Infinite-Dimensional Scaling Limits of Stochastic Networks

by

Mohammadreza Aghajani

B. S., Sharif University of Technology, 2008

M.S., Carniegie-Mellon University, 2010

Sc. M., Brown University, 2011

A Dissertation submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in the Division of Applied Mathematics at Brown University

Providence, Rhode Island

May 2016

This dissertation by Mohammadreza Aghajani is accepted in its present form
by the Division of Applied Mathematics as satisfying the
dissertation requirement for the degree of Doctor of Philosophy.

Date ______________ ____

Kavita Ramanan, Director

Recommended to the Graduate Council

Date ______________ ____

Haya Kaspi, Reader

Date ______________ ____

Philippe Robert, Reader

Approved by the Graduate Council

Date ______________ ____

Peter M. Weber, Dean of the Graduate School

Vitæ

Mohammadreza (Reza) Aghajani was born on September 19, 1986, in Mashhad, Iran. After finishing high school in Mashhad, he successfully passed Iran's national university entrance exam, where he ranked 48th among more than 400000 participants. He attended Sharif University of Technology in Tehran, and as an exceptional talent, he was permitted to study in two majors. He received his B.Sc. in Electrical Engineering and B.Sc. in Mathematics, in 2008. He then pursued his graduate studies abroad and attended Carnegie Mellon University in Pittsburgh, PA, where he received his M.Sc. in Electrical and Computer engineering in 2010, under supervision of Prof. Bruno Sinopoli. He started his Ph.D. in the Division of Applied Mathematics at Brown University in September 2010. After finishing his course work and receiving a M.Sc. in Applied Mathematics in 2010, he started his research on scaling limits of stochastic networks, under supervision of Prof. Kavita Ramanan. He has presented his results and contributions in several scientific conferences. Reza has received the Sigma Xi Award for excellence in research and high potential for future contributions, and the Dunmu Ji Award in recognition of his particularly original and independent thesis.

Acknowledgements

First, I would like to thank my advisor, Professor Kavita Ramanan, who changed the course of my scientific life. I am very grateful to her for all her help, support, and guidance over the last seven years, and for all the things that I learnt from her, which are beyond just mathematical theorems.

Then I would like to thank my committee members, Professor Philippe Robert and Professor Haya Kaspi, for traveling thousands of miles to attend my defense, for their constructive feedback, and for fruitful discussions. In addition, I would like to thank Philippe for his hospitality during my visit to INRIA, Paris-Rocquencourt in summer 2014.

I would also like to thank all my teachers during the last 24 years. I particularly thank my high school Physics teacher, Prof. Miri, who introduced me to the world of science, and I wish I would have listened to his advises more carefully. I am also grateful to my college Mathematics teachers, Prof. Kambiz Mahmoudian and Prof. Reza Moghadasi, who allured me into the beauty of Mathematics. I would also like to thank my M.Sc. supervisor Prof. Bruno Sinopoli, who supported me when I needed it the most.

Next, I would like to thank all the members of the Division of Applied mathematics, who made my Ph.D. studies here a pleasant experience. In particular, I thank Helen Li, with whom I had a wonderful collaboration, and Steven Kim, for many intellectual discussions, and for telling me what O.G. stands for. I am also grateful to Hanie Mirzaei, Delaram Motamed Vaziri and Mohammad Moghadasi for their hospitality and support upon my arrival to Providence, and to my Iranian friends in Rhode Island at the time, Amir Banari, Telli Davoodi, Mohsen Zayernoori, Akbar Bagri, Sobhan Naderi, Tayebe Tajali Bakhsh, Sara Kadkhodaei, Maryam Jouzi and Gelare Azadi, for all the good memories and their invaluable support. I specially thank Nakisa Mohammadifard, for her great support and for making everything so much more fun.

Finally, I would like to thank my two lovely sisters, Toktam and Aida, and my precious mom, for her unconditional support, and for making all these possible through her many sacrifices, and it is to her that I dedicate this thesis.

Contents

List of Tables

List of Figures

Abstract of "Infinite-Dimensional Scaling Limits of Stochastic Networks"

by Mohammadreza Aghajani, Ph.D., Brown University, May 2016

Large-scale stochastic networks arise in a variety of real world applications ranging from telecommunications, service systems, and computer networks to health care services and biological systems. Such networks are typically too complex and not amenable to exact analysis. However, it is often possible to provide qualitative and quantitative insight into network performance, in both the transient and equilibrium regimes, through tractable approximations.

Finite-dimensional Markov processes have been used extensively in the literature to describe the scaling limits of stochastic networks under simplifying assumptions such as exponential service distributions. However, statistical analysis has shown that the service distribution is typically non-exponential. For networks with general service distributions, the dimension of the system descriptor typically grows with size of the network, and hence any scaling limit is inherently infinite dimensional. Thus, a significant challenge is to find a suitable representation of the network that leads to a tractable scaling limit.

The focus of this thesis is on developing a mathematical framework for the analysis of infinite-dimensional scaling limits of many-server stochastic networks. Specifically, we have focused on two rather different problems for two kinds of large-scale stochastic networks – a many-server queueing model (also known as the "$GI/GI/N$" model) and a randomized load balancing model, both in the presence of general service distributions.

For the many-server queueing model, we show that the limit of normalized fluctuations of the network can be captured by a non-standard Stochastic Partial Differential Equation (SPDE). We also establish ergodicity of the solution to this SPDE via a novel construction of an asymptotic (equivalent) coupling,

as traditional Harris recurrent techniques are often inapplicable in the infinite-dimensional setting. Our result resolves an open problem on a many-server queueing model originally raised by Halfin and Whitt in 1981. For the randomized load-balancing network, we introduce a new representation using interacting measure-valued processes, and show that the mean transient behavior of the network can be approximated by a countably infinite coupled system of partial integro-differential equations (PDEs). This provides a significant extension of the ODE method used in the case of an exponential service distribution.

CHAPTER 1

Introduction

1.1. Large-Scale Stochastic Networks

Stochastic networks arise in a variety of real world applications ranging from telecommunications, service systems such as call centers, and computer networks to health care services and biological systems. Such networks are typically too complex and not amenable to exact analysis. However, it is often possible to provide insight into network performance, in both the transient and equilibrium regimes, through tractable approximations. The accuracy of these approximations in a suitable network parameter regime can be rigorously justified through limit theorems.

Large-scale many server networks are of particular importance due to the surge in applications such as cloud computing, server farms and large-scale distributed storage systems. A natural approximation is to consider the asymptotics of these networks as the system size grows to infinity. This approximation can be carried out to various levels of accuracy: hydrodynamic (or fluid) limits approximate the mean behavior of the network by a deterministic evolution equation, while diffusion limits capture fluctuations of the state of the network around its mean behavior.

Until recently, scaling limits of many-server networks were established mostly under the assumption of exponential service distributions, where finite-dimensional Markovian representations are readily available for both the original network state process and its scaling limit. However, statistical analysis has shown that the service distribution is typically non-exponential [**16, 21, 72, 67**]. For networks with general service distributions, the dimension of the Markovian system descriptor grows with size, and hence any scaling limit is inherently infinite dimensional.

Thus, finding a suitable representation of the network that leads to a tractable scaling limit is a significant challenge, and analysis of the infinite-dimensional limit process requires new techniques.

This thesis has focused on developing a mathematical framework for analysis of infinite-dimensional scaling limits of many-server stochastic networks. Specifically, two rather different problems for two kinds of large-scale stochastic networks are studied: a many-server queueing model (also known as the "$GI/GI/N$" model) and a randomized load balancing model, both in the presence of general service distributions. In both cases, it turns out that a measure-valued or other infinite-dimensional state representation plays a key role. Measure-valued processes have also proved useful for the study of single-server networks with non head-of-the-line policies, including earliest-deadline-first, processor sharing or shortest-remaining-processing-time. Nevertheless, in almost all of the above, a sort of "state space collapse" occurs in heavy traffic, and the limiting infinite-dimensional diffusion limit can be expressed as the deterministic mapping of a one-dimensional stochastic process as, for example, in [**29, 70, 71, 45, 46, 63**]. However, as discussed in Part I, in the heavy-traffic approximation we established for the $GI/GI/N$ queue, the limiting diffusion process is truly infinite-dimensional. On the other hand, previous fluid limits have mainly been described by a single measure-valued process, or a coupled pair of measure-valued processes. But for the randomized load balancing model described in Part II, the fluid limit is an infinite sequence of interacting measure-valued processes. This leads to a new set of challenges, and is discussed in Part II.

1.2. The $GI/GI/N$ Queueing System in the Halfin-Whitt Regime

1.2.1. Description of Problem. Parallel server queues are powerful and flexible models which have been used in a great diversity of applications, including call centers [**2, 41, 103, 42**], health care [**7, 19**], and transportation [**77, 92**]. A standard model in queueing theory is the so-called $GI/GI/N$ queue, which consists of a network of N parallel servers to which a common stream of jobs arrive according to a renewal process. Jobs have independent and identically distributed (i.i.d.) service times, and, if all servers are busy on arrival, jobs wait in a common queue to be processed in a First-Come-First-Serve (FCFS) manner by the next available server (we refer the reader to [**3**] for a more detailed description of this model). Motivated by applications in call centers, data centers and health care, a particular focus in recent years has been on characterizing steady state quantities such as, for example, the mean steady-state number of jobs in system $X^{(N)}$, and the steady state probability P_c, of the wait to exceed a certain threshold $c \geq 0$. Since an exact characterization of this quantity appears infeasible, the goal has rather been to obtain provably good approximations to this quantity, which are accurate in the limit as the number of servers N goes to infinity.

