then

$$\lim_{N \rightarrow \infty} \|\mathbf{g}_{\gamma^N}^N - \mathbf{g}_\gamma\|_{\mathbb{S}, T} = 0, \quad \lim_{N \rightarrow \infty} \|\mathbf{g}_{\mu^N}^N - \mathbf{g}_\mu\|_{\mathbb{S}, T} = 0,$$

by Assumption 3.3, where $\|\mathbf{g}\|_{\mathbb{S}, T} = \sup_{0 \leq t \leq T} \|\mathbf{g}(t, \mathbf{h})\|$.

Theorem 6.1. *Assume that the functions $\mathbf{h}_1, \mathbf{h}_2$ are Lipschitz continuous with respect to the constants C_{Lip}^1 and C_{Lip}^2 , i.e. $\|\mathbf{h}_1(\mathbf{x}_1) - \mathbf{h}_1(\mathbf{x}_2)\| \leq C_{Lip}^1 \|\mathbf{x}_1 - \mathbf{x}_2\|$, $\|\mathbf{h}_2(\mathbf{x}_1) - \mathbf{h}_2(\mathbf{x}_2)\| \leq C_{Lip}^2 \|\mathbf{x}_1 - \mathbf{x}_2\|$, for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^L$. In addition $\|\mathbf{h}_1(\mathbf{x})\| \leq \|\mathbf{x}\|$, $\|\mathbf{h}_2(\mathbf{x})\| \leq \|\mathbf{x}\|$, for any $\mathbf{x} \in \mathbb{R}^L$, and $\mathbf{h}_1(0) = \mathbf{h}_2(0) = \mathbf{0}$. Then the integral representation characterized by*

$$\Phi : \mathbb{R}_+^L \times \mathbb{D}([0, \infty), \mathbb{R}^L) \times (\mathbb{M}^{L \times L})^2 \times \mathbb{D}([0, \infty), \mathbb{R}^L)^2 \rightarrow \mathbb{D}([0, \infty), \mathbb{R}^L)$$

mapping $(\mathbf{C}, \mathbf{Y}, \mathbf{P}_1, \mathbf{P}_2, \mathbf{Z}_\gamma, \mathbf{Z}_\mu)$ into $\mathbf{X} \equiv \Phi(\mathbf{C}, \mathbf{Y}, \mathbf{P}_1, \mathbf{P}_2, \mathbf{Z}_\gamma, \mathbf{Z}_\mu)$ has a unique solution. Moreover, the map Φ is continuous in $\mathbb{D}([0, \infty), \mathbb{R}^L)$ endowed with the topology of uniform convergence on compact sets or the Skorokhod \mathbb{J}_1 topology.

7. Proof of fluid approximation

In the beginning of the section, we represent the fluid scaled system length processes $\bar{\mathbf{Q}}^N$ in the following manner:

$$\begin{aligned} \bar{\mathbf{Q}}^N(t) = & \bar{\mathbf{Q}}^N(0) + \bar{\mathbf{M}}^{N,E}(t) + \left(\bar{\mathbf{M}}^{N,A}(t)^T - \bar{\mathbf{M}}^{N,A}(t)\right) \cdot \mathbf{1} + \left(\bar{\mathbf{M}}^{N,S}(t)^T - \bar{\mathbf{M}}^{N,S}(t)\right) \cdot \mathbf{1} \\ & - \bar{\mathbf{M}}_0^{N,A}(t) - \bar{\mathbf{M}}_0^{N,S}(t) + \bar{\mathbf{R}}^{N,E}(t) + \left(\bar{\mathbf{R}}^{N,A}(t)^T - \bar{\mathbf{R}}^{N,A}(t)\right) \cdot \mathbf{1} \\ & + \left(\bar{\mathbf{R}}^{N,S}(t)^T - \bar{\mathbf{R}}^{N,S}(t)\right) \cdot \mathbf{1} - \bar{\mathbf{R}}_0^{N,A}(t) - \bar{\mathbf{R}}_0^{N,S}(t), \end{aligned} \tag{7.1}$$

where the fluid scaled martingales are defined by $\bar{M}_i^{N,E}(t) = \frac{M_i^{N,E}(t)}{N}$, $\bar{M}_{ij}^{N,A}(t) = \frac{M_{ij}^{N,A}(t)}{N}$, $\bar{M}_{ij}^{N,S}(t) = \frac{M_{ij}^{N,S}(t)}{N}$, $i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\}$ and the fluid scaled Lebesgue-Stieltjes integral processes are defined by

$$\begin{aligned} \bar{R}_i^{N,E}(t) &= \int_0^t \frac{1}{N} dZ_{\lambda_i^N}^N(s), \quad i \in \mathbb{L}, \quad \bar{R}_{ij}^{N,A}(t) = \int_0^t \left(\bar{Q}_i^N(s) - \bar{K}_i^N(s)\right)^+ \bar{\psi}_{ij}^N(\bar{Q}_i^N(s-)) dZ_{\gamma_i^N}^N(s), \\ \bar{R}_{ij}^{N,S}(t) &= \int_0^t \left(\bar{Q}_i^N(s) \wedge \bar{K}_i^N(s)\right) \bar{\phi}_{ij}^N(\bar{Q}_i^N(s-)) dZ_{\mu_{ij}^N}^N(s), \quad i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\}. \end{aligned} \tag{7.2}$$

In addition, we introduce the processes

$$\begin{aligned} \bar{R}_i^E(t) &= \int_0^t \bar{\lambda}_i^\pi(s) ds, \quad i \in \mathbb{L}, \quad \bar{R}_{ij}^A(t) = \int_0^t \left(\bar{Q}_i(s) - \bar{K}_i(s)\right)^+ \bar{\psi}_{ij}(\bar{Q}_i(s-)) \bar{\gamma}_i^\pi(s) ds, \\ \bar{R}_{ij}^S(t) &= \int_0^t \left(\bar{Q}_i(s) \wedge \bar{K}_i(s)\right) \bar{\phi}_{ij}(\bar{Q}_i(s-)) \bar{\mu}_{ij}^\pi(s) ds, \quad i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\}, \end{aligned} \tag{7.3}$$

so that we can express the \mathbb{R}^L -valued fluid limiting process $\bar{\mathbf{Q}}$ of Theorem 3.2 in the following manner

$$\bar{\mathbf{Q}}(t) = \bar{\mathbf{Q}}(0) + \int_0^t \bar{\boldsymbol{\lambda}}^\pi(s) ds + \left(\bar{\mathbf{R}}^A(t)^T - \bar{\mathbf{R}}^A(t)\right) \cdot \mathbf{1} + \left(\bar{\mathbf{R}}^S(t)^T - \bar{\mathbf{R}}^S(t)\right) \cdot \mathbf{1} - \bar{\mathbf{R}}_0^A(t) - \bar{\mathbf{R}}_0^S(t). \tag{7.4}$$

As a consequence of Theorem 5.6 and Assumption 3.3, the deterministic approximations of the integral processes of $\{\mathbf{Z}_{\lambda^N}^N(t), t \geq 0\}$, $\{\mathbf{Z}_{\gamma^N}^N(t), t \geq 0\}$ and $\{\mathbf{Z}_{\mu^N}^N(t), t \geq 0\}$ are given by

$$\left\{ \int_0^t \bar{\boldsymbol{\lambda}}^\pi(s) ds, t \geq 0 \right\}, \quad \left\{ \int_0^t \bar{\boldsymbol{\gamma}}^\pi(s) ds, t \geq 0 \right\}, \quad \left\{ \int_0^t \bar{\boldsymbol{\mu}}^\pi(s) ds, t \geq 0 \right\},$$

respectively. Before proceeding to prove the main Theorem 3.2, we establish the trivial convergences of the fluid scaled martingales $\bar{M}_i^{N,E}, \bar{M}_{ij}^{N,A}, \bar{M}_{ij}^{N,S}, i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\}$ in the next lemma.

Lemma 7.1. *Under Assumption 3.2–3.3, Assumption 3.5, for all $T > 0$ and $\epsilon > 0$ the following convergences hold in $\mathbb{D}([0, \infty, \mathbb{R})$.*

$$\limsup_{N \rightarrow \infty} \mathbb{P} \left(\sup_{0 \leq t \leq T} \left\| \overline{\mathbf{M}}^{N,E}(t) \right\| > \epsilon \right) = 0, \quad \limsup_{N \rightarrow \infty} \mathbb{P} \left(\sup_{0 \leq t \leq T} \left\| \overline{\mathbf{M}}^{N,I}(t) \right\| > \epsilon \right) = 0, \quad I = A, S.$$

Proof of Theorem 3.2: For every $i \in \mathbb{L}$ and $j \in \mathbb{L} \cup \{0\}$, consider the differences

$$\begin{aligned} \overline{R}_{ij}^{N,A}(t) - \overline{R}_{ij}^A(t) &= \int_0^t \left(\left(\overline{Q}_i^N(s) - \overline{K}_i^N(s) \right)^+ - \left(\overline{Q}_i(s) - \overline{K}_i(s) \right)^+ \right) \overline{\psi}_{ij}^N \left(\overline{Q}_i^N(s-) \right) dZ_{\gamma_i^N}^N(s) \\ &\quad + \int_0^t \left(\overline{Q}_i(s) - \overline{K}_i(s) \right)^+ \left(\overline{\psi}_{ij}^N \left(\overline{Q}_i^N(s-) \right) - \overline{\psi}_{ij} \left(\overline{Q}_i(s-) \right) \right) dZ_{\gamma_i^N}^N(s) \\ &\quad + \int_0^t \left(\overline{Q}_i(s) - \overline{K}_i(s) \right)^+ \left(\overline{\psi}_{ij} \left(\overline{Q}_i^N(s-) \right) - \overline{\psi}_{ij} \left(\overline{Q}_i(s-) \right) \right) dZ_{\gamma_i^N}^N(s) \\ &\quad + \int_0^t \left(\overline{Q}_i(s) - \overline{K}_i(s) \right)^+ \overline{\psi}_{ij} \left(\overline{Q}_i(s-) \right) \left(dZ_{\gamma_i^N}^N(s) - \overline{\gamma}_i^\pi(s) ds \right). \end{aligned} \quad (7.5)$$

and

$$\begin{aligned} \overline{R}_{ij}^{N,S}(t) - \overline{R}_{ij}^S(t) &= \int_0^t \left(\overline{Q}_i^N(s) \wedge \overline{K}_i^N(s) - \overline{Q}_i(s) \wedge \overline{K}_i(s) \right) \overline{\phi}_{ij}^N \left(\overline{Q}_i^N(s-) \right) dZ_{\mu_i^N}^N(s) \\ &\quad + \int_0^t \left(\overline{Q}_i(s) \wedge \overline{K}_i(s) \right) \left(\overline{\phi}_{ij}^N \left(\overline{Q}_i^N(s-) \right) - \overline{\phi}_{ij} \left(\overline{Q}_i(s-) \right) \right) dZ_{\mu_i^N}^N(s) \\ &\quad + \int_0^t \left(\overline{Q}_i(s) \wedge \overline{K}_i(s) \right) \left(\overline{\phi}_{ij} \left(\overline{Q}_i^N(s-) \right) - \overline{\phi}_{ij} \left(\overline{Q}_i(s-) \right) \right) dZ_{\mu_i^N}^N(s) \\ &\quad + \int_0^t \left(\overline{Q}_i(s) \wedge \overline{K}_i(s) \right) \overline{\phi}_{ij} \left(\overline{Q}_i(s-) \right) \left(dZ_{\mu_i^N}^N(s) - \overline{\mu}_i^\pi ds \right). \end{aligned} \quad (7.6)$$