In many applications such as call centers, the tradeoff between the quality of service and the cost of providing the service is at a desirable balance when the network operates in the Quality- and Efficiency-Driven (Q.E.D.) regime, that is, when the steady state probability that a job needs to wait for service is bounded away from both 0 and 1. A design question is how the number of servers should be scaled with the arrival rate, as the network size grows, in order for the network to operate in the Q.E.D. regime. For the case of exponential service time distribution, this problem is solved in the seminal work of Halfin and Whitt [**51**]. In particular, they studied a sequence of $GI/M/N$ systems in which the mean arrival rate has the form $\lambda_N = N - \beta\sqrt{N} + o(\sqrt{N})$ for some $\beta > 0$, and showed that the sequence of centered and renormalized processes $\widehat{X}^{(N)} = (X^{(N)} - N)/\sqrt{N}$ converges weakly on finite time intervals to a positive recurrent diffusion X with a constant negative drift when $X > 0$ and an Ornstein-Uhlenbeck type restoring drift when $X < 0$. Moreover, they also showed [**51**, Proposition 1 and Corollary 2] that, as the number of servers N goes to infinity, the invariant distribution of $\widehat{X}^{(N)}$ converges to the (unique) invariant distribution of the diffusion X. Since the exact form of this distribution can be easily identified, this provides a useful and explicit approximation for the steady state probability of a positive waiting time in an N-server network, which is indeed bounded away from 0 and 1.

However, as mentioned above, statistical analysis has shown that the service distribution is typically non-exponential [**16**]. An analogous result for generally distributed service times was posed as an open problem in [**51**], which remained unsolved except for a few very specific distribution families [**101, 56, 38**].

In analogy with the exponential case, a natural conjecture would be that this limit is equal to the unique stationary distribution of the process X obtained as the limit (over finite intervals) of the scaled processes $\widehat{X}^{(N)}$, assuming this can be shown to exist. However, while for exponential service distributions both the process $\widehat{X}^{(N)}$ and its limit X are Markovian, this is no longer true for more general service distributions. Indeed, although convergence (over finite time intervals) of the sequence of scaled processes $\widehat{X}^{(N)}$ has been established for various classes of service distributions (see, e.g., [**86, 76, 38, 62**]), the obtained limit is not Markovian, with the exception of the work in [**62**]. This makes characterization of the stationary distribution of X challenging. It is easy to see that, except for special classes of distributions (e.g., phase-type distributions, as considered in [**86**]) any Markovian limit process will be infinite-dimensional. One key challenge is then to find a suitable representation of the limit process that leads to a tractable Markovian process whose invariant distribution can be analyzed.

On the other hand, analysis of the diffusion limit requires new techniques. Due to the infinite-dimensional nature of the limit Markov process, standard methods for establishing ergodicity of finite-dimensional Markov processes such as positive Harris recurrence are not well suited to this setting. Moreover, the equations governing the dynamics of the limit process turn out, in many cases, to be non-standard (in particular, due to the presences of non-linear boundary conditions), and hence, other techniques such as the dissipativity method used for studying ergodicity of nonlinear Stochastic Partial Differential Equations (SPDEs) (see, e.g.

[**24, 78**]) also appear not immediately applicable. Another key challenge is therefore to invoke the appropriate techniques to establish ergodicity for the limit process. Both of these challenges are addressed in this thesis.

1.2.2. Prior Work. Study of performance of parallel server queueing models has a rich history in Applied Probability, Operations Research, and Communications Theory literature, among other related fields (see [**22, 53**] for a survey), and goes back to the pioneering work of Erlang [**31**] and Pollaczek [**84**]. The $GI/GI/N$ model is also widely studied (see references in [**99**]), and in particular, its stability is proved in [**65**], under the condition that the traffic intensity ρ is strictly less than one. A summary of result on the $GI/GI/N$ model can also be found in [**3**].

The aforementioned question of the "right" scaling of the arrival rate with to the number of servers for a network in Q.E.D. regime was first studied by Erlang [**32**] and Jagerman [**54**]. As stated, the solution for the case of exponential service distribution is given by Halfin and Whitt [**51**], where they also established an approximation of the steady state distribution of $\widehat{X}^{(N)}$, by the invariant distribution of a diffusion process $\widehat{X}$. The problem of obtaining an analogous approximation for general service distributions was posed as an open problem in [**51**, Section 4], and has remained unsolved except for a few specific distributions.

The convergence, over finite time, of the sequence of scaled process $\widehat{X}^{(N)}$ has been established for various classes of service distributions (see, e.g., [**86, 76, 38, 62**]), with the most general results obtained by Puhalskii and Reed in [**87, 85**]. The obtained limit is not Markovian, with the exception of the work in [**62**]. In contrast, there are relatively few studies of diffusion limits of the steady-state distribution of many-server networks (i.e. on infinite horizon). Whitt [**101**] extended the result in [**51**] to the H_2^* service time distribution (i.e. mixture of an exponential distribution and a point mass at 0). The case of deterministic service times is considered in [**56**], and was generalized in [**38**], to consider the case when the service distribution has finite support. In [**56**] the limit steady state distribution of the scaled number of jobs in system is characterized as the supremum of a sum of independent normal random variables, and in [**38**] the limit steady state distribution is shown to be the stationary distribution to a specific Markov chain. Most recently, Dai et al. [**25**] considered phase-type service distributions. A significant recent advance is the work of Gamarnik and Goldberg [**37**], which uses general bounds for the FCFS $GI/GI/N$ queue to show that when the service distribution G has $2+\epsilon$ finite moments and satisfies other minor technical conditions, the sequence of scaled steady state queue lengths $\{\widehat{Q}^N\}_{N\in\mathbb{N}}$ is tight, and related large deviations bounds are obtained [**37, 43**]. This result is extended to the case of multiple customer classes in [**40**]. However, the missing element in converting the tightness result of [**37**] to a convergence result was the identification and unique characterization of a candidate limit distribution.

In [**61**], a Markovian state representation for the system was introduced in terms of the pair $(X^{(N)},\nu^{(N)})$, where $\nu^{(N)}$ is a finite measure on $[0,\infty)$ that has a unit mass at the age (amount of time spent in service) of each job in service, and a fluid limit was established. This representation proved useful for performance analysis and optimal control of fluid limits of the model and several generalizations (see, e.g., [**57, 58, 4**]), and an associated distribution-valued diffusion-scaled limit was obtained (in the subcritical, critical and supercritical regimes) by Kaspi and Ramanan in [**62**]. However, the limit turns out not to be Markov on its own, unless one imposes stringent assumptions on the service distribution. Instead, it was shown in [**62**] that for a fairly general class of service distributions, this limit could be "Markovianized" with the addition of a third component that takes values in the space of continuous functions on $[0,\infty)$. While this result was useful in that it represented the limit process X as an Itô process with a constant diffusion coefficient (equal to the one obtained for the exponential distribution), it still leads to a three-component distribution-valued Markov process that does not appear to be very tractable.

Diffusion scale Approximation of steady-state distributions have been established in other settings. In the conventional heavy-traffic setting, Gurvich [**48**] systematically generalized the results in [**39**] to multi-class queueing networks. Katsuda [**63**] proves some interchange limit results for a multi-class single-server queue with feedback under various disciplines. Ye and Yao [**102**] studied interchange limit results in a head-of-line bandwidth sharing model with two customer classes and two servers. Moreover, Other works in the simpler setting of infinite-server queues include [**69, 83, 27, 88**]. In particular, Decreusefond and Moyal [**27**] and Reed and Talreja [**88**] obtained diffusion limits that are tempered distribution-valued Ornstein Uhlenbeck processes, and the latter also identified the invariant distribution of the process. Finally, similar heavy traffic approximation results in alternative queuing models have been obtained by various authors. Queueing systems with abandonments are studied by Garnett et al. [**42**] and Zeltyn and Mandelbaum [**103**] in the case of exponentially distribution service times, and by Kang and Ramanan [**57, 58**], among others, in the case of generally distributed service times. Diffusion-scale tightness of invariant distributions in networks with multiple customer classes, multiple server pools, and subcritical load are proved in [**94**]. Also, diffusion approximation for stationary distribution in Jackson networks are obtained in [**17, 39**].

1.2.3. Contributions. Part I of this thesis contains a series of contributions which lead to resolving the aforementioned open problem on diffusion approximation for the $GI/GI/N$ model in Halfin-Whitt regime, originally raised by Halfin and Whitt in [**51**] and posed again by Gamarnik and Goldberg in [**37**]. Although this result is of independent interest, the broader intension is to introduce a new framework for the asymptotic analysis of large-scale stochastic networks, using measure-valued and other infinite-dimensional processes, and to showcase the effectiveness of new techniques in analysis of their scaling limits. Here is a summary of the contributions in Part I.

(1) **A new representation for the $GI/GI/N$ queueing system.** Inspired by a measure-valued representation in **[61]**, we introduce a new infinite-dimensional representation for the $GI/GI/N$ queueing system. Specifically, we consider a representation $(X^{(N)}, Z^{(N)})$, where $X^{(N)}$ is, the number of jobs in the N-server network, and $Z^{(N)}$ is a functional of the measure valued state variable $\nu^{(N)}$ used in **[61]**, which takes values in a Hilbert functional space $\mathbb{H}^1(0,\infty)$ (i.e. the space of absolutely continuous, square integrable functions on $(0,\infty)$ whose density is also square integrable). The key feature of this representation is that the appended variable $Z^{(N)}$ contains just enough information, so the sequence of scaled and renormalized processes $\{\widehat{Y}^{(N)} = (\widehat{X}^{(N)}, \widehat{Z}^{(N)})\}$ converges, in the Halfin-Whitt asymptotic regime, to a infinite-dimensional *Markov* process.