By (7.1), (7.4), and the decomposition representations (7.5), (7.6), we finally get

$$\begin{aligned} \overline{Q}^N(t) - \overline{Q}(t) &= \overline{Q}^N(0) - \overline{Q}(0) + \overline{Y}^N(t) \\ &\quad + \int_0^t \left(\overline{\psi} \left(\overline{Q}^N(s-) \right)^T - \overline{\psi} \left(\overline{Q}(s-) \right)^T \right) \mathbf{diag} \left(\overline{Q}(s) - \overline{K}(s) \right)^+ dZ_{\gamma^N}^N(s) \\ &\quad + \int_0^t \left(\overline{\phi} \left(\overline{Q}^N(s-) \right)^T - \overline{\phi} \left(\overline{Q}(s-) \right)^T \right) \mathbf{diag} \left(\overline{Q}(s) \wedge \overline{K}(s) \right) dZ_{\mu^N}^N(s) \\ &\quad - \int_0^t \left(\mathbf{I} - \overline{\psi}^N \left(\overline{Q}^N(s-) \right)^T \right) \mathbf{diag} \left(\mathbf{h}_1 \left(\overline{Q}^N(s) - \overline{Q}(s) \right) \right) dZ_{\gamma^N}^N(s) \\ &\quad - \int_0^t \left(\mathbf{I} - \overline{\phi}^N \left(\overline{Q}^N(s-) \right)^T \right) \mathbf{diag} \left(\mathbf{h}_2 \left(\overline{Q}^N(s) - \overline{Q}(s) \right) \right) dZ_{\mu^N}^N(s), \end{aligned} \quad (7.7)$$

where $\mathbf{h}_1(\overline{Q}^N(t) - \overline{Q}(t)) = (\overline{Q}^N(t) - \overline{K}^N(t))^+ - (\overline{Q}(t) - \overline{K}(t))^+$, $\mathbf{h}_2(\overline{Q}^N(t) - \overline{Q}(t)) = (\overline{Q}^N(t) \wedge \overline{K}^N(t)) - (\overline{Q}(t) \wedge \overline{K}(t))$, and

$$\begin{aligned} \overline{Y}^N(t) &= \overline{\mathbf{M}}^{N,E}(t) + \left(\overline{\mathbf{M}}^{N,A}(t)^T - \overline{\mathbf{M}}^{N,A}(t) \right) \cdot \mathbf{1} + \left(\overline{\mathbf{M}}^{N,S}(t)^T - \overline{\mathbf{M}}^{N,S}(t) \right) \cdot \mathbf{1} \\ &\quad - \overline{\mathbf{M}}_0^{N,A}(t) - \overline{\mathbf{M}}_0^{N,S}(t) + \overline{\mathbf{R}}^{N,E}(t) - \int_0^t \overline{\lambda}^\pi(s) ds \\ &\quad + \int_0^t \left(\overline{\psi}^N \left(\overline{Q}^N(s-) \right)^T - \overline{\psi} \left(\overline{Q}(s-) \right)^T \right) \mathbf{diag} \left(\overline{Q}(s) - \overline{K}(s) \right)^+ dZ_{\gamma^N}^N(s) \\ &\quad - \int_0^t \left(\mathbf{I} - \overline{\psi} \left(\overline{Q}(s-) \right)^T \right) \mathbf{diag} \left(\overline{Q}(s) - \overline{K}(s) \right)^+ \left(dZ_{\gamma^N}^N(s) - \overline{\gamma}^\pi(s) ds \right) \end{aligned}$$

$$\begin{aligned}
 &+ \int_0^t \left(\bar{\phi}^N \left(\bar{\mathbf{Q}}^N(s-) \right)^T - \bar{\phi} \left(\bar{\mathbf{Q}}^N(s-) \right)^T \right) \mathbf{diag} \left(\bar{\mathbf{Q}}(s) \wedge \bar{\mathbf{K}}(s) \right) d\mathbf{Z}_{\mu^N}^N(s) \\
 &- \int_0^t \left(\mathbf{I} - \bar{\phi} \left(\bar{\mathbf{Q}}(s-) \right)^T \right) \mathbf{diag} \left(\bar{\mathbf{Q}}(s) \wedge \bar{\mathbf{K}}(s) \right) \left(d\mathbf{Z}_{\mu^N}^N(s) - \bar{\boldsymbol{\mu}}^\pi(s) ds \right).
 \end{aligned}$$

It follows from Lemma 7.1, Theorem 5.6 along with Assumptions 3.3, 3.5 that for any $T > 0$ and $\epsilon > 0$, $\limsup_{N \rightarrow \infty} \left(\sup_{0 \leq t \leq T} \|\bar{\mathbf{Y}}^N(t)\| > \epsilon \right) = 0$. Using the argument of Theorem 6.1, and applying the Lipschitz continuity of the probability matrices $\bar{\psi}(\cdot)$ and $\bar{\phi}(\cdot)$ and Gronwall’s inequality, it follows from (7.7) that $\limsup_{N \rightarrow \infty} \left(\sup_{0 \leq t \leq T} \|\bar{\mathbf{Q}}^N(t) - \bar{\mathbf{Q}}(t)\| > \epsilon \right) = 0$ for any $T > 0$ and $\epsilon > 0$, which essentially concludes the requisite convergence. \square

8. Proof of diffusion approximation

In this section, we present the detailed proof of the multi-scale diffusion approximation Theorem 3.4 of the system length process within two-time-scale time-varying many-server heavy-traffic regime. Based on the fluid trajectories of $\bar{\mathbf{Q}}$ from Theorem 3.2, the system experiences subcritical, critical or supercritical loading phases (Remark 3.3), and according wise the diffusion scaled length process $\hat{\mathbf{Q}}^N$ converges $\mathbb{D}([0, \infty), \mathbb{R}^L)$.

Summary of proof of Theorem 3.4 Firstly, we divide the main proof into several parts as the final representation (8.6) of the diffusion scaled system length process $\hat{\mathbf{Q}}^N$ consists of certain martingales as well as semimartingales. The representation (8.6) of $\hat{\mathbf{Q}}^N$ necessitates the derivation of the joint convergences of the martingales and semimartingales separately. This joint convergence follows from the multi-dimensional Gaussian convergence of the martingales in Lemma 8.1 and the Skorokhod convergence of the semimartingales in Proposition 8.2 presented in Appendix A. Finally, after deriving all the convergence results, we proceed to main proof and conclude the joint convergence among martingales and semimartingales. In particular, we adapt the martingale central limit theorem by virtue of Ethier and Kurtz (1986) to prove the requisite joint convergence. In the end, the application of the continuous mapping theorem in view of Theorem 6.1 together with the Lipschitz property of $\bar{\psi}(\cdot)$, $\bar{\phi}(\cdot)$ completes the proof.

We define the diffusion scaled martingales by $\widehat{M}_i^{N,E}(t) = \frac{M_i^{N,E}(t)}{\sqrt{N}}$, $\widehat{M}_{ij}^{N,A}(t) = \frac{M_{ij}^{N,A}(t)}{\sqrt{N}}$, $\widehat{M}_{ij}^{N,S}(t) = \frac{M_{ij}^{N,S}(t)}{\sqrt{N}}$ $i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\}$ with trajectories in $\mathbb{D}([0, \infty), \mathbb{R})$. Additionally, define the diffusion scaled integral processes

$$\begin{aligned}
 \widehat{R}_i^{N,E}(t) &= N^{1-\delta} \left(\bar{R}_i^{N,E}(t) - \bar{R}_i^E(t) \right), \quad i \in \mathbb{L}, \\
 \widehat{R}_{ij}^{N,A}(t) &= N^{1-\delta} \left(\bar{R}_{ij}^{N,A}(t) - \bar{R}_{ij}^A(t) \right), \quad i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\}, \\
 \widehat{R}_{ij}^{N,S}(t) &= N^{1-\delta} \left(\bar{R}_{ij}^{N,S}(t) - \bar{R}_{ij}^S(t) \right), \quad i \in \mathbb{L}, j \in \mathbb{L} \cup \{0\}.
 \end{aligned}$$

Consequently, the representations (7.1) and (7.4) result the diffusion scaled representation of $\hat{\mathbf{Q}}^N(t)$ as follows:

$$\begin{aligned}
 \widehat{\mathbf{Q}}^N(t) &= \widehat{\mathbf{Q}}^N(0) + N^{1/2-\delta} \left(\widehat{\mathbf{M}}^{N,E}(t) + \left(\widehat{\mathbf{M}}^{N,A}(t)^T - \widehat{\mathbf{M}}^{N,A}(t) \right) \cdot \mathbf{1} + \left(\widehat{\mathbf{M}}^{N,S}(t)^T - \widehat{\mathbf{M}}^{N,S}(t) \right) \cdot \mathbf{1} \right. \\
 &\quad \left. - \widehat{\mathbf{M}}_0^{N,A}(t) - \widehat{\mathbf{M}}_0^{N,S}(t) \right) + \widehat{\mathbf{R}}^{N,E}(t) + \left(\widehat{\mathbf{R}}^{N,A}(t)^T - \widehat{\mathbf{R}}^{N,A}(t) \right) \cdot \mathbf{1} \\
 &\quad + \left(\widehat{\mathbf{R}}^{N,S}(t)^T - \widehat{\mathbf{R}}^{N,S}(t) \right) \cdot \mathbf{1} - \widehat{\mathbf{R}}_0^{N,A}(t) - \widehat{\mathbf{R}}_0^{N,S}(t).
 \end{aligned} \tag{8.1}$$