(2) **An SPDE characterization of the limit process.** Next, we identify the diffusion limit process of the sequence $\{\widehat{Y}^{(N)}\}$. Under suitable conditions on the service time distribution, we characterize the components X and Z of the limit process to be the unique solution of a coupled pair of stochastic equations driven by a Brownian motion and an independent space-time white noise. Specifically, X satisfies an Itô equation with a constant diffusion coefficient and a drift that depends on Z, while Z is an $\mathbb{H}^1(0,\infty)$-valued process driven by a space-time white noise that satisfies a (non-standard) stochastic partial differential equation (SPDE) on the domain $(0,\infty)$. The two components are further linked by a nonlinear boundary condition: at each time $t \geq 0$, $Z(t,0)$ is a (deterministic) nonlinear function of $X(t)$ (See Definition 3.1). This SPDE characterization of $(X, Z(\cdot))$ facilitates the use of tools from stochastic calculus to compute performance measures of interest and natural generalizations would potentially also be useful for studying diffusion control problems for many-server networks, in a manner analogous to Brownian control problems that have been studied in the context of single-server networks **[52]**.

(3) **Ergodicity of the infinite-dimensional limit process.** Our next contribution is to establish the uniqueness of the invariant distribution of the limit (Markov) process $Y = (X, Z)$ (see Theorem 3.26). As mentioned above, standard methods for establishing ergodicity of finite-dimensional Markov processes such as positive Harris recurrence are not well suited to this setting due to the infinite-dimensional nature of the state space. Other techniques such as the dissipativity method used for studying ergodicity of nonlinear SPDEs (see, e.g. [**24, 78**]) also appear not immediately applicable due to the non-standard form of the equations, in particular, the presence of the non-linear boundary condition. Instead, we adopt the asymptotic (equivalent) coupling approach developed in a series of papers by Hairer, Mattingly, Scheutzow and co-authors (see, e.g., [**30, 49, 6, 50**] and references therein) to show that (X, Z) has a unique invariant distribution. The asymptotic equivalent coupling that we construct has a somewhat different flavor from that used in previous works, and entails the analysis of a certain renewal equation. We believe that SPDEs with this kind of boundary condition are likely to arise in the study of scaling limits and control of other parallel server networks, and so our constructions could be useful in a broader context.

(4) **Diffusion approximation of the stationary distribution of** $GI/GI/N$ **queue in Halfin-Whitt regime.** Finally, we show that, in the Halfin-Whitt asymptotic regime, the limit of the sequence $\{\widehat{\pi}^{(N)}\}$ of the stationary distributions of $\{\widehat{Y}^{(N)} = (\widehat{X}^{(N)}, \widehat{Z}^{(N)})\}$ exists, and converges to the unique invariant distribution of the limit process $Y = (X, Z)$, thus resolving (for a large class of service distributions) in the affirmative the question of convergence of the scaled and centered steady-state distributions of $GI/GI/N$ queues in the Halfin-Whitt regime posed in [**51**] (see also [**37**]). We prove this by first, showing the tightness of the sequence of stationary processes $\{\widehat{\pi}^{(N)}\}$, and then, showing that every subsequential limit of $\{\widehat{\pi}^{(N)}\}$ is in

$$
\begin{tikzcd}
V_t^{(N)} \arrow[rr, "\text{Proposition 2.9}", "t \to \infty"'] \arrow[d, "\widehat{\Psi}"] & & V_\infty^{(N)} \arrow[d, "\widehat{\Psi}"] \\
\widehat{Y}_t^{(N)} \arrow[rr, "\text{Theorem 2.7}", "t \to \infty"'] & & \widehat{\pi}^{(N)}
\end{tikzcd}
$$

FIGURE 1. Relation between the representation $\widehat{Y}^{(N)}$ and the Markovian state process $V^{(N)}$ of **[62]**.

fact the unique invariant distribution of the diffusion limit process Y, and hence, uniquely characterizing the limit. Although the tightness for the positive tail of $\widehat{X}^{(N)}$-marginal, $\widehat{X}_\infty^{(N)+}$, of $\{\widehat{\pi}^{(N)}\}$ is proved in **[37]**, we need in addition to establish tightness of the $\widehat{Z}^{(N)}$-marginal in its function space $\mathbb{H}^1(0,\infty)$. This involves a fair amount of computations (see Section 4.1.3), and required us to strengthen the tightness result in **[37]** and to show, instead, the uniform boundedness of $\widehat{X}_\infty^{(N)+}$ in $\mathbb{L}^1$ (Section 4.1.2).

1.2.4. Outline. Part I of the thesis is organized as follows. In chapter 2, we define, for every $N \in \mathbb{N}$, the representation process $Y^{(N)}$ and its centered and renormalized version $\widehat{Y}^{(N)}$, and study long-time behavior and other properties of these processes. The main result of this chapter is Theorem 2.7, which shows the convergence of $\widehat{Y}_t^{(N)}$ to its stationary distribution $\widehat{\pi}^{(N)}$, as $t \to \infty$. Since we have not been able to show that $\widehat{Y}^{(N)}$ is a Markov process, we proved this theorem using the relation between $\widehat{Y}^{(N)}$ and the Markovian state process $V^{(N)}$ introduced in **[61]**. Indeed, we show that the $\widehat{Y}^{(N)}$ satisfies $\widehat{Y}_t^{(N)} = \widehat{\Psi}(V_t^{(N)}), t \geq 0$, for a (deterministic) continuous function $\widehat{\Psi}$ (see (2.2.13)). The proof of Theorem 2.7 then follows from the ergodicity of the state process $V^{(N)}$ (Proposition 2.9) and an application of the continuous mapping theorem (see Figure 1).

$$
\begin{tikzcd}
\widehat{Y}_t^{(N)} \arrow[rr, "\text{Theorem 2.7}", "t \to \infty"'] \arrow[d, "\text{Proposition 4.19}"', "N \to \infty"] & & \widehat{\pi}^{(N)} \arrow[d, "N \to \infty"', "\text{Theorem 4.1}"] \\
Y_t \arrow[rr, dashed, "\text{uniqueness of invariant distribution}", "\text{Theorem 3.26}"'] & & \pi
\end{tikzcd}
$$

FIGURE 2. Summary of results

In Chapter 3, we introduce an SPDE that characterizes the diffusion limit process Y of the sequence of representation processes $\{\widehat{Y}^{(N)}\}_{N\in\mathbb{N}}$. Then we present a fairly explicit from for the unique solution Y of this SPDE (Definition 3.3), and show that Y is an infinite-dimensional Feller Markov process. The main result of this Chapter, Theorem 3.26, shows the uniqueness of invariant distribution of Markov process Y.

Finally in Chapter 4, we prove the convergence of the sequence of stationary distributions $\widehat{\pi}^{(N)}$ (Theorem 4.1). Tightness of this sequence is proved in Section 4.1 (Theorem 4.2). Moreover, to show that every subsequential limit of $\{\widehat{\pi}^{(N)}\}$ is an invariant distribution of the limit process Y, we first establish a (finite-dimensional) diffusion limit theorem for $\{\widehat{Y}_t^{(N)}\}$ at each finite time $t \in [0, \infty)$ under certain conditions on the sequence of initial conditions (Proposition 4.19), and then invoke this result for the stationary version of $\widehat{Y}^{(N)}$ (i.e., when $\widehat{Y}_0^{(N)}$ is distributed as $\widehat{\pi}^{(N)}$). Theorem 4.1 is then proved at the end of Section 4.2.3.

Figure 2 summarizes our main results in Part I.

1.3. Hydrodynamic Limits of Randomized Load-Balancing Networks

1.3.1. Description of Problem. Load balancing is an effective method to improve the performance and reliability of networks by optimizing resource use. With growth in the use of server farms and computer clusters, large-scale load balancing networks appear in a variety of applications, including internet services such as high-traffic web sites, high-bandwidth File Transfer Protocol sites, Domain Name System (DNS) servers, and databases, as well as cloud computing and communications systems. An important problem is the design and analysis of load-balancing algorithms that aim to improve network performance. This is particularly challenging for large-scale networks, where it is not feasible to implement classical algorithms like *join-the-shortest-queue*, which incur high communication overhead and computational cost. In this context, randomized algorithms provide an attractive alternative.

A classical model of randomized load balancing is the so-called "supermarket model" that was introduced in Vvydenskaya et al **[96]** and Mitzenmacher **[82]**, and studied under the exponential service time distribution assumption (see also the static balls-and-bins analog studied in Azar et al. [**5**]). In this model, jobs arrive to a network of N parallel single-server queues according to a Poisson process with rate λN. Upon arrival of a job, d queues are sampled independently and uniformly at random and the job is routed to the shortest queue amongst those sampled. Jobs within each queue are served according to a first-come first-serve (FCFS) scheduling rule, their service times are assumed to be independent and exponentially distributed with unit mean, and also independent of the arrival process. Jobs leave the network on completion of service. When $d = 1$, this model reduces to a system of N independent $M/M/1$ queues (i.e., single-server queues with Poisson arrivals and exponential service times), for which it is a classical result that under the stability condition $\lambda < 1$ the stationary distribution of the length of a single queue has an exponential tail (see Proposition III.3.2 in [**3**]). When $d \geq 2$ and $\lambda < 1$, the stationary distribution of a typical queue is not exactly computable, but it was shown in [**96**] that as the number of servers goes to infinity, the probability $P_\ell^{(N)}$ that the length of a typical queue in the N-server network exceeds ℓ converges to $S_\ell = \lambda^{\alpha_\ell}$, where $\alpha_\ell = (d^\ell - 1)/(d-1)$. Therefore, the limit distribution of a typical queue has a double exponential tail. This shows that introducing just a little bit of random choice leads to a significant improvement in performance, a phenomenon that has been dubbed the "power of two choices" and has led to substantial interest in this class of randomized load balancing schemes.

The analysis of the $SQ(d)$ model in the case of exponential service times is carried out using a slight extension of the so-called "ODE method". This proceeds by first representing the dynamics of the N-server network by a Markov process $S^{(N)} = (S_\ell^{(N)}; \ell \geq 1)$, where $S_\ell^{(N)}(t)$ represents the number of queues that have ℓ or more jobs at time t, and then showing that, as $N \to \infty$, the sequence of suitably scaled Markov processes converges weakly (on finite time intervals) to the unique solution of a countable system of coupled $[0,1]$-valued ordinary differential equations (ODEs). This limit result is obtained by a simple application of Kurtz's theorem (see Theorem 11.2.1 in [**33**]), generalized to countable state spaces. Insight into the equilibrium behavior is obtained by first showing that, under the stability condition $\lambda < 1$, each Markov process $S^{(N)}$ is ergodic, then characterizing the limit of the sequence of scaled stationary distributions as the unique equilibrium point of this system of ODEs [**96**], and finally explicitly identifying this equilibrium point.

However, in most real-world applications, service times are typically not exponentially distributed. For example, statistical analyses suggest that service times follow a Log-Normal distribution in (medium-scale) call centers [**16**], a Gamma distribution in Automatic Teller Machines [**67**], and studies of content download in Amazon S3 suggest that download times follow a shifted exponential distribution [**21, 72**]. In the case of general service times, in order to describe the evolution of the system it is not sufficient to keep track of the fraction of queues with ℓ jobs at any time. For each job in service, one has also to keep track of additional information such as its age (the amount of time the job has spent in service) or its residual service time. In the system with N servers, this requires keeping track of N additional random variables, and thus the dimension of the Markovian state representation grows with N, which is not convenient for obtaining a limit theorem.

1.3.2. Prior Work. For the case of exponential service time distributions, various modifications of the supermarket model have been studied, that incorporate features of relevance in applications such as load migration, load stealing and thresholds (see, e.g., [**35**], [**81**] and [**60**]). Other theoretical results on the model include a propagation of chaos result [**44**], analysis of the maximum equilibrium queue length [**74**] in an N-server network, and strong approximations [**75**].

Although results on the exponential service time model were obtained almost two decades ago, we are not aware of any rigorous prior work on the transient behavior of the supermarket model with general service time distributions. Till recently, there was little work even on the equilibrium behavior, with the exception of the work of Foss and Chernova [**36**], who used fluid limits to show that the natural condition $\lambda < 1$ is sufficient for stability of the N-server model with general service distributions. However, recent progress on the equilibrium behavior in the case of Poisson arrivals was made in a nice series of papers by Bramson et al. [**10, 11, 12**]. (Some of the results in [**36, 10, 11, 12**] have the nice feature that they are valid for more general systems, including more general scheduling disciplines within each queue, but we restrict our discussion to the case of the supermarket model.) The general approach adopted in the papers [**10, 11, 12**] is to first establish the ansatz that any fixed number of queues at stationarity is mutually asymptotically independent (in the spirit of a propagation-of-chaos result at equilibrium), and to then show that the stationary distributions converge, as $N \to \infty$, to the unique solution of a certain fixed point equation, in the spirit of the cavity method from statistical physics. The ansatz was established in [**11**] for the class of service distributions that have a decreasing hazard rate function. When combined with uniform bounds on the tails of the stationary queue length distributions obtained in [**9**], the scaled stationary queue lengths of the N-server systems were then shown to converge to the unique solution of a certain fixed point equation. The latter equation was analyzed in [**12**] for the sub-class of power law distributions with exponent $-\beta$ for $\beta > 1$, to uncover the interesting phenomenon that the limiting stationary distribution has a doubly exponential tail if $\beta > d/(d-1)$, an exponential tail if $\beta = d/(d-1)$ and a power law tail if $\beta < d/(d-1)$.

The results in [**10, 11, 12**] assume Poisson arrival streams and the convergence results in [**11, 12**] require service distributions with decreasing hazard rate functions. This raises the natural question of whether the approach of [**10, 11, 12**] can be used to analyze the equilibrium under more general assumptions than Poisson arrivals and service distributions with decreasing hazard rates. Such a generalization would entail a proof of the ansatz in greater generality which, according to [**11, 12**], appears to be a difficult problem.

1.3.3. Contributions. The main objective of Part II is to introduce a mathematical framework for the analysis of stochastic networks with randomized load balancing in the presence of general service times, with the specific aim of applying this framework to characterize hydrodynamic limits of supermarket model with general service times. A key component of our approach is a certain particle representation of the model that leads to a Markovian description of the dynamics in terms of interacting measure-valued stochastic processes. Below is a list of our contributions and results.

(1) **A new interactive measure-valued representation for load-balancing networks.** We introduce a particle representation of the model that allows one to represent the Markovian dynamics of all N-server systems on a common state space. Specifically, we represent the state of an N-server network at any time t in terms of a sequence of measures $\nu^{(N)}(t) = (\nu_\ell^{(N)}(t); \ell \geq 1)$, where $\nu_\ell^{(N)}(t)$ is the measure that has a "particle" representing a unit delta mass at the age of each job that is in service at a queue

of length greater than or equal to ℓ at time t. This yields a description of the dynamics of the general supermarket model with N servers in terms of a coupled system of interacting measure-valued stochastic processes. The framework introduced here generalizes a simpler representation introduced in Part I for the $GI/GI/N$ queueing model, and can be adapted to study a larger class of randomized load-balancing schemes of interest.

(2) **A hydrodynamic limit theorem for the new representation.** Under mild conditions on the service time distribution and general arrival processes (Assumptions I-III), we show that, as $N \to \infty$, the corresponding sequence of scaled interacting measure-valued stochastic processes converges weakly (with respect to a suitable topology) to the unique solution $(\nu_\ell\ ; \ell \geq 1)$ of a countably infinite system of coupled deterministic measure-valued evolution equations. Despite being measure-valued, this countable system of coupled equations can be partially solved in an explicit fashion, leading to a tractable and intuitive description of the limit. In contrast with previous work, this result enables the study of (approximate) transient behavior of large-scale load-balancing networks with general service distributions. It worth emphasizing that, as depicted in Chapter 8, the time it takes for the network to reach stationarity is even longer under certain non-exponential service distributions, and therefore, understanding transient behavior of these networks is even more crucial in the presence of general service distribution.

(3) **A PDE representation of the hydrodynamic limit.** The measure-valued hydrodynamic limit theorem implies, in particular, the convergence of the fraction of queues with length greater than or equal to ℓ, to a functional $(S_\ell; \ell \geq 1)$ of the measure-valued limit process $(\nu_\ell; \ell)$. However, $(\nu_\ell; \ell \geq 1)$ contains more information than just S_ℓ, and if one is only interested the limit of queue length distribution, a simpler description of the limit process is available . We show that the dynamics of $(S_\ell; \ell \geq 1)$ can be

described using a system of coupled partial differential equations (PDEs), which we call hydrodynamic PDEs. This provides a direct generalization of the ODE description obtained in **[96, 81]** when the service time distribution is exponential. The PDEs can be used to approximate not only the queue length distribution but also other quality of service (QoS) parameters such as the virtual waiting time.

(4) **Uniqueness of the solution to the hydrodynamic measure-valued equations and PDEs.** A key challenge is to prove the uniqueness of the solution to the countable system of coupled measure-valued equations that characterize the hydrodynamic limit. This proof is non-standard, and is carry out by imposing a carefully chosen norm on the space of infinite vectors of measure-valued processes, and invoking properties of certain renewal equations. Similarly, as the hydrodynamic PDEs consists of a countable system of non-standard partial integro-differential equations, establishing the uniqueness of the solution to these equations requires a new approach.

(5) **Numerical solution to the hydrodynamic PDEs and insight into network performance.** We devise and use a stable numerical scheme to approximate the solution to the hydrodynamic PDEs, which provides a computationally efficient alternative to Monte Carlo simulations that could be useful for studying the performance and design of large networks. We show that the obtained numerical approximations provides a good match with simulations, and illustrate how this approximation can be used to shed insight into transient phenomena of practical relevance.

1.3.4. Outline. Part II of the thesis is organized as follows. First in Chapter 5, we study in details an N-server network with general service distribution and under the "supermarket" randomized algorithm. We introduce a sequence of interacting measure-valued processes $(\nu_\ell^{(N)}; \ell \geq 1)$ that describes the state of the network at each time. Moreover, we describe the dynamics of the network, by deriving equations that govern the evolution of these measure-valued state process. Finally, we establish martingale decompositions for certain stochastic processes that describe arrivals and departures to and from the network. These decompositions are essential for establishing hydrodynamic limits later in Chapter 7.

Basic assumptions imposed on the arrival process and service time distribution are also presented in this chapter (see Assumption I-III).