Note that, by the given hypothesis, the routing substochastic zero-diagonal matrices $\psi^N(\cdot)$ and $\phi^N(\cdot)$ satisfy the following relations

$$\sum_{j \in \mathbb{L}} \bar{\psi}_{ij}^N(\cdot) + \bar{\psi}_{i0}^N(\cdot) = 1, \sum_{j \in \mathbb{L}} \hat{\psi}_{ij}^N(\cdot) + \hat{\psi}_{i0}^N(\cdot) = 0, \sum_{j \in \mathbb{L}} \bar{\phi}_{ij}^N(\cdot) + \bar{\phi}_{i0}^N(\cdot) = 1, \sum_{j \in \mathbb{L}} \hat{\phi}_{ij}^N(\cdot) + \hat{\phi}_{i0}^N(\cdot) = 0.$$

In order to analyze the Lebesgue-Stieltjes integral processes $\hat{\mathbf{R}}_1^{N,A}$ ($\hat{\mathbf{R}}_0^{N,A}$) and $\hat{\mathbf{R}}_1^{N,S}$ ($\hat{\mathbf{R}}_0^{N,S}$) in (8.1), by view of definitions (7.2) and (7.3), we define

$$\begin{aligned} \hat{\mathbf{R}}_1^{N,A}(t) &= -N^{1-\delta-\beta/2} \int_0^t \left(\mathbf{I} - \left(\bar{\psi}^N \left(\bar{\mathbf{Q}}^N(s-) \right) \right)^T \right) \mathbf{diag} \left(\bar{\mathbf{Q}}(s) - \bar{\mathbf{K}}(s) \right)^+ d\hat{\mathbf{Z}}_{\gamma^N}^N(s), \\ \hat{\mathbf{R}}_2^{N,A}(t) &= \int_0^t \left(\hat{\psi}^N \left(\bar{\mathbf{Q}}^N(s-) \right) \right)^T \mathbf{diag} \left(\bar{\mathbf{Q}}(s) - \bar{\mathbf{K}}(s) \right)^+ \bar{\gamma}^{N,\pi}(s) ds, \\ \hat{\mathbf{R}}_3^{N,A}(t) &= - \int_0^t \left(\mathbf{I} - \left(\bar{\psi} \left(\bar{\mathbf{Q}}(s-) \right) \right)^T \right) \mathbf{diag} \left(\bar{\mathbf{Q}}(s) - \bar{\mathbf{K}}(s) \right)^+ \bar{\gamma}^{N,\pi}(s) ds, \end{aligned} \tag{8.2}$$

and

$$\begin{aligned} \hat{\mathbf{R}}_1^{N,S}(t) &= -N^{1-\delta-\beta/2} \int_0^t \left(\mathbf{I} - \left(\bar{\phi}^N \left(\bar{\mathbf{Q}}^N(s-) \right) \right)^T \right) \mathbf{diag} \left(\bar{\mathbf{Q}}(s) \wedge \bar{\mathbf{K}}(s) \right) d\hat{\mathbf{Z}}_{\mu^N}^N(s), \\ \hat{\mathbf{R}}_2^{N,S}(t) &= \int_0^t \left(\hat{\phi}^N \left(\bar{\mathbf{Q}}^N(s-) \right) \right)^T \mathbf{diag} \left(\bar{\mathbf{Q}}(s) \wedge \bar{\mathbf{K}}(s) \right) \bar{\mu}^{N,\pi}(s) ds, \\ \hat{\mathbf{R}}_3^{N,S}(t) &= - \int_0^t \left(\mathbf{I} - \left(\bar{\phi} \left(\bar{\mathbf{Q}}(s-) \right) \right)^T \right) \mathbf{diag} \left(\bar{\mathbf{Q}}(s) \wedge \bar{\mathbf{K}}(s) \right) \hat{\mu}^{N,\pi}(s) ds. \end{aligned} \tag{8.3}$$

such that

$$\begin{aligned} & \left(\hat{\mathbf{R}}^{N,A}(t)^T - \hat{\mathbf{R}}^{N,A}(t) \right) \cdot \mathbf{1} - \hat{\mathbf{R}}_0^{N,A}(t) \\ &= \sum_{j=1}^3 \hat{\mathbf{R}}_j^{N,A}(t) + N^{1-\delta} \int_0^t \left(\bar{\psi} \left(\bar{\mathbf{Q}}^N(s-) \right) - \bar{\psi} \left(\bar{\mathbf{Q}}(s-) \right) \right)^T \mathbf{diag} \left(\bar{\mathbf{Q}}(s) - \bar{\mathbf{K}}(s) \right)^+ \bar{\gamma}^{N,\pi}(s) ds, \\ & \quad - \int_0^t \left(\mathbf{I} - \bar{\psi}^N \left(\bar{\mathbf{Q}}^N(s-) \right) \right)^T \mathbf{diag} \left(\mathbf{h}_1 \left(\hat{\mathbf{Q}}^N(s) \right) \right) d\mathbf{Z}_{\gamma^N}^N(s) \\ & \left(\hat{\mathbf{R}}^{N,S}(t)^T - \hat{\mathbf{R}}^{N,S}(t) \right) \cdot \mathbf{1} - \hat{\mathbf{R}}_0^{N,S}(t) \\ &= \sum_{j=1}^3 \hat{\mathbf{R}}_j^{N,S}(t) + N^{1-\delta} \int_0^t \left(\bar{\phi} \left(\bar{\mathbf{Q}}^N(s-) \right) - \bar{\phi} \left(\bar{\mathbf{Q}}(s-) \right) \right)^T \mathbf{diag} \left(\bar{\mathbf{Q}}(s) \wedge \bar{\mathbf{K}}(s) \right) \bar{\mu}^{N,\pi}(s) ds, \\ & \quad - \int_0^t \left(\mathbf{I} - \bar{\phi}^N \left(\bar{\mathbf{Q}}^N(s-) \right) \right)^T \mathbf{diag} \left(\mathbf{h}_2 \left(\hat{\mathbf{Q}}^N(s) \right) \right) d\mathbf{Z}_{\mu^N}^N(s), \end{aligned}$$

Similarly, the decomposition of $\hat{\mathbf{R}}^{N,E}$ into $\hat{\mathbf{R}}_1^{N,E}$ and $\hat{\mathbf{R}}_2^{N,E}$ yields,

$$\hat{\mathbf{R}}_1^{N,E}(t) = N^{1-\delta-\beta/2} \hat{\mathbf{Z}}_{\lambda^N}^N(t), \hat{\mathbf{R}}_2^{N,E}(t) = \int_0^t \hat{\lambda}^{N,\pi}(s) ds. \tag{8.4}$$

By given hypotheses, $\bar{\psi}(\cdot)$, $\bar{\psi}_0(\cdot)$ and $\bar{\phi}(\cdot)$, $\bar{\phi}_0(\cdot)$ are differentiable and by Taylor's approximations, we obtain for $\bar{\chi}(\cdot) = \bar{\psi}(\cdot)$, $\bar{\psi}_0(\cdot)$, $\bar{\phi}(\cdot)$, $\bar{\phi}_0(\cdot)$,

$$N^{1-\delta} \left(\bar{\chi} \left(\bar{\mathbf{Q}}^N(t) \right) - \bar{\chi} \left(\bar{\mathbf{Q}}(t) \right) \right) = \langle \nabla \bar{\chi} \left(\bar{\mathbf{Q}}(t) \right), \hat{\mathbf{Q}}^N(t) \rangle + o \left(N^{1-\delta} \|\bar{\mathbf{Q}}^N(t) - \bar{\mathbf{Q}}(t)\| \right), \tag{8.5}$$

where for the remainder term, $\frac{o(N^{1-\delta} \|\bar{\mathbf{Q}}^N(t) - \bar{\mathbf{Q}}(t)\|)}{N^{1-\delta} \|\bar{\mathbf{Q}}^N(t) - \bar{\mathbf{Q}}(t)\|} \xrightarrow{\mathbb{P}} 0$ uniformly on compact sets as $N \rightarrow \infty$.

Here, $\langle \nabla \mathbf{P}(\mathbf{x}), \mathbf{y} \rangle := \sum_{i=1}^L \frac{\partial}{\partial x_i} \mathbf{P}(\mathbf{x}) \cdot y_i$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^L$ and $\mathbf{P}(\cdot) \in \mathbb{M}^{L \times L}$.

In view of the decompositions (8.2), (8.3), (8.4) and (8.5), the representation (8.1) reduces into

$$\begin{aligned} \widehat{Q}^N(t) = & \widehat{Q}^N(0) + N^{1/2-\delta} \left(\widehat{M}^{N,E}(t) + (\widehat{M}^{N,A}(t))^T - \widehat{M}^{N,A}(t) \right) \cdot \mathbf{1} + (\widehat{M}^{N,S}(t))^T - \widehat{M}^{N,S}(t) \cdot \mathbf{1} \\ & - \widehat{M}_0^{N,A}(t) - \widehat{M}_0^{N,S}(t) + \sum_{j=1}^2 \widehat{R}_j^{N,E}(t) + \sum_{j=1}^3 \widehat{R}_j^{N,A}(t) + \sum_{j=1}^3 \widehat{R}_j^{N,S}(t) \\ & + \int_0^t \langle \nabla \bar{\psi}(\bar{Q}(s-)), \widehat{Q}^N(s) \rangle^T \text{diag}(\bar{Q}(s) - \bar{K}(s))^+ \bar{\gamma}^{N,\pi}(s) \, ds, \\ & + \int_0^t \langle \nabla \bar{\phi}(\bar{Q}(s-)), \widehat{Q}^N(s) \rangle^T \text{diag}(\bar{Q}(s) \wedge \bar{K}(s)) \bar{\mu}^{N,\pi}(s) \, ds, \\ & - \int_0^t \left(\mathbf{I} - \bar{\psi}^N(\bar{Q}^N(s-))^T \right) \text{diag}(\mathbf{h}_1(\widehat{Q}^N(s))) \, d\mathbf{Z}_{\gamma^N}^N(s) \\ & - \int_0^t \left(\mathbf{I} - \bar{\phi}^N(\bar{Q}^N(s-))^T \right) \text{diag}(\mathbf{h}_2(\widehat{Q}^N(s))) \, d\mathbf{Z}_{\mu^N}^N(s), \end{aligned} \tag{8.6}$$

where $\mathbf{h}_1(\widehat{Q}^N(t)) = (\widehat{Q}^N(t) - \widehat{K}^N(t)) \mathbf{1}(\bar{Q}(t) > \bar{K}(t)) + (\widehat{Q}^N(t) - \widehat{K}^N(t))^+ \mathbf{1}(\bar{Q}(t) = \bar{K}(t))$, and $\mathbf{h}_2(\widehat{Q}^N(t)) = \widehat{Q}^N(t) \mathbf{1}(\bar{Q}(t) < \bar{K}(t)) + \widehat{Q}^N(t) \wedge \widehat{K}^N(t) \mathbf{1}(\bar{Q}(t) = \bar{K}(t)) + \widehat{K}^N(t) \mathbf{1}(\bar{Q}(t) > \bar{K}(t))$.