In Chapter 6, we describe the hydrodynamic limit of the Markovian state process. This limit is characterized as the solution to a system of coupled measure-valued equations, which we call "measure-valued hydrodynamic equations". The uniqueness of the solution to these equations is also established in this chapter (see Theorem 6.3). Moreover, we introduce a reduced version of measure-valued hydrodynamic equations, which is a system of coupled partial integro-differential equations (PDEs) that we call hydrodynamic PDEs. We study these PDEs under an additional smoothness assumption on the service time distribution (see Assumption IV in Section 6.2), and in particular, we prove the uniqueness of the solution to these PDEs.

In Chapter 7, we establish the main result of Part II, that is a hydrodynamic limit theorem for the state representation of the load balancing network described in Chapter 5. Specifically, we prove that the scaled interacting measure-valued representation $(\nu_\ell^{(N)}; \ell \geq 1)$ defined in Section 5.1.2, converges to the unique solution of the measure-valued hydrodynamic equations of Section 6.1.1 (see Theorem 7.1 below).

The measure-valued hydrodynamic limit theorem implies, in particular, the convergence of the fraction of queues with length greater than or equal to ℓ, to a functional $(S_\ell;\ell\geq 1)$ of the measure-valued limit process $(\nu_\ell;\ell)$, although $(\nu_\ell;\ell\geq 1)$ contains more information than just S_ℓ. We show that if one is only interested the limit of queue length distribution, one can avoid the measure-valued equations, and characterize the limit only using the hydrodynamic PDEs. Moreover, we show in this chapter that the PDEs can be used to approximate not only the queue length distribution but also other quality of service (QoS) parameters such as the virtual waiting time.

Finally, in Chapter 8, we discuss the application of our result in understanding the behavior of stochastic networks in practice. We devise and use a stable numerical scheme to approximate the solution to the hydrodynamic PDEs, and show that the obtained numerical approximations provides a good match with Monte Carlo simulations. This numerical approximation, hence, provides a computationally efficient alternative to Monte Carlo simulations that could be useful for studying the performance and design of large networks, Moreover, we illustrate how this approximation can be used to shed insight into transient phenomena of practical relevance (see Section 8.2).

1.4. Some Common Notation

The following notation will be used throughout the paper. $\mathbb{Z}$, $\mathbb{Z}_+$ and $\mathbb{N}$ are the sets of integers, nonnegative integers and positive integers, respectively. Also, $\mathbb{R}$ is set of real numbers and $\mathbb{R}_+$ the set of nonnegative real numbers. For $a,b\in\mathbb{R}$, $a\wedge b$ and $a\vee b$ denote the maximum and maximum of a and b, respectively. Also, $a^+\doteq a\vee 0$ and $a^-\doteq -(a\wedge 0)$, and hence $a=a^+-a^-$. For a set B, $\mathbb{1}_B\left(\cdot\right)$ is the indicator function of the set B (i.e., $\mathbb{1}_B\left(x\right)=1$ if $x\in B$ and $\mathbb{1}_B\left(x\right)=0$ otherwise). Moreover, with a slight abuse of notation, on every domain V, $\mathbf{1}$ denotes the constant function equal to 1 on V.

1.4.1. Function Spaces. For every $n\in\mathbb{N}$ and subset $V\subset\mathbb{R}^n$, $\mathbb{C}(V)$, $\mathbb{C}_b(V)$ and $\mathbb{C}_c(V)$ are respectively, the space of continuous functions on V, the space of bounded continuous functions on V and the space of continuous functions with compact support on V. We use $\|f\|_\infty$ to denote the supremum of $|f(s)|$, $s \in V$. For $f \in \mathbb{C}_{\mathbb{R}}[0,\infty)$ and $T > 0$, $\|f\|_T$ denotes the supremum of $|f(s)|$, $s \in [0,T]$. A function for which $\|f\|_T < \infty$ for every $T < \infty$ is said to be locally bounded. Also, $\mathbb{C}^0[0,\infty)$ denotes the subspace of functions $f \in \mathbb{C}_{\mathbb{R}}[0,\infty)$ with $f(0) = 0$, $\mathbb{C}^1[0,\infty)$ denotes the set of functions $f \in \mathbb{C}_{\mathbb{R}}[0,\infty)$ for which the derivative, denoted by f', exists and is continuous on $[0,\infty)$, and $\mathbb{C}^1_b[0,\infty)$ represents the subset of functions in $\mathbb{C}^1[0,\infty)$ that are bounded and have bounded derivative.

For every Polish space $\mathbb{X}$, $\mathbb{C}_{\mathbb{X}}[0,\infty)$ denotes the set of continuous $\mathbb{X}$-valued functions on $[0,\infty)$, and $\mathbb{D}_{\mathbb{X}}[0,\infty)$ is the set of $\mathbb{X}$-valued functions on $[0,\infty)$ which are right continuous and have left limits. Endowed with the Skorokhod topology, $\mathbb{D}_{\mathbb{X}}[0,\infty)$ is a complete separable metric space. However, we sometimes equip $\mathbb{D}_{\mathbb{R}}[0,\infty)$ with the topology of uniform convergence on compact sets. For every function $f \in \mathbb{D}_{\mathbb{X}}[0,\infty)$ and $T \geq 0$, $w'(f,\delta,T)$ is the modulus of continuity of f in $\mathbb{D}_{\mathbb{X}}[0,\infty)$; see (3.6.2) in **[33]** for a precise definition of w'. Furthermore, for every function $f \in \mathbb{D}_{\mathbb{R}}[0,\infty)$, $[f]$ denotes the quadratic variation of the f, defined by

$$[f]_t = \lim_{|P|\to 0} \sum_{k=1}^{n} \left(f(t_k) - f(t_{k-1})\right)^2,$$

where the limit is taken over all partitions $P = \{t_0 = 0, t_1, ..., t_n = t\}$ $[0,t]$ with $|P| \doteq \max_{k=1,...,n} |t_k - t_{k-1}|$. For a $\mathbb{D}_{\mathbb{R}}[0,\infty)$-valued stochastic process, the limit is defined in the sense of convergence in probability.

Moreover, let $\mathbb{L}^1(0,\infty)$, $\mathbb{L}^2(0,\infty)$, and $\mathbb{L}^\infty(0,\infty)$, denote, respectively, the spaces of integrable, square-integrable and essentially bounded functions on $(0,\infty)$ with their corresponding standard norms. The space $\mathbb{L}^2$ is a Hilbert space with the inner product

$$\langle f, g\rangle_{\mathbb{L}^2} = \int_0^\infty f(x)g(x)dx.$$

Also, $\mathbb{L}^1_{\text{loc}}(0,\infty)$ denotes the space of locally integrable functions on $[0,\infty)$. The space $\mathbb{H}^1(0,\infty)$ denotes the space of square integrable functions f on $(0,\infty)$ whose weak derivative f' exists and is also square integrable. $\mathbb{H}^1(0,\infty)$ equipped with the inner product

$$\langle f, g \rangle_{\mathbb{H}^1} = \langle f, g \rangle_{\mathbb{L}^2} + \langle f', g' \rangle_{\mathbb{L}^2}, \tag{1.4.1}$$

and the corresponding norm

$$\|f\|_{\mathbb{H}^1} = \left(\|f\|^2_{\mathbb{L}^2(0,\infty)} + \|f'\|^2_{\mathbb{L}^2(0,\infty)} \right)^{\frac{1}{2}},$$

is a separable Hilbert space [**15**, Proposition 8.1]. Recall that every function $f \in \mathbb{H}^1(0,\infty)$ is absolutely continuous and its density coincides with its weak derivative [**34**, Problem 5 on Page 290]. Throughout this paper, the density of an absolutely continuous function f is denoted by f'.

1.4.2. Measure Spaces. For every subset V of $\mathbb{R}$ or $\mathbb{R}^2$ endowed with Borel sigma-algebra, let $\mathbb{M}_F(V)$ (respectively, $\mathbb{M}_{\leq 1}(V)$) be the space of finite positive measures (respectively, probability measures) in V. For $\mu \in \mathbb{M}_F(V)$ and any bounded Borel-measurable function f on V, we denote the integral of f with respect to μ, in Part I by

$$\mu(f) \doteq \int_V f(x)\mu(dx),$$

and in Part II by

$$\langle f, \mu \rangle \doteq \int_V f(x)\mu(dx).$$

We can extent the bracket notation for the signed measures, and for every measure μ with representation $\mu = \mu^+ - \mu^-; \mu^+, \mu^- \in \mathbb{M}_F[0,\infty)$, define $\langle f, \mu \rangle \doteq \langle f, \mu^+ \rangle - \langle f, \mu^- \rangle$. We equipped $\mathbb{M}_F(V)$ and $\mathbb{M}_{\leq 1}(V)$ with the weak topology: $\mu_n \Rightarrow \mu$ if and only if $\langle f, \mu_n \rangle \to \langle f, \mu \rangle$ for all $f \in \mathbb{C}_b[0,\infty)$. Moreover, the Prohorov's metric d_P on $\mathbb{M}_F(V)$ (see page 72 in [**8**]) induces the same topology and $(\mathbb{M}_F(V), d_P)$ is a Polish space (Theorem 6.8 in Chapter 1 of [**8**]). We also denote $\mathbb{M}_D[0,\infty)(V)$ to be the subspace of $\mathbb{M}_F[0,\infty)$ consisting of measures of the form $\sum_{i=1}^m \delta_{x_i}$ for some $m \in \mathbb{N}$ and $x_i \in V, i = 1, ..., m$.

Also, we denote by $\mathbb{M}(V)$ the space of Radon measures, that is, the ones which assign finite measures to every relatively compact subset of V. Alternatively, identifying a Radon measure by the function $\mu(\varphi) \doteq \int_V \varphi(x)dx$ on

$\mathbb{C}_{\mathbb{R}}[0,\infty)_c(V)$, one can define a Radon measures on V as a linear real-valued mapping on $\mathbb{C}_{\mathbb{R}}[0,\infty)_c(V)$ such that for every compact set $\mathcal{K} \subset V$, there exist a finite $C_{\mathcal{K}}$ such that

$$\mu(\varphi) \leq C_{\mathcal{K}} \|\varphi\|_{\infty}, \forall \varphi \text{ with } \mathrm{supp}(\varphi) \subset \mathcal{K}.$$

Part I

Diffusion Approximations for Many-Server Queues in Heavy Traffic

CHAPTER 2

A Representation for the N-Server Queueing Network

In this Chapter, we introduce a new infinite-dimensional representation for a $GI/GI/N$ queueing network with general service distributions, which leads to a tractable Markov Process in the diffusion limit. Specifically, we consider a representation $(X^{(N)}, Z^{(N)})$, where $X^{(N)}$ is, the number of jobs in the N-server network, and $Z^{(N)}$ is a function-valued process that contains just enough information so that the limit of scaled and renormalized process $(\widehat{X}^{(N)}, \widehat{Z}^{(N)})$ in the Halfin-Whitt asymptotic regime is an infinite-dimensional Markov process. Long-time behavior and other properties of this process is also studied in this Chapter.

In Section 2.1, we describe the $GI/GI/N$ queueing model in details, and state our assumptions on service time distribution and arrival process. The new representation $Y^{(N)} = (X^{(N)}, Z^{(N)})$ is defined in Section 2.2, and its connection to the Markovian state process $V^{(N)}$ used in **[61]** is demonstrated. We show in Theorem 2.7 that as $t \to \infty$, the process $Y_t^{(N)}$ converges to its stationary distribution. Finally, in Section 2.3, we establish a fluid limit theorem for $Y^{(N)}$, and describe the dynamics of the centered and diffusion-scaled process $\widehat{Y}^{(N)}$.

2.1. Description of the N-Server Network

This section contains the basic definitions and descriptions of the $GI/GI/N$ queueing model. In Section 2.1.1, we state the required assumptions on the arrival process and service time distribution. Then in Section 2.1.2, we recall the Markovian State process $V^{(N)}$ introduce in **[61]**

2.1.1. Basic Definitions and Assumptions. We consider a $GI/GI/N$ model in which a common stream of jobs arriving according to a renewal process to a network of N parallel servers with a common infinite-capacity queue. Servers follow the First-Come-First-Serve (FCFS) service policy and are non-idling, in the sense that no server is idling when there is a job waiting in the queue. Let $E^{(N)}$ denote the cumulative arrival process, that is, $E_t^{(N)}$ represents the number of jobs that arrived to the system in the interval $[0,t]$. We assume that $E^{(N)}$ is a renewal process. Specifically, let $\tilde{E}$ be a renewal process with delay $\tilde{u}_0$ and renewal times $\{\tilde{u}_n; n \in \mathbb{N}\}$, with common cumulative distribution function (c.d.f.) $\tilde{G}_E$, which has mean 1 and finite variance $\tilde{\sigma}_A^2$.

ASSUMPTION I. For every $N \in \mathbb{N}$, the arrival process $E^{(N)}$ satisfies $E_t^{(N)} = \tilde{E}(\lambda^{(N)}t)$, where

$$\lambda^{(N)} \doteq N - \beta\sqrt{N}. \tag{2.1.1}$$

We further assume that $\tilde{G}_E$ is non-lattice and has full support (i.e. $\tilde{G}_E(x) < 1$ for all $x > 0$) and finite $(2+\epsilon)$ moment, for some $\epsilon > 0$. In other words, $E^{(N)}$ is a renewal process with delay $u_0^{(N)} = \tilde{u}_0/N$ and inter-arrival times $u_n^{(N)} = \tilde{u}_n/N, n \in \mathbb{N}$, with c.d.f. $G_E^{(N)}(x) \doteq \tilde{G}_E(Nx)$, mean $1/\lambda^{(N)}$, and finite $(2+\epsilon)$ moment.

The condition (2.1.1) captures the Halfin-Whitt asymptotic regime, where the arrival rate scales with the number of servers N as $N - \mathcal{O}(\sqrt{N})$.

REMARK 2.1. The full support assumption imposed on the inter-arrival distribution is only used to show ergodicity of a Markovian state representation of the sub-critical $GI/GI/N$ queue (specifically, in the proof of Proposition 2.9). However, this is not a necessary result for ergodicity, and our main results will continue to hold under weaker conditions on the inter-arrival distribution, as long as they guarantee the ergodicity of $\{Y_t^{(N)}; t \geq 0\}$.

The sequence of service times, denoted by $\{v_j; j \in \mathbb{Z}\}$, are assumed to be i.i.d. and independent of the arrival process, with a common c.d.f. G. Define $\overline{G} = 1 - G$ to be the complementary c.d.f.

ASSUMPTION II. The service time c.d.f. G satisfies the following conditions:

(a) G has a finite mean equal to one, and is absolutely continuous with continuous density g.

(b) G has finite $2+\epsilon$ moment for some $\epsilon > 0$, that is, $\overline{G}(x) = \mathcal{O}(x^{-(2+\epsilon)})$, as $x \to \infty$.

Assumptions I and II, except for G having a density, where also imposed in **[37]** (they also assumed other technical conditions on the service time distribution, which are verified when G has a density, see proof of Proposition 4.5). We impose one additional condition on G. Recall that when G has a density g, the hazard rate function h of G is defined as

$$h(x) \doteq \frac{g(x)}{1-G(x)} = \frac{g(x)}{\overline{G}(x)}, \tag{2.1.2}$$

for x in the support of the distribution of G. In Assumption III below, we will require h to be bounded, which (since the hazard rate function h is not integrable on $(0,\infty)$), automatically implies that G has support $[0,\infty)$.

ASSUMPTION III. The hazard rate function h lies in $\mathbb{C}_b^1[0,\infty)$.

REMARK 2.2. Assumption III implies that

$$H \doteq \sup_{x\in[0,\infty)} h(x) < \infty.$$

Since $g = \overline{G}h$ and $h' = g'/\overline{G} + g^2/\overline{G}^2$, Assumption III also implies that the probability density function (p.d.f.) g is differentiable with a continuous derivative g', and

$$H_2 \doteq \sup_{x\in[0,\infty)} |h_2(x)| < \infty, \qquad \text{where} \qquad h_2(x) \doteq \frac{g'(x)}{\overline{G}(x)} = h'(x) - h^2(x)\,; x \geq 0.$$

REMARK 2.3. It follows from Assumption II.a that $\overline{G}$ is integrable, and since it is also bounded by 1, it must be that $\overline{G} \in \mathbb{L}^1(0,\infty) \cap \mathbb{L}^2(0,\infty)$. Then Assumption III implies that g and g' also lie in $\mathbb{L}^1(0,\infty) \cap \mathbb{L}^2(0,\infty)$, and hence, both $\overline{G}$ and g lie in $\mathbb{H}^1(0,\infty)$. Furthermore, $\int_\cdot^\infty \overline{G}(x)dx$ is integrable by Assumption II.b is bounded by one by Assumption II.a, and hence, $\int_\cdot^\infty \overline{G}(x)dx$ also lies in $\mathbb{L}^1(0,\infty) \cap \mathbb{L}^2(0,\infty)$.

Finally, to prove the main result, we need the following additional condition on the service time distribution.

ASSUMPTION IV. The service time distribution satisfies the following conditions.

(a) G has finite $(3+\epsilon)$ moment for some $\epsilon > 0$, that is, $\overline{G}(x) = \mathcal{O}(x^{-(3+\epsilon)})$, as $x \to \infty$.

(b) The function g' is absolutely continuous with bounded density g'' which satisfies $g''(x) = \mathcal{O}(x^{-(2+\epsilon)})$ as $x \to \infty$.

REMARK 2.4. The function $\int_{\cdot}^{\infty} \int_{s}^{\infty} \overline{G}(x)dx\ ds$ is integrable under Assumption IV.a.

2.1.2. The Markovian State Descriptor $V^{(N)}$. In **[61]**, the $GI/GI/N$ queue was represented by the following state descriptor $V^{(N)} = (V_t^{(N)}; t \geq 0)$, where $V^{(N)}$ is a three-component process:

$$V_t^{(N)} \doteq (R_t^{(N)}, X_t^{(N)}, \nu_t^{(N)}). \tag{2.1.3}$$

The first component $R^{(N)}$ is the forward recurrence time of the arrival process $E^{(N)}$:

$$R_t^{(N)} \doteq \inf\{u > t : E_u^{(N)} > E_t^{(N)}\} - t, \quad t \geq 0, \tag{2.1.4}$$

and note that $R_0^{(N)} = u_0^{(N)}$ (see Assumption I). When the arrival process is Poisson, this component can be omitted. The second component, $X_t^{(N)}$, is the number of jobs in the system at time $t \geq 0$, and the third component, $\nu_t^{(N)}$, is the measure on $[0, \infty)$ that is a sum of delta masses, each at the age (i.e. the amount of time spent thus far in service) of a job currently in service. Given $N \in \mathbb{N}$, for every $t \geq 0$, $V_t^{(N)}$ takes values in the space

$$\mathcal{V}^{(N)} \doteq \{(r, x, \mu) \in \mathbb{R}_+ \times \mathbb{N} \times \mathbb{M}_D[0, \infty) : x \wedge \mu(\mathbf{1}) = N\}, \tag{2.1.5}$$

where recall from Section 1.4 that $\mathbb{M}_F[0, \infty)$ is the set of finite positive measures on $[0, \infty)$, and $\mathbb{M}_D[0, \infty)$ is a subspace of $\mathbb{M}_F[0, \infty)$ consisting of measures of the form

$\sum_{i=1}^{m} \delta_{x_i}$ for some $m \in \mathbb{N}$ and $x_i \in \mathbb{R}, i = 1, ..., m$. The process $V^{(N)}$ is explicitly constructed in [**57**, Appendices A].