Before proceeding to the complete proof of Theorem 3.4, we divide the complete proof into two parts. Firstly, we derive the Gaussian approximation of the martingales of (8.6) in Lemma 8.1, and secondly the diffusion approximation of the integral processes $(\widehat{R}_i^{N,E}, \widehat{R}_j^{N,I}, i = 1, 2, j = 1, 2, 3, 4, I = A, S)$ is obtained in Proposition 8.2, and the latter is presented in Appendix A.

Lemma 8.1. *Suppose the fluid approximation of Theorem 3.2 is satisfied. Then, under Assumption 3.2–Assumption 3.4, the joint process $(\widehat{M}^{N,E}, \widehat{M}_0^{N,A}, \widehat{M}_0^{N,S}, \widehat{M}^{N,A}, \widehat{M}^{N,S})$ converges in distribution to $(\{B^E(t), t \geq 0\}, \{B^A(t), t \geq 0\}, \{B_0^S(t), t \geq 0\}, \{B^A(t), t \geq 0\}, \{B^S(t), t \geq 0\})$ in $\mathbb{D}([0, \infty), \mathbb{R}^L)^3 \times \mathbb{D}([0, \infty), \mathbb{M}^{L \times L})^2$ as $N \rightarrow \infty$, where $B^E(t) = ((B_i^E \circ \bar{R}_i^E)(t))_{i=1}^L$, $B_0^A(t) = ((B_{i0}^A \circ \bar{R}_{i0}^A)(t))_{i=1}^L$, $B_0^S(t) = ((B_{i0}^S \circ \bar{R}_{i0}^E)(t))_{i=1}^L$, $B^A(t) = ((B_{ij}^A \circ \bar{R}_{ij}^A)(t))_{i,j=1}^L$, $B^S(t) = ((B_{ij}^S \circ \bar{R}_{ij}^E)(t))_{i,j=1}^L$, with $(B_i^E)_{i=1}^L, (B_{i0}^A)_{i=1}^L, (B_{i0}^S)_{i=1}^L, (B_{ij}^A)_{i,j=1}^L, (B_{ij}^S)_{i,j=1}^L$ being mutually independent \mathbb{R}^L and $\mathbb{M}^{L \times L}$ -valued standard Brownian motions.*

Our next aim is to show the joint convergence of $(\widehat{R}_1^{N,E}, \widehat{R}_2^{N,E}, \widehat{R}_1^{N,I}, \widehat{R}_2^{N,I}, \widehat{R}_3^{N,I}, I = A, S)$ as $N \rightarrow \infty$. By definitions, (8.2), (8.3) and (8.4), the convergences of $\widehat{R}_1^{N,E}, \widehat{R}_1^{N,A}$ and $\widehat{R}_1^{N,S}$, are entirely relied on the convergence of the semimartingales $\widehat{Z}_{\lambda^N}^N, \widehat{Z}_{\gamma^N}^N, \widehat{Z}_{\mu^N}^N$ derived in Theorem 5.6. In the next step, to apply Theorem 5.8, we adapt the semimartingale decomposition

$$\widehat{Z}_{f^N}^N = \mathbf{V}_{f^N}^N + \mathbf{G}_{f^N}^N(t), \quad f^N = \lambda^N(\cdot), \gamma^N(\cdot), \mu^N(\cdot), \tag{8.7}$$

where $\mathbf{V}_{\lambda^N}^N, \mathbf{V}_{\gamma^N}^N, \mathbf{V}_{\mu^N}^N$ in $\mathbb{D}([0, \infty), \mathbb{R}^L)$ are \mathcal{F}^N -local square-integrable martingales, and Theorem 5.6 yields the convergences $(\mathbf{V}_{\lambda^N}^N, \mathbf{V}_{\gamma^N}^N, \mathbf{V}_{\mu^N}^N)$,

$$\begin{aligned} \langle \mathbf{V}_{\lambda^N}^N \rangle(t) & \xrightarrow{\mathbb{P}} \mathbf{C}_{\lambda\lambda}(t), & \langle \mathbf{V}_{\gamma^N}^N \rangle(t) & \xrightarrow{\mathbb{P}} \mathbf{C}_{\gamma\gamma}(t), & \langle \mathbf{V}_{\mu^N}^N \rangle(t) & \xrightarrow{\mathbb{P}} \mathbf{C}_{\mu\mu}(t), \\ \langle \mathbf{V}_{\lambda^N}^N, \mathbf{V}_{\gamma^N}^N \rangle(t) & \xrightarrow{\mathbb{P}} \mathbf{C}_{\lambda\gamma}(t), & \langle \mathbf{V}_{\lambda^N}^N, \mathbf{V}_{\mu^N}^N \rangle(t) & \xrightarrow{\mathbb{P}} \mathbf{C}_{\lambda\mu}(t), & \langle \mathbf{V}_{\gamma^N}^N, \mathbf{V}_{\mu^N}^N \rangle(t) & \xrightarrow{\mathbb{P}} \mathbf{C}_{\gamma\mu}(t), \end{aligned}$$

and $\mathbf{G}_{f^N}^N$ is the finite-variation process. The representation of $\mathbf{G}_{f^N}^N$ can be understood by the additional term in (A.3). By definition, the decomposition (A.3) leads to

$$\sup_{0 \leq t \leq T} \|\widehat{Z}_{f^N}^N(t) - \mathbf{V}_{f^N}^N(t)\| \xrightarrow{\mathbb{P}} 0, \text{ as } N \rightarrow \infty. \tag{8.8}$$

Next, we define

$$\begin{aligned}\widehat{\mathbf{R}}_1^E(t) &= \mathbf{1}(\beta \leq 1) \mathbf{W}_\lambda(t), \\ \widehat{\mathbf{R}}_1^A(t) &= -\mathbf{1}(\beta \leq 1) \int_0^t \left(\mathbf{I} - (\overline{\psi}(\overline{\mathbf{Q}}(s)))^T \right) \text{diag}(\overline{\mathbf{Q}}(s) - \overline{\mathbf{K}}(s))^+ d\mathbf{W}_\gamma(s), \\ \widehat{\mathbf{R}}_2^A(t) &= \int_0^t \left(\widehat{\psi}(\overline{\mathbf{Q}}(s-)) \right)^T \text{diag}(\overline{\mathbf{Q}}(s) - \overline{\mathbf{K}}(s))^+ \widehat{\gamma}^\pi(s) ds, \\ \widehat{\mathbf{R}}_3^A(t) &= -\int_0^t \left(\mathbf{I} - (\overline{\psi}(\overline{\mathbf{Q}}(s-)))^T \right) \text{diag}(\overline{\mathbf{Q}}(s) - \overline{\mathbf{K}}(s))^+ \widehat{\gamma}^\pi(s) ds, \\ \widehat{\mathbf{R}}_1^S(t) &= -\mathbf{1}(\beta \leq 1) \int_0^t \left(\mathbf{I} - (\overline{\phi}(\overline{\mathbf{Q}}(s)))^T \right) \text{diag}(\overline{\mathbf{Q}}(s) \wedge \overline{\mathbf{K}}(s)) d\mathbf{W}_\mu(s), \\ \widehat{\mathbf{R}}_2^S(t) &= \int_0^t \left(\widehat{\phi}(\overline{\mathbf{Q}}(s)) \right)^T \text{diag}(\overline{\mathbf{Q}}(s) \wedge \overline{\mathbf{K}}(s)) \widehat{\mu}^\pi(s) ds, \\ \widehat{\mathbf{R}}_3^S(t) &= -\int_0^t \left(\mathbf{I} - (\overline{\phi}(\overline{\mathbf{Q}}(s)))^T \right) \text{diag}(\overline{\mathbf{Q}}(s) \wedge \overline{\mathbf{K}}(s)) \widehat{\mu}^\pi(s) ds,\end{aligned}$$

Proposition 8.2. *Under Assumptions 3.3–3.6, the process $(\widehat{\mathbf{R}}_1^{N,E}, \widehat{\mathbf{R}}_2^{N,E}, \widehat{\mathbf{R}}_1^{N,A}, \widehat{\mathbf{R}}_2^{N,A}, \widehat{\mathbf{R}}_3^{N,A}, \widehat{\mathbf{R}}_1^{N,S}, \widehat{\mathbf{R}}_2^{N,S}, \widehat{\mathbf{R}}_3^{N,S})$ jointly converges in distribution to $(\widehat{\mathbf{R}}_1^E, \int_0^\cdot \widehat{\lambda}^\pi(s) ds, \widehat{\mathbf{R}}_1^A, \widehat{\mathbf{R}}_2^A, \widehat{\mathbf{R}}_3^A, \widehat{\mathbf{R}}_1^S, \widehat{\mathbf{R}}_2^S, \widehat{\mathbf{R}}_3^S)$ with trajectories in $\mathbb{D}([0, \infty), \mathbb{R}^L)^8$, where with $\mathbf{W}_\lambda, \mathbf{W}_\gamma, \mathbf{W}_\mu$ being the \mathbb{R}^L -valued correlated Gaussian process with covariance $\mathbf{C}_{fg}(t)$, with $\mathbf{f}_i(t) = \lambda_i(t), \gamma_i(t), \mu_i(t), \mathbf{g}_j(t) = \lambda_j(t), \gamma_j(t), \mu_j(t), i, j \in \mathbb{L}$.*

Proof of Theorem 3.4: We begin the proof by establishing the joint convergence of

$$(\widehat{\mathbf{Q}}^N(0), \widehat{\mathbf{M}}^{N,E}, \widehat{\mathbf{M}}_0^{N,A}, \widehat{\mathbf{M}}_0^{N,S}, \widehat{\mathbf{M}}^{N,A}, \widehat{\mathbf{M}}^{N,S}, \widehat{\mathbf{R}}_1^{N,E}, \widehat{\mathbf{R}}_2^{N,E}, \widehat{\mathbf{R}}_1^{N,A}, \widehat{\mathbf{R}}_2^{N,A}, \widehat{\mathbf{R}}_3^{N,A}, \widehat{\mathbf{R}}_1^{N,S}, \widehat{\mathbf{R}}_2^{N,S}, \widehat{\mathbf{R}}_3^{N,S})$$

in $\mathbb{R}^L \times \mathbb{D}([0, \infty), \mathbb{R}^L)^3 \times \mathbb{D}([0, \infty), \mathbb{M}^{L \times L})^2 \times \mathbb{D}([0, \infty), \mathbb{R}^L)^8$.

Knowing this $1/2 - \delta = \min\{(\beta - 1)/2, 0\}$, the joint convergence of the martingales

$$N^{1/2-\delta}(\widehat{\mathbf{M}}^{N,E}, \widehat{\mathbf{M}}_0^{N,A}, \widehat{\mathbf{M}}_0^{N,S}, \widehat{\mathbf{M}}^{N,A}, \widehat{\mathbf{M}}^{N,S})$$

to the Gaussian process $\mathbf{1}(\beta \geq 1)(\mathbf{B}^E, \mathbf{B}_0^A, \mathbf{B}_0^S, \mathbf{B}^A, \mathbf{B}^S)$ follows from Lemma 8.1. On the other hand, Proposition 8.2 shows that convergence of $(\widehat{\mathbf{R}}_1^{N,E}, \widehat{\mathbf{R}}_2^{N,E}, \widehat{\mathbf{R}}_1^{N,A}, \widehat{\mathbf{R}}_2^{N,A}, \widehat{\mathbf{R}}_3^{N,A}, \widehat{\mathbf{R}}_1^{N,S}, \widehat{\mathbf{R}}_2^{N,S}, \widehat{\mathbf{R}}_3^{N,S})$ with trajectories in $\mathbb{D}([0, \infty), \mathbb{R}^L)^8$.

Without loss of generality, in the rest of the proof, we assume $\widehat{\mathbf{M}}^{N,I}(t)$ for $\mathbb{M}^{L \times L}$ -valued matrices $\widehat{\mathbf{M}}^{N,I}(t)$ on $\mathbb{D}([0, \infty), \mathbb{R}^L)$, $I = A, S$. In order to prove the joint convergence of

$$(N^{1/2-\delta}(\widehat{\mathbf{M}}^{N,E}, \widehat{\mathbf{M}}_0^{N,A}, \widehat{\mathbf{M}}_0^{N,S}, \widehat{\mathbf{M}}^{N,A}, \widehat{\mathbf{M}}^{N,S}), \widehat{\mathbf{R}}_1^{N,E}, \widehat{\mathbf{R}}_1^{N,A}, \widehat{\mathbf{R}}_1^{N,S})$$

in $\mathbb{D}([0, \infty), \mathbb{R}^L)^8$, we consider the process

$$\begin{aligned}& \left(N^{1/2-\delta}(\widehat{\mathbf{M}}^{N,E}, \widehat{\mathbf{M}}_0^{N,A}, \widehat{\mathbf{M}}_0^{N,S}, \widehat{\mathbf{M}}^{N,A}, \widehat{\mathbf{M}}^{N,S}), \right. \\ & \left. N^{1-\delta-\beta/2} \left(\mathbf{V}_{\lambda^N}^N, \int_0^\cdot \left(\mathbf{I} - (\overline{\psi}^N(\overline{\mathbf{Q}}^N(s)))^T \right) \text{diag}(\overline{\mathbf{Q}}(s) - \overline{\mathbf{K}}(s))^+ d\mathbf{V}_{\gamma^N}^N(s), \right. \right. \\ & \left. \left. \int_0^\cdot \left(\mathbf{I} - (\overline{\phi}^N(\overline{\mathbf{Q}}^N(s)))^T \right) \text{diag}(\overline{\mathbf{Q}}(s) \wedge \overline{\mathbf{K}}(s)) d\mathbf{V}_{\mu^N}^N(s) \right) \right)\end{aligned}$$

which is an \mathcal{F}^N -local square-integrable martingale with trajectories in $\mathbb{D}([0, \infty), \mathbb{R}^L)$ ⁸. Note that the semimartingale decomposition in view of (8.7) results

$$\left[N^{1/2-\delta} \widehat{\mathbf{M}}^{N,I}, N^{1-\delta-\beta/2} \mathbf{V}_{\mathbf{f}^N}^N \right] (t) = N^{3/2-2\delta-\beta/2} \left(\left[\widehat{\mathbf{M}}^{N,I}, \widehat{\mathbf{Z}}_{\mathbf{f}^N}^N \right] (t) + \left[\widehat{\mathbf{M}}^{N,I}, \mathbf{G}_{\mathbf{f}^N}^N \right] (t) \right),$$

where $I = E, A, S$, and $\mathbf{f}^N = \boldsymbol{\lambda}^N(\cdot), \boldsymbol{\gamma}^N(\cdot), \boldsymbol{\mu}^N$.

Since $\widehat{\mathbf{Z}}_{\mathbf{f}^N}^N$ is a continuous-time finite-variation process, it yields $\left[\widehat{\mathbf{M}}^{N,I}, \widehat{\mathbf{Z}}_{\mathbf{f}^N}^N \right] (t) = 0$ for any $t \geq 0$. Therefore, we remain with $N^{3/2-2\delta-\beta/2} \left[\widehat{\mathbf{M}}^{N,I}, \mathbf{G}_{\mathbf{f}^N}^N \right] (t)$ which converges to zero in probability, as $N \rightarrow \infty$, for all $t \geq 0$. Consequently,

$$\langle N^{1/2-\delta} \widehat{\mathbf{M}}^{N,I}, N^{1-\delta-\beta/2} \mathbf{V}_{\mathbf{f}^N}^N \rangle (t) \xrightarrow{\mathbb{P}} \mathbf{1}(\beta <, =, > 1) \mathbf{0},$$

as $N \rightarrow \infty$. Moreover the jumps of $N^{1/2-\delta} \widehat{\mathbf{M}}^{N,I}$ and $N^{1-\delta-\beta/2} \mathbf{V}_{\mathbf{f}^N}^N$ are not greater than $N^{-\delta}$ and $N^{1-\delta-\beta}$, respectively, which vanishes to zero as N grows large, as $\delta, 1 - \delta - \beta = \min\{-\beta/2, 1/2 - \beta\} < 0$. Therefore, as a consequence of the martingale central limit theorem followed by (8.8), the above joint martingale converges in distribution to $(\mathbf{1}(\beta \geq 1)(\mathbf{B}^E, \mathbf{B}_0^A, \mathbf{B}_0^S, \mathbf{B}^A, \mathbf{B}^S), \mathbf{1}(\beta \leq 1)(\mathbf{W}_\lambda, \int_0^\cdot (\mathbf{I} - (\overline{\psi}(\overline{\mathbf{Q}}(s)))^T) \text{diag}(\overline{\mathbf{Q}}(s) - \overline{\mathbf{K}}(s))^+ d\mathbf{W}_\gamma(s), \int_0^\cdot (\mathbf{I} - (\overline{\phi}(\overline{\mathbf{Q}}(s)))^T) \text{diag}(\overline{\mathbf{Q}}(s) \wedge \overline{\mathbf{K}}(s)) d\mathbf{W}_\mu(s))$. Finally, including the deterministic convergence of the remaining process $(\widehat{\mathbf{Q}}^N(0), \widehat{\mathbf{R}}_2^{N,E}, \widehat{\mathbf{R}}_2^{N,A}, \widehat{\mathbf{R}}_3^{N,A}, \widehat{\mathbf{R}}_2^{N,S}, \widehat{\mathbf{R}}_3^{N,S})$ resulting from the given hypotheses, the convergence relation (8.8) implies the complete joint converges.

In the end, we apply the same analogy used earlier in the proof of Theorem 3.2 in the representation (8.6) to show the continuity of the integral map Φ by virtue of (6.1). Using the Lipschitz continuity of \mathbf{h}_1 and \mathbf{h}_2 , the direct consequence of the continuity mapping theorem via Theorem 6.1 yields the weak convergence of $\widehat{\mathbf{Q}}^N$ and completes the proof. \square

Appendix A. Convergence of the integral processes with respect to external process

Proof of Theorem 5.6: Firstly, consider the individual process $\overline{\mathbf{Z}}_{\mathbf{f}^N}^N, \mathbf{f}^N(\cdot) = (\boldsymbol{\lambda}_i^N(\cdot))_{i=1}^L, (\boldsymbol{\gamma}_i^N(\cdot))_{i=1}^L, (\boldsymbol{\mu}_i^N(\cdot))_{i=1}^L$, and define $B_i^N(t, \mathbf{z}, \mathbf{h}) = (\mathbf{f}_i^N(t))^T (\mathbf{h} - \boldsymbol{\pi}), i \in \mathbb{L}$. Consider the corresponding \mathbb{R}^L -valued function $\mathbf{B}^N(t, \mathbf{z}, \mathbf{h}) = (B_i^N(t, \mathbf{z}, \mathbf{h}))_{i=1}^L$. Therefore by definition (5.19),

$$d\overline{\mathbf{Z}}_{\mathbf{f}^N}^N(t) = \mathbf{B}^N(t, \overline{\mathbf{Z}}_{\mathbf{f}^N}^N(t), \mathbf{H}^N(t)) dt, \quad \overline{\mathbf{Z}}_{\mathbf{f}^N}^N(0) = \mathbf{0}.$$

Take $\mathbf{f}_i(\cdot) = \boldsymbol{\lambda}_i(\cdot), \boldsymbol{\gamma}_i(\cdot), \boldsymbol{\mu}_i(\cdot)$ and define $B_i(t, \mathbf{z}, \mathbf{h}) = (\mathbf{f}_i(t))^T (\mathbf{h} - \boldsymbol{\pi})$, and the \mathbb{R}^L -valued function $\mathbf{B}(t, \mathbf{z}, \mathbf{h}) = (B_i(t, \mathbf{z}, \mathbf{h}))_{i=1}^L$. By Assumption 3.3, it follows that $\|\mathbf{B}^N(t, \mathbf{z}, \mathbf{h}) - \mathbf{B}(t, \mathbf{z}, \mathbf{h})\| \rightarrow 0$ as $N \rightarrow \infty$ uniformly on compact sets. Now, consider this particular form of $\mathbf{B}^N(t, \mathbf{z}, \mathbf{h})$ and $\mathbf{A}^N(t, \mathbf{z}, \mathbf{h}) = \mathbf{0}$ in (5.1). It is easy to verify that for each $N \in \mathbb{N}$, the functions $\mathbf{A}^N(t, \mathbf{z}, \mathbf{h})$ and $\mathbf{B}^N(t, \mathbf{z}, \mathbf{h})$ satisfy the regularity Assumptions 5.1–5.8. Therefore, Theorem 5.1 implies that $\overline{\mathbf{Z}}_{\mathbf{f}^N}^N$ converges in probability to $\{\mathbf{B}_0(t), t \geq 0\}$ uniformly on compact sets, where

$$d\mathbf{B}_0(t) = \overline{\mathbf{B}}^\pi(t, \mathbf{B}_0(t)) dt, \quad \mathbf{B}_0(0) = \mathbf{0}.$$

By definition, $\overline{\mathbf{B}}^\pi(t, \mathbf{z}) = \sum_{\mathbf{h}} \mathbf{B}(t, \mathbf{z}, \mathbf{h}) \pi(\mathbf{h})$, which turns out to be trivial vector $\mathbf{0}$, and hence concludes the convergence of $\overline{\mathbf{Z}}_{\mathbf{f}^N}^N$ in $\mathbb{C}([0, \infty), \mathbb{R}^L)$ to the zero process, for each $\mathbf{f}^N(\cdot)$. Consequently, the joint deterministic trivial convergence of $(\overline{\mathbf{Z}}_{\boldsymbol{\lambda}^N}^N, \overline{\mathbf{Z}}_{\boldsymbol{\gamma}^N}^N, \overline{\mathbf{Z}}_{\boldsymbol{\mu}^N}^N)$ is established.