We now express $V^{(N)}$ more explicitly in terms of certain primitives of the queue. Jobs are index by $j \in \mathbb{Z}$. A job j arrives to the system at time $\alpha_j^{(N)}$, and if an idle server is available at that time, the job immediately enters service. Otherwise, the job joins the back of the queue and enters service at a later time $\beta_j^{(N)}$. When its service is completed, the job departs from the system at time $\gamma_j^{(N)} \doteq \beta_j^{(N)} + v_j$, where $\{v_j\}$ is the i.i.d. sequence of service time requirements. We assign non-positive indices $j \leq 0$ to the jobs that are initially in service, and positive indices $j \geq 1$ to other jobs, either initially in queue or arrived after time 0, and in the order of their arrival. Note that there are initially $X_0^{(N)}$ jobs in system, and by the non-idling assumption, $X_0^{(N)} \wedge N$ of them are initially in service. Hence, $j_0^{(N)} \doteq -(X_0^{(N)} \wedge N) + 1$ is the smallest job index.

For every job j, we define the age process $\{a_j^{(N)}(t); t \geq 0\}$ as

$$(2.1.6) \qquad a_j^{(N)}(t) \doteq \begin{cases} [t - \beta_j^{(N)}(t)] \vee 0 & \text{if } t \leq \beta_j^{(N)} + v_j, \\ v_j & \text{otherwise.} \end{cases} \qquad t \geq 0,$$

that is, the age of a job is zero before service entry, then grows linearly with rate 1 until it reaches v_j at the departure time, and remains constant afterwards. Also, define $\{K_t^{(N)}; t \geq 0\}$ to be the cumulative service entry process, that is, $K_t^{(N)}$ is the total number of jobs that entered service during the interval $[0, t]$. Note that $K_t^{(N)}$ is then the largest job index that has entered service by the time t. The measure $\nu_t^{(N)}$ can be expressed as

$$(2.1.7) \qquad \nu_t^{(N)} \doteq \sum_{j=j_0^{(N)}}^{K_t^{(N)}} \delta_{a_j^{(N)}(t)} 1_{\{a_j^{(N)}(t) < v_j\}}.$$

Note that for every $t \geq 0$, $\nu_t^{(N)}(\mathbf{1})$ is the number of jobs in service at time t, and therefore, the non-idling assumption yields

$$(2.1.8) \qquad \nu_t^{(N)}(\mathbf{1}) = N \wedge X_t^{(N)},$$

or equivalently,

$$N - \nu_t^{(N)}(\mathbf{1}) = (N - X_t^{(N)})^+. \tag{2.1.9}$$

Although the results of this paper do not depend on the particular rule used to assign jobs to servers, for technical purposes, we define the station process $\{\kappa_j^{(N)}(t); t \geq 0\}$ for every job j as follows: $\kappa_j^{(N)}(t)$ is equal to the index $i \in \{1, ..., N\}$ of the server at which the job j receives/received service if it has already entered service, and is equal to 0, otherwise. For $t \geq 0$ define

$$\tilde{\mathcal{F}}_t^{(N)} \doteq \sigma\left(R^{(N)}(s), a_j^{(N)}(s), \kappa_j^{(N)}(s); s \in [0,t], j = -N+1, ... -1, 0, 1, ...\right),$$

and let $\{\mathcal{F}_t^{(N)}; t \geq 0\}$ be the associated right-continuous filtration that is complete with respect to $\mathbb{P}$. It has been shown in [**57**, Lemma A.1, Lemma B.1] that $\{V_t^{(N)}, \mathcal{F}_t^{(N)}; t \geq 0\}$ is a càdlàg Feller Markov process.

2.2. An Infinite-Dimensional Representation

In This Section, we define our representation $Y^{(N)}$. The main result of this section is Theorem 2.7, which states that for every $N \in \mathbb{N}$, the suitably centered and renormalized process $\widehat{Y}^{(N)}$ lies in a state space $\mathcal{Y}$ (that does not depend on N), and as $t \to \infty$, $\widehat{Y}_t^{(N)}$ converges to its stationary distribution $\widehat{\pi}^{(N)}$.

The state space $\mathcal{Y}$ is defined in Section 2.2.1, and is shown to be a closed subspace of the Polish space $\mathbb{R} \times \mathbb{H}^1(0,\infty)$. In Section 2.2.2, the process $Y^{(N)}$ and its centred and renormalized version $\widehat{Y}^{(N)}$ are defined. The connection between the process $Y^{(N)}$ and the Markovian state process $V^{(N)}$ that is used in [**61**] is demonstrated in 2.2.3, and in particular, $\widehat{Y}_t^{(N)}$ is shown to be a continuous function of $V_t^{(N)}$ for each $t \geq 0$ (see (2.2.15)). Finally in Section 2.2.4, we prove Theorem 2.7, by first establishing the ergodicity of the Markovian state process $V^{(N)}$ using an existing result on stability of subcritical $GI/GI/N$ queues (see Proposition 2.9), and then using the relation between $\widehat{Y}^{(N)}$ and $V^{(N)}$.

2.2.1. The State Space $\mathcal{Y}$. Define the space $\mathcal{Y}$ as

$$\mathcal{Y} \doteq \left\{(x, f) \in \mathbb{R} \times \mathbb{H}^1(0, \infty) : f(0) = x \wedge 0\right\}, \tag{2.2.1}$$

where, for $f \in \mathbb{H}^1(0,\infty)$, $f(0)$ denotes $\lim_{r\downarrow 0} f(r)$, which is well defined since functions in $\mathbb{H}^1(0,\infty)$ are continuous on $[0,\infty)$, using the following theorem.

LEMMA 2.5. The following statements hold.

(a) Every function $f \in \mathbb{H}^1(0,\infty)$ has a version that is uniformly continuous on $(0,\infty)$. Therefore, $f(0) = \lim_{r\downarrow 0} f(r)$ is well defined and f can be continuously extended to $[0,\infty)$, that is, $f \in \mathbb{C}_{\mathbb{R}}[0,\infty)$.

(b) The embedding $I : \mathbb{H}^1(0,\infty) \mapsto \mathbb{C}_{\mathbb{R}}[0,\infty)$ is continuous.

(c) For every $t \geq 0$, the mapping $f \mapsto f(t+\cdot)$ is a continuous mapping from $\mathbb{H}^1(0,\infty)$ into itself. Also, for every $f \in \mathbb{H}^1(0,\infty)$, the translation mapping $t \mapsto f(t+\cdot)$ from $[0,\infty)$ to $\mathbb{H}^1(0,\infty)$ is continuous. Moreover,

$$\lim_{t\to\infty} \|f(t+\cdot)\|_{\mathbb{H}^1} = 0. \tag{2.2.2}$$

PROOF. We start with the proof of property a. As discussed in Section 1.4, $f \in \mathbb{H}^1(0,\infty)$ is absolutely continuous with square integrable density f'. Thus, using Hölder's inequality, for every $s, t \in (0,\infty)$ with $|t-s| \leq \delta$,

$$|f(t) - f(s)|^2 = \left|\int_s^t f'(u)du\right|^2 \leq |t-s| \int_s^t f'(u)^2 du \leq \delta \|f'\|_{\mathbb{L}^2},$$

and therefore, f is uniformly continuous on $(0,\infty)$. Consequently, the limit $f(0) \doteq \lim_{r\downarrow 0} f(r)$ exists and f has a continuous extension to $[0,\infty)$.

To show property b, note that for any $f_1, f_2 \in \mathbb{H}^1(0,\infty)$ and every $0 < T < \infty$, the restriction of $f_1 - f_2$ to $(0,T)$ lies in $\mathbb{H}^1(0,T)$. By Morrey's inequality ([**34**, Theorem 5 on page 269]), there exists a constant $C_T < \infty$ such that

$$\|f_1 - f_2\|_T \leq C_T \|f_1 - f_2\|_{\mathbb{H}^1(0,T)} \leq C_T \|f_1 - f_2\|_{\mathbb{H}^1(0,\infty)}. \tag{2.2.3}$$

Since $\mathbb{C}_{\mathbb{R}}[0,\infty)$ is equipped with the topology of uniform convergence on compact sets, this proves the continuity stated in property b.