Next, to derive the diffusion convergence, we first establish the diffusion convergence of each process $\widehat{\mathbf{Z}}_{\mathbf{f}^N}^N$, for every $\mathbf{f}^N(\cdot)$, in $\mathbb{C}([0, \infty), \mathbb{R}^L)$. Likewise the earlier case, define $A_i^N(t, \mathbf{z}, \mathbf{h}) =$

$(\mathbf{f}_i^N(t))^T(\mathbf{h} - \boldsymbol{\pi})$, $i \in \mathbb{L}$, and the \mathbb{R}^L -valued function $\mathbf{A}^N(t, \mathbf{z}, \mathbf{h}) = (A_i^N(t, \mathbf{z}, \mathbf{h}))_{i=1}^L$. Thus, in view of (5.20), we obtain

$$d\widehat{\mathbf{Z}}_{\mathbf{f}^N}^N(t) = N^{\beta/2} \mathbf{A}^N(t, \widehat{\mathbf{Z}}_{\mathbf{f}^N}^N(t), \mathbf{H}^N(t)) dt, \quad \widehat{\mathbf{Z}}_{\mathbf{f}^N}^N(0) = \mathbf{0}.$$

Again, take $A_i(t, \mathbf{z}, \mathbf{h}) = (\mathbf{f}_i(t))^T(\mathbf{h} - \boldsymbol{\pi})$ with $\mathbf{f}_i(\cdot)$, and consider the \mathbb{R}^L -valued function $\mathbf{A}(t, \mathbf{z}, \mathbf{h}) = (A_i(t, \mathbf{z}, \mathbf{h}))_{i=1}^L$. Therefore, Assumption 3.3 yields the convergence $\|\mathbf{A}^N(t, \mathbf{z}, \mathbf{h}) - \mathbf{A}(t, \mathbf{z}, \mathbf{h})\| \rightarrow 0$ uniformly on compact sets as $N \rightarrow \infty$. Substitute the given $\mathbf{A}^N(t, \mathbf{z}, \mathbf{h})$ and $\mathbf{B}^N(t, \mathbf{z}, \mathbf{h}) = \mathbf{0}$ in (5.1). Under regularity Assumptions 5.1–5.8, Theorem 5.1 ensures that $\widehat{\mathbf{Z}}_{\mathbf{f}^N}^N$ converges in distribution to the diffusion process $\{\mathbf{W}_{\mathbf{f}}(t), t \geq 0\}$, which is the solution of SDE

$$d\mathbf{W}_{\mathbf{f}}(t) = \overline{\mathbf{A}}^{1,\boldsymbol{\pi}}(t, \mathbf{W}_{\mathbf{f}}(t)) d\mathbf{W}(t) + \overline{\mathbf{A}}^{2,\boldsymbol{\pi}}(t, \mathbf{W}_{\mathbf{f}}(t)) dt, \quad \mathbf{W}_{\mathbf{f}}(0) = \mathbf{0}.$$

Calculating the coefficients $\overline{\mathbf{A}}^{1,\boldsymbol{\pi}}(t, \mathbf{z})$ and $\overline{\mathbf{A}}^{2,\boldsymbol{\pi}}(t, \mathbf{z})$ using the formula (5.5), the diffusion approximation of $\widehat{\mathbf{Z}}_{\mathbf{f}^N}^N$ can be concluded. Analogous to the previous proof, the requisite joint convergence of $(\widehat{\mathbf{Z}}_{\boldsymbol{\lambda}^N}^N, \widehat{\mathbf{Z}}_{\boldsymbol{\gamma}^N}^N, \widehat{\mathbf{Z}}_{\boldsymbol{\mu}^N}^N)$ can be established by extending into higher-dimension with trajectories in $\mathbb{C}([0, \infty), \mathbb{R}^{3L})$. \square

Now, we discuss the sufficient predictable uniformly tight (P-UT) condition in order to ensure the weak convergence of the integral processes with respect to semimartingale $\widehat{\mathbf{Z}}_{\boldsymbol{\lambda}^N}^N, \widehat{\mathbf{Z}}_{\boldsymbol{\gamma}^N}^N, \widehat{\mathbf{Z}}_{\boldsymbol{\mu}^N}^N$ as we mentioned in Section 5

Definition A.1. For each $N \in \mathbb{N}$, given an adapted \mathbb{R}^L -valued stochastic process $\{\mathbf{y}^N(t), t \geq 0\}$ with respect to some history $\{\mathcal{Y}^N(t), t \geq 0\}$, the family $\{\mathbf{y}^N, N \in \mathbb{N}\}$ is said to be P-UT if the family of random variables

$$\left\{ \left| \sum_{i=1}^L \mathcal{A}_i^N \cdot \mathcal{Y}_i^N(t) \right| : \{(\mathcal{A}_i^N(t))_{1 \leq i \leq L}, t \geq 0\} \text{ is } \mathcal{Y}^N\text{-predictable with } |\mathcal{A}_i^N| \leq 1, N \in \mathbb{N} \right\} \quad (\text{A.1})$$

is tight in \mathbb{R} for every $t \geq 0$.

P-UT property of $\mathbf{Z}_{\mathbf{f}^N}^N$: Knowing Definition (A.1), first we demonstrate that the family $\{\mathbf{Z}_{\mathbf{f}^N}^N, N \in \mathbb{N}\}$ does not satisfy P-UT condition when $\mathbf{Z}_{\mathbf{f}^N}^N = \widehat{\mathbf{Z}}_{\mathbf{f}^N}^N$, which suffices to show that each sequence $\{\widehat{\mathbf{Z}}_{\mathbf{f}_i^N}^N, N \in \mathbb{N}\}$ is not P-UT for $i = 1, 2, \dots, L$. According to the proof of Theorem 5.1, by (5.16) the semimartingale decomposition yields that

$$\begin{aligned} \widehat{\mathbf{Z}}_{\mathbf{f}_i^N}^N(t) = & M_{g_i^N}^N(t) - N^{-\beta/2} \left(\theta_{g_i}^N(t, \widehat{\mathbf{Z}}_{\mathbf{f}^N}^N(t), \mathbf{H}^N(t)) - \theta_{g_i}^N(0, \mathbf{0}, \mathbf{h}) \right) - N^{-\beta} \left(\phi_{g_i}^N(t, \widehat{\mathbf{Z}}_{\mathbf{f}^N}^N(t), \mathbf{H}^N(t)) \right. \\ & \left. - \phi_{g_i}^N(0, \mathbf{0}, \mathbf{h}) \right) + \int_0^t \left(\mathcal{Q}_x \phi_{g_i}^N(u, \widehat{\mathbf{Z}}_{\mathbf{f}^N}^N(u), \mathbf{H}^N(u)) + \alpha_{g_i}^N(u, \widehat{\mathbf{Z}}_{\mathbf{f}^N}^N(u), \mathbf{H}^N(u)) \right. \\ & \left. + \boldsymbol{\Upsilon}_{g_i}^N(u, \mathbf{X}^N(u), \mathbf{H}^N(u)) \right) du, \end{aligned} \quad (\text{A.2})$$

In the above expression (A.2), for given g , the martingale $\{M_{g_i^N}^N, N \in \mathbb{N}\}$ is P-UT in virtue of Proposition VI.6.13 in Jacod and Shiryaev (2003) as the predictable entity

$$\langle M_{g_i^N}^N \rangle(t) = \int_0^t \Gamma_{g_i^N}^N(u, \widehat{\mathbf{Z}}_{\mathbf{f}^N}^N(u), \mathbf{H}^N(u)) du$$

is \mathbb{C} -tight for every $t \geq 0$, $i = 1, 2, \dots, L$. Due to the acceleration of the rate matrix $N^\beta \mathcal{Q}$ of the jump process \mathbf{H}^N , the entity $\int_0^t \mathcal{A}^N(s) N^{-\beta/2} d\theta_{g_i}^N(s, \widehat{\mathbf{Z}}_{\mathbf{f}^N}^N(s), \mathbf{H}^N(s))$ turns out to be $\mathcal{O}(N^{\beta/2})$ by definition (5.7), which does not converge as N grows large for any \mathcal{G}^N -predictable process \mathcal{A}^N with $|\mathcal{A}^N| \leq 1$. In the same analogy, the entity $\int_0^t \mathcal{A}^N(s) N^{-\beta} d\phi_{g_i}^N(s, \widehat{\mathbf{Z}}_{\mathbf{f}^N}^N(s), \mathbf{H}^N(s))$ is \mathbb{C} -tight in \mathbb{R} for any $t \geq 0$. The last integral process in the representation (A.2) is also P-UT as a result of

boundedness of the entities $\mathcal{Q}_x \phi_{g_i}^N(t, \mathbf{z}, \mathbf{h}), \alpha_{g_i}^N(t, \mathbf{z}, \mathbf{h}), \Upsilon_{g_i}^N(t, \mathbf{z}, \mathbf{h})$. Therefore, due to the presence of $N^{-\beta/2} \theta_{g_i}^N(t, \mathbf{z}, \mathbf{h})$ term in the expression (A.2), the family $\{\widehat{\mathbf{Z}}_{\mathbf{f}^N}^N, N \in \mathbb{N}\}$ is not P-UT. Whereas using the similar analogy, the family $\{\overline{\mathbf{Z}}_{\mathbf{f}^N}^N, N \in \mathbb{N}\}$ satisfy the P-UT criterion since the semimartingale decomposition of $\overline{\mathbf{Z}}_{\mathbf{f}^N}^N$ in view of (A.2) does not admit the term $N^{-\beta/2} \theta_{g_i}^N(t, \mathbf{z}, \mathbf{h})$.

Proof of Theorem 5.8: In the earlier paragraph, we already discussed that the family $\{\widehat{\mathbf{Z}}_{\mathbf{f}^N}^N, N \in \mathbb{N}\}$ does not satisfy the necessary P-UT criterion for given $\mathbf{f}^N(\cdot) = (\lambda_i^N(\cdot))_{i=1}^L, (\gamma_i^N(\cdot))_{i=1}^L, (\mu_i^N(\cdot))_{i=1}^L$ in order to conclude the assertion of the above theorem. However, if we can approximate the semimartingale $(\widehat{\mathbf{Z}}_{\lambda^N}^N, \widehat{\mathbf{Z}}_{\gamma^N}^N, \widehat{\mathbf{Z}}_{\mu^N}^N)$ by another semimartingale $\{(\mathbf{Y}_{\lambda^N}^N(t), \mathbf{Y}_{\gamma^N}^N(t), \mathbf{Y}_{\mu^N}^N(t)), t \geq 0\}$ with trajectories in $\mathbb{D}([0, \infty), \mathbb{R}^{3L})$ satisfying the P-UT criterion in such a way that $\sup_{0 \leq t \leq T} \|\widehat{\mathbf{Z}}_{\mathbf{f}^N}^N(t) - \mathbf{Y}_{\mathbf{f}^N}^N(t)\| \xrightarrow{\mathbb{P}} 0$ as $N \rightarrow \infty$ and converging in distribution to the same limiting process $(\mathbf{W}_\lambda, \mathbf{W}_\gamma, \mathbf{W}_\mu)$ as $(\widehat{\mathbf{Z}}_{\lambda^N}^N, \widehat{\mathbf{Z}}_{\gamma^N}^N, \widehat{\mathbf{Z}}_{\mu^N}^N)$ does, the requisite convergence will be ensured by Theorem VI.6.22 of Jacod and Shiryaev (2003).

Analogous to the previous proof, given $\mathbf{f}^N(\cdot)$, we take $\mathbf{A}^N(t, \mathbf{z}, \mathbf{h}) = (\mathbf{f}_i^N(t)^T (\mathbf{h} - \boldsymbol{\pi}))_{i=1}^L$, and $\mathbf{B}^N(t, \mathbf{z}, \mathbf{h}) = \mathbf{0}$ in (5.1). Considering $g(\mathbf{z}) = z_i, i = 1, 2, \dots, L$, in the definitions (5.7), (5.8), and (5.9), we deduce the \mathbb{R}^L -valued functions

$$\boldsymbol{\theta}_z^N(t, \mathbf{z}, \mathbf{h}) = \left(\sum_v \mathcal{D}(\mathbf{h}, \mathbf{v}) (\mathbf{f}_i^N(t))^T (\mathbf{v} - \boldsymbol{\pi}) \right)_{i=1}^L, \quad \alpha_z^N(t, \mathbf{z}, \mathbf{h}) = \phi_z^N(t, \mathbf{z}, \mathbf{h}) = \mathbf{0}.$$

Therefore, the semimartingale representation of $\widehat{\mathbf{Z}}_{\mathbf{f}^N}^N$ in view of (5.16) yields

$$\begin{aligned} \widehat{\mathbf{Z}}_{\mathbf{f}^N}^N(t) &= \mathbf{M}_z^N(t) - N^{-\beta/2} \left(\boldsymbol{\theta}_z(t, \widehat{\mathbf{Z}}_{\mathbf{f}^N}^N(t), \mathbf{H}^N(t)) - \boldsymbol{\theta}_z(0, \mathbf{0}, \mathbf{h}) \right) \\ &\quad + N^{-\beta/2} \int_0^t \frac{\partial}{\partial s} \boldsymbol{\theta}_z(s, \widehat{\mathbf{Z}}_{\mathbf{f}^N}^N(s), \mathbf{H}^N(s)) ds, \end{aligned} \tag{A.3}$$

with $\widehat{\mathbf{Z}}_{\mathbf{f}^N}^N(0) = \mathbf{0}$, where \mathbf{M}_z^N is an \mathbb{R}^L -valued \mathcal{G}^N -square-integrable martingale corresponding to the vector valued function $\mathbf{g}^N(t, \mathbf{z}, \mathbf{h}) = \mathbf{z} + N^{-\beta/2} \boldsymbol{\theta}_z^N(t, \mathbf{z}, \mathbf{h})$, analogous to the constructions defined in the proof of Theorem 5.1. Here, every (i, j) th component of $\mathbb{M}^{L \times L}$ -valued predictable quadratic variation process associated to the martingale \mathbf{M}_z^N is given by

$$\langle M_{z_i}^N, M_{z_j}^N \rangle (t) = \int_0^t \Gamma_{g_i^N, g_j^N}^N(u, \widehat{\mathbf{Z}}_{\mathbf{f}^N}^N(u), \mathbf{H}^N(u)) du,$$

where

$$\Gamma_{g_i^N, g_j^N}^N(t, \mathbf{z}, \mathbf{h}) = \mathcal{L}_t^N(g_i^N \cdot g_j^N)(t, \mathbf{z}, \mathbf{h}) - g_i^N(t, \mathbf{z}, \mathbf{h}) \mathcal{L}_t^N g_j^N(t, \mathbf{z}, \mathbf{h}) - g_j^N(t, \mathbf{z}, \mathbf{h}) \mathcal{L}_t^N g_i^N(t, \mathbf{z}, \mathbf{h}),$$

with $\mathcal{L}_t^N = N^\beta \mathcal{Q} + N^{\beta/2} \mathbf{A}^N(t, \mathbf{z}, \mathbf{h}) \cdot \nabla_z$. Using the relations (5.13) and (5.14), it is easy to verify that, for each $i, j = 1, 2, \dots, L$,

$$\begin{aligned} \mathcal{L}_t^N g_i^N(t, \mathbf{z}, \mathbf{h}) &= 0, \\ \mathcal{L}_t^N (g_i^N \cdot g_j^N)(t, \mathbf{z}, \mathbf{h}) &= \mathcal{Q} \left(\theta_{z_i}^N(t, \mathbf{z}, \mathbf{h}) \cdot \theta_{z_j}^N(t, \mathbf{z}, \mathbf{h}) \right) + \mathbf{A}^N(t, \mathbf{z}, \mathbf{h}) \cdot \nabla_z (z_i \theta_{z_j}^N(t, \mathbf{z}, \mathbf{h}) + z_j \theta_{z_i}^N(t, \mathbf{z}, \mathbf{h})). \end{aligned} \tag{A.4}$$

Note that, the entity $\theta_{z_i}^N(t, \mathbf{z}, \mathbf{h})$ is i th component of the vector $\boldsymbol{\theta}_z^N(t, \mathbf{z}, \mathbf{h})$, and independent of \mathbf{z} variable, and hence applying Theorem 9.6.1 in Liptser and Shiryaev (1989) followed by (A.4) and Assumption 3.3 we obtain

$$\lim_{N \rightarrow \infty} \left| \int_0^t \mathcal{L}_u^N (g_i^N \cdot g_j^N) \left(u, \widehat{\mathbf{Z}}_{\mathbf{f}^N}^N(u), \mathbf{H}^N(u) \right) du - \int_0^t \overline{A}_{i,j}^\pi \left(u, \widehat{\mathbf{Z}}_{\mathbf{f}^N}^N(u) \right) du \right| = 0, \quad \mathbb{P}_\pi - \text{a.s.},$$

for every $t \geq 0$. Here, the first term on the right hand side of (A.4) asymptotically vanishes to zero after taking average with respect to the invariant distribution π . As a result, the second term of (A.4) eventually contributes to the limit $\bar{A}_{i,j}^\pi(t, z) = \sum_{\mathbf{v}} \pi(\mathbf{v}) \sum_{\mathbf{y}} \mathcal{D}(\mathbf{v}, \mathbf{y}) (A_i(t, z, \mathbf{v}) A_j(t, z, \mathbf{y}) + A_j(t, z, \mathbf{v}) A_i(t, z, \mathbf{y}))$, and consequently, $\mathbf{A}(t, z, \mathbf{h}) = ((\mathbf{f}_i(t))^T (\mathbf{h} - \boldsymbol{\pi}))_{i=1}^L$ results $\bar{A}_{i,j}^\pi(t, z) = \sum_{k=1}^d \sum_{l=1}^d \pi_k \mathcal{D}_{kl} \int_0^t (f_{ik}(s) f_{jl}(s) + f_{il}(s) f_{jk}(s)) ds$, which is the (i, j) th element of the covariance matrix $\mathbf{C}_f(t)$ of the limiting process \mathbf{W}_f according to Theorem 5.6, given $\mathbf{f}(\cdot) = \boldsymbol{\lambda}(\cdot), \boldsymbol{\gamma}(\cdot), \boldsymbol{\mu}_i(\cdot)$.

In the next step, choose $\mathbf{Y}_{f_N}^N(t) = \mathbf{M}_z^N(t), t \geq 0$. Clearly, by definition $\sup_{0 \leq t \leq T} \|\hat{\mathbf{Z}}_{f_N}^N(t) - \mathbf{Y}_{f_N}^N(t)\| \xrightarrow{\mathbb{P}} 0$ as $N \rightarrow \infty$. Also, the family $\{\mathbf{Y}_{f_N}^N, N \in \mathbb{N}\}$ is P-UT by reason of Theorem VI.6.13 in Jacod and Shiryaev (2003). In order to prove the assertion it suffices to show that $\mathbf{Y}_{f_N}^N$ converges in distribution to the Gaussian process \mathbf{W}_f , as $N \rightarrow \infty$, and as an extensive consequence $(\mathbf{Y}_{\lambda_N}^N, \mathbf{Y}_{\gamma_N}^N, \mathbf{Y}_{\mu_N}^N)$ converges in distribution to $(\mathbf{W}_\lambda, \mathbf{W}_\gamma, \mathbf{W}_\mu)$ in $\mathbb{D}([0, \infty), \mathbb{R}^{3L})$. It is already shown that the predictable entity $\langle \mathbf{Y}_{f_i}^N, \mathbf{Y}_{f_j}^N \rangle(t)$ converges in probability to $\int_0^t \bar{A}_{i,j}^\pi(s) ds$ for each $t \geq 0$ as $N \rightarrow \infty$. Additionally, the jumps of $\mathbf{Y}_{f_N}^N$ is not greater than $N^{-\beta/2}$, and the latter quantity vanishes to zero eventually. Therefore, by the classical martingale central limit theorem we obtain $\mathbf{Y}_{f_N}^N \Rightarrow \mathbf{W}_f$ as $N \rightarrow \infty$. The proof is complete. \square

Proof of Proposition 8.2: First we begin with the deterministic convergence of the joint process $(\hat{\mathbf{R}}_2^{N,E}, \hat{\mathbf{R}}_2^{N,A}, \hat{\mathbf{R}}_3^{N,A}, \hat{\mathbf{R}}_2^{N,S}, \hat{\mathbf{R}}_3^{N,S})$ to $(\int_0^\cdot \hat{\boldsymbol{\lambda}}^\pi(s) ds, \hat{\mathbf{R}}_2^A, \hat{\mathbf{R}}_3^A, \hat{\mathbf{R}}_2^S, \hat{\mathbf{R}}_3^S)$, which are direct consequence of Assumptions 3.3–3.6. It remains to prove the Skorokhod convergence of $(\hat{\mathbf{R}}_1^{N,E}, \hat{\mathbf{R}}_1^{N,A}, \hat{\mathbf{R}}_1^{N,S})$. To prove this joint convergence in the Skorokhod product space, we define the matrices

$$\begin{aligned} \boldsymbol{\kappa}_1^E(t) &= \text{diag}(\mathbf{1}), \boldsymbol{\kappa}_1^{N,A}(t) = -(\mathbf{I} - (\bar{\boldsymbol{\psi}}^N(\bar{\mathbf{Q}}^N(t)))^T) \text{diag}(\bar{\mathbf{Q}}(t) - \bar{\mathbf{K}}(t))^+, \\ \boldsymbol{\kappa}_1^{N,S}(t) &= -(\mathbf{I} - (\bar{\boldsymbol{\phi}}^N(\bar{\mathbf{Q}}^N(t)))^T) \text{diag}(\bar{\mathbf{Q}}(t) \wedge \bar{\mathbf{K}}(t)). \end{aligned}$$

Each $\mathbb{M}^{L \times L}$ -valued random process $\boldsymbol{\kappa}_1^E, \boldsymbol{\kappa}_1^{N,A}, \boldsymbol{\kappa}_1^{N,S}$ is \mathcal{F}^N -predictable locally bounded process, and by the given hypotheses and Theorem 3.2, it, respectively, converges in probability to $\boldsymbol{\kappa}_1^E, -(\mathbf{I} - (\bar{\boldsymbol{\psi}}(\bar{\mathbf{Q}}(\cdot)))^T) \text{diag}(\bar{\mathbf{Q}}(\cdot) - \bar{\mathbf{K}}(\cdot))^+, -(\mathbf{I} - (\bar{\boldsymbol{\phi}}(\bar{\mathbf{Q}}(\cdot)))^T) \text{diag}(\bar{\mathbf{Q}}(\cdot) \wedge \bar{\mathbf{K}}(\cdot))$ uniformly on compact sets, as $N \rightarrow \infty$.

It follows from Theorem 5.6 and the relation in (8.8) that the square-integrable martingale $(\mathbf{V}_{\lambda_N}^N, \mathbf{V}_{\gamma_N}^N, \mathbf{V}_{\mu_N}^N)$ converges in distribution to $(\mathbf{W}_\lambda, \mathbf{W}_\gamma, \mathbf{W}_\mu)$ in $\mathbb{D}([0, \infty), \mathbb{R}^{3L})$ as $N \rightarrow \infty$. Moreover, it satisfy the P-UT condition. Knowing that $1 - \delta - \beta/2 = \min\{0, (1 - \beta)/2\}$, taking $\boldsymbol{\kappa}^N(t)$ as $\boldsymbol{\kappa}_1^E(t) \oplus \boldsymbol{\kappa}_1^{N,A}(t) \oplus \boldsymbol{\kappa}_1^{N,S}(t)$ in $\mathbb{D}([0, \infty), \mathbb{M}^{3L \times 3L})$, it follows from Theorem 5.8 that $\boldsymbol{\kappa}_-^N \cdot (\mathbf{V}_{\lambda_N}^N, \mathbf{V}_{\gamma_N}^N, \mathbf{V}_{\mu_N}^N)$ jointly converges in distribution to $\mathbf{1}(\beta \leq 1)(\hat{\mathbf{R}}_1^E, \hat{\mathbf{R}}_1^A, \hat{\mathbf{R}}_1^S)$ in $\mathbb{D}([0, \infty), \mathbb{R}^{3L})$. The convergence (8.8) together with the previous deterministic joint convergence complete the proof. \square

Acknowledgements

The authors would like to thank the anonymous referee and the Associate Editor for their careful reading and detailed comments, which have immensely helped to improve the paper.

References

Anderson, D., Blom, J., Mandjes, M., Thorsdottir, H., and de Turck, K. A functional central limit theorem for a Markov-modulated infinite-server queue. *Methodol. Comput. Appl. Probab.*, **18** (1), 153–168 (2016). DOI: [10.1007/s11009-014-9405-8](https://doi.org/10.1007/s11009-014-9405-8).

- Arapostathis, A., Das, A., Pang, G., and Zheng, Y. Optimal control of Markov-modulated multiclass many-server queues. *Stoch. Syst.*, **9** (2), 155–181 (2019). DOI: [10.1287/stsy.2019.0029](https://doi.org/10.1287/stsy.2019.0029).
- Arapostathis, A., Pang, G., and Zheng, Y. Exponential ergodicity and steady-state approximations for a class of Markov processes under fast regime switching. *Adv. in Appl. Probab.*, **53** (1), 1–29 (2021). DOI: [10.1017/apr.2020.47](https://doi.org/10.1017/apr.2020.47).
- Budhiraja, A., Dupuis, P., and Ganguly, A. Large deviations for small noise diffusions in a fast markovian environment. *Electron. J. Probab.*, **23**, Paper No. 112, 33 (2018). DOI: [10.1214/18-ejp228](https://doi.org/10.1214/18-ejp228).
- Budhiraja, A. and Liu, X. Multiscale diffusion approximations for stochastic networks in heavy traffic. *Stochastic Process. Appl.*, **121** (3), 630–656 (2011). DOI: [10.1016/j.spa.2010.10.009](https://doi.org/10.1016/j.spa.2010.10.009).
- Ethier, S. N. and Kurtz, T. G. *Markov processes. Characterization and convergence*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York (1986). ISBN 0-471-08186-8. DOI: [10.1002/9780470316658](https://doi.org/10.1002/9780470316658).
- Halfin, S. and Whitt, W. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.*, **29** (3), 567–588 (1981). DOI: [10.1287/opre.29.3.567](https://doi.org/10.1287/opre.29.3.567).
- Jacod, J. and Shiryaev, A. N. *Limit theorems for stochastic processes*, volume 288 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, second edition (2003). ISBN 3-540-43932-3. DOI: [10.1007/978-3-662-05265-5](https://doi.org/10.1007/978-3-662-05265-5).
- Jansen, H., Mandjes, M., De Turck, K., and Wittevrongel, S. Diffusion limits for networks of Markov-modulated infinite-server queues. *Perform. Eval.*, **135**, 102039 (2019). DOI: [10.1016/j.peva.2019.102039](https://doi.org/10.1016/j.peva.2019.102039).
- Kurtz, T. G. Averaging for martingale problems and stochastic approximation. In *Applied stochastic analysis (New Brunswick, NJ, 1991)*, volume 177 of *Lect. Notes Control Inf. Sci.*, pp. 186–209. Springer, Berlin (1992). ISBN 3-540-55296-0. DOI: [10.1007/BFb0007058](https://doi.org/10.1007/BFb0007058).
- Liptser, R. S. and Shiryaev, A. N. *Theory of martingales*, volume 49 of *Mathematics and its Applications (Soviet Series)*. Kluwer Academic Publishers Group, Dordrecht (1989). ISBN 0-7923-0395-4. DOI: [10.1007/978-94-009-2438-3](https://doi.org/10.1007/978-94-009-2438-3).
- Liu, Y. and Whitt, W. The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Syst.*, **71** (4), 405–444 (2012). DOI: [10.1007/s11134-012-9291-0](https://doi.org/10.1007/s11134-012-9291-0).
- Mandelbaum, A., Massey, W. A., and Reiman, M. I. Strong approximations for Markovian service networks. *Queueing Systems Theory Appl.*, **30** (1-2), 149–201 (1998). DOI: [10.1023/A:1019112920622](https://doi.org/10.1023/A:1019112920622).
- Pang, G., Talreja, R., and Whitt, W. Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probab. Surv.*, **4**, 193–267 (2007). DOI: [10.1214/06-PS091](https://doi.org/10.1214/06-PS091).
- Pang, G. and Yao, D. D. Heavy-traffic limits for a many-server queueing network with switchover. *Adv. in Appl. Probab.*, **45** (3), 645–672 (2013). DOI: [10.1239/aap/1377868533](https://doi.org/10.1239/aap/1377868533).
- Papanicolaou, G. C. Asymptotic analysis of transport processes. *Bull. Amer. Math. Soc.*, **81**, 330–392 (1975). DOI: [10.1090/S0002-9904-1975-13744-X](https://doi.org/10.1090/S0002-9904-1975-13744-X).
- Papanicolaou, G. C., Stroock, D., and Varadhan, S. R. S. Martingale approach to some limit theorems. In *Proc. 1976 Duke Conf. on Turbulence*, Duke Univ. Math. Ser., Vol. III, pp. Paper No. 6, ii+120 (1977). MR461684.
- Puhalskii, A. A. On the $M_t/M_t/K_t + M_t$ queue in heavy traffic. *Math. Methods Oper. Res.*, **78** (1), 119–148 (2013). DOI: [10.1007/s00186-013-0435-8](https://doi.org/10.1007/s00186-013-0435-8).
- Reiman, M. I. Open queueing networks in heavy traffic. *Math. Oper. Res.*, **9** (3), 441–458 (1984). DOI: [10.1287/moor.9.3.441](https://doi.org/10.1287/moor.9.3.441).
- Sen, A. and Selvaraju, N. Diffusion approximation of an infinite-server queue under Markovian environment with rapid switching. *Statist. Probab. Lett.*, **195**, Paper No. 109778, 11 (2023). DOI: [10.1016/j.spl.2023.109778](https://doi.org/10.1016/j.spl.2023.109778).
- Skorokhod, A. V. *Asymptotic methods in the theory of stochastic differential equations*, volume 78 of *Translations of Mathematical Monographs* (1989). DOI: [10.1090/mmono/078](https://doi.org/10.1090/mmono/078).

-
- Stroock, D. W. and Varadhan, S. R. S. *Multidimensional diffusion processes*. Classics in Mathematics. Springer-Verlag, Berlin, 1997 edition (2006). ISBN 978-3-540-28998-2; 3-540-28998-4. [MR2190038](#).
- Sun, X. and Whitt, W. Delay-based service differentiation with many servers and time-varying arrival rates. *Stoch. Syst.*, **8** (3), 230–263 (2018). DOI: [10.1287/stsy.2018.0015](https://doi.org/10.1287/stsy.2018.0015).
- Whitt, W. Time-Varying Queues. *QMSM*, **1** (2), 79–164 (2018). <https://par.nsf.gov/biblio/10120248>.