We now establish the claims stated in property c. The first claim follows from the following elementary inequality: for every $f_1, f_2 \in \mathbb{H}^1(0,\infty)$,

$$\begin{aligned} \|f_1(t+\cdot) - f_2(t+\cdot)\|_{\mathbb{H}^1}^2 &= \int_t^\infty (f_1(u) - f_2(u))^2 du + \int_t^\infty (f_1'(u) - f_2'(u))^2 du \\ &\leq \|f_1 - f_2\|_{\mathbb{H}^1}. \end{aligned}$$

For the second claim, fix $f \in \mathbb{L}^2(0,\infty)$ and $\epsilon > 0$. Since $\mathbb{C}_c(0,\infty)$ is dense in $\mathbb{L}^2(0,\infty)$, there exists a function $\tilde{f}$ that is uniformly continuous on $(0,\infty)$ such that $\|f - \tilde{f}\|_{\mathbb{L}^2} \leq \epsilon$. Therefore, for every $t, t_0 \in (0,\infty)$,

$$\begin{aligned} \|f(t+\cdot) - f(t_0+\cdot)\|_{\mathbb{L}^2} &\leq \|f(t+\cdot) - \tilde{f}(t+\cdot)\|_{\mathbb{L}^2} + \|\tilde{f}(t+\cdot) - \tilde{f}(t_0+\cdot)\|_{\mathbb{L}^2} \\ &\quad + \|\tilde{f}(t_0+\cdot) - f(t_0+\cdot)\|_{\mathbb{L}^2} \\ &\leq 2\epsilon + \|\tilde{f}(t+\cdot) - \tilde{f}(t_0+\cdot)\|_{\mathbb{L}^2}. \end{aligned}$$

Taking the limit as $t \to t_0$ on both sides of the last inequality, by the uniform continuity of $\tilde{f}$ and the dominated convergence theorem, we have

$$\lim_{t\to t_0} \|f(t+\cdot) - f(t_0+\cdot)\|_{\mathbb{L}^2} \leq 2\epsilon + \lim_{t\to t_0} \|\tilde{f}(t+\cdot) - \tilde{f}(t_0+\cdot)\|_{\mathbb{L}^2} = 2\epsilon.$$

Since $\epsilon > 0$ is arbitrary, this shows that the translation map $t \mapsto f(t+\cdot)$ is continuous in $\mathbb{L}^2(0,\infty)$. If $f \in \mathbb{H}^1[0,\infty)$, then $f' \in \mathbb{L}^2(0,\infty)$ and so the above argument also shows that the map $t \mapsto f'(t+\cdot)$ is continuous in $\mathbb{L}^2(0,\infty)$, which proves the continuity of $t \mapsto f(t+\cdot)$ in $\mathbb{H}^1(0,\infty)$. Finally, by definition,

$$\|f(t+\cdot)\|_{\mathbb{L}^2}^2 = \int_0^\infty f^2(t+x)dx = \int_t^\infty f^2(x)dx.$$

Since $f \in \mathbb{L}^2(0,\infty)$, the right-hand side above converges to zero as $t \to \infty$. Similarly, since $f' \in \mathbb{L}^2$, $\lim_{t\to\infty} \|f'(t+\cdot)\|_{\mathbb{L}^2} = 0$, and (2.2.2) follows. $\square$

COROLLARY 2.6. $\mathcal{Y}$ is a closed subspace of $\mathbb{R} \times \mathbb{H}^1(0,\infty)$ and hence, is a Polish space.

PROOF. Note that the mapping $f \mapsto f(0)$ is continuous on $\mathbb{C}_{\mathbb{R}}[0,\infty)$, and hence continuous on $\mathbb{H}^1(0,\infty)$, by part b of Lemma 2.5. Since $x \mapsto x \wedge 0$ is also continuous on $\mathbb{R}$, the set

$$\mathcal{Y} = \{(x,f) \in \mathbb{R} \times \mathbb{H}^1(0,\infty) : f(0) - x \wedge 0 = 0\}$$

is the pre-image of the close set $\{0\}$ under a continuous map, and hence is closed. $\square$

We equip $\mathcal{Y}$ with the corresponding Borel σ-algebra.

2.2.2. Definition of the State Process $Y^{(N)}$. For every $t \geq 0$, recall that $X_t^{(N)}$ is the total number of jobs in the system. As mentioned in the introduction, we introduce a novel representation for the $GI/GI/N$ queue, in which we append a function-valued variable $Z^{(N)}$ to the number of jobs $X^{(N)}$. This representation, when suitably centered and re-scaled, leads to a Markovian limit process in the Halfin-Whitt regime, and enables us to prove convergence of the steady-state distribution of $X^{(N)}$.

For every $t \geq 0$, let $\mathcal{S}^{(N)}(t)$ denote the set of indices of jobs that are receiving service at time t, and for every job $j \in \mathcal{S}^{(N)}(t)$, let recall that $a_j^{(N)}(t)$ is the age of that job at time t. Define the process $\{Z_t^{(N)}; t \geq 0\}$ where for each $t \geq 0$, $r \mapsto Z_t^{(N)}(r)$ is the random function on $[0,\infty)$ given as

(2.2.4)
$$Z_t^{(N)}(r) \doteq \sum_{j \in \mathcal{S}^{(N)}(t)} \frac{\overline{G}(a_j^{(N)}(t)+r)}{\overline{G}(a_j^{(N)}(t))} = \sum_{j=j_0^{(N)}}^{K_t^{(N)}} \frac{\overline{G}(a_j^{(N)}(t)+r)}{\overline{G}(a_j^{(N)}(t))} 1_{\{a_j^{(N)}(t)<v_j\}}. ; \qquad r \geq 0,$$

Note that for every job $j \in \mathcal{S}^{(N)}(t)$, the quantity $\overline{G}(a_j^{(N)}(t)+r)/\overline{G}(a_j^{(N)}(t))$ is the conditional probability that the job j has not yet departed the network by the time $t+r$, given its current age $a_j^{(N)}(t)$. Therefore, $Z_t^{(N)}(r)$ is, roughly speaking, the conditional expected fraction of jobs receiving service at time t, that will still be in the system at time $t+r$, given the ages of all jobs at time t.

Now define the process $\{Y_t^{(N)}; t \geq 0\}$ as

$$Y_t^{(N)} \doteq (X_t^{(N)}, Z_t^{(N)}). \tag{2.2.5}$$

As shown in Theorem 2.7, $Y_t^{(N)}$ takes values in a the space $\mathcal{Y}$, defined as in Section 2.2.1.

The centered and diffusion-scaled version of the state representation $Y^{(N)}$ is defined as follows. For $H = X, Z, Y$, define

$$\widehat{H}^{(N)} \doteq \frac{H^{(N)} - N\overline{H}}{\sqrt{N}}, \tag{2.2.6}$$

where $\overline{H}$ is the fluid limit, defined as

$$\overline{X} \doteq 1, \qquad \overline{Z}(r) \doteq \int_0^\infty \overline{G}(x+r)dx; \;\; r \geq 0, \quad \text{and} \quad \overline{Y} \doteq (\overline{X}, \overline{Z}). \tag{2.2.7}$$

The definition of $\overline{Y}$ is inspired by the fluid limit theorem established in [**61**, Theorem 3.7] and will be justified in Section 2.3.1 (see Corollary 2.11.) Our first main result is on the long-time behavior of $\widehat{Y}^{(N)}$ for a fixed N.

THEOREM 2.7. Suppose Assumptions I-III hold. Then for every $N \in \mathbb{N}$ and $t \geq 0$, $\widehat{Y}_t^{(N)} \in \mathcal{Y}$. Moreover, there exists a measure $\widehat{\pi}^{(N)}$ on $\mathcal{Y}$ such that for every initial condition $(u_0^{(N)}, X_0^{(N)}, a_j^{(N)}(0); j \in \mathcal{S}^{(N)}(0))$, $\widehat{Y}_t^{(N)} \Rightarrow \widehat{\pi}^{(N)}$ in $\mathcal{Y}$ as $t \to \infty$.

PROOF. See the end of Section 2.2.4. $\square$

2.2.3. Relation between $Y^{(N)}$ and $V^{(N)}$. The representation $Y^{(N)}$ defined in (2.2.5) can be written as a functional of the Markovian state variable $V^{(N)}$. For brevity of notation, define a family $\{\Phi_t, t \geq 0\}$ of operators on the functions on $[0,\infty)$ as follows: for every $t \geq 0$,

$$(\Phi_t f)(x) = f(x+t)\,\frac{1-G(x+t)}{1-G(x)}, \quad x \in [0,\infty). \tag{2.2.8}$$

It is straightforward to see that bounded continuous functions are mapped in to bounded continuous functions by Φ_t, $\Phi_0 f = f$, and the following semigroup property holds:

$$\Phi_t \Phi_s = \Phi_{t+s}, \qquad t,s \geq 0. \tag{2.2.9}$$

Next, motivated by the definition (2.2.4) of $Z^{(N)}$, define the mapping $\mathcal{T}$ on $\mathbb{M}_F[0,\infty)$ by

$$(2.2.10) \qquad (\mathcal{T}\mu)(r) = \mu(\Phi_r \mathbf{1}), \qquad r \geq 0.$$

The following lemma studies properties of $\mathcal{T}$.

LEMMA 2.8. Suppose Assumptions II and III hold. Then, for every $\mu \in \mathbb{M}_D[0,\infty)$, $\mathcal{T}\mu$ lies in $\mathbb{H}^1(0,\infty)$ and has density

$$(2.2.11) \qquad (\mathcal{T}\mu)'(r) = -\mu(\Phi_r h), \qquad r \geq 0.$$

Also, the mapping $\mathcal{T} : \mathbb{M}_D[0,\infty) \to \mathbb{H}^1(0,\infty)$ is continuous.

PROOF. For the first assertion, let $\mu = \sum_{j=1}^m \delta_{a_j}$, for some $m \in \mathbb{N}$ and $a_j \in \mathbb{R}_+$, $j = 1, ..., m$, and let $z = \mathcal{T}\mu$. Then we have

$$z(r) = \mu(\Phi_r \mathbf{1}) = \sum_{j=1}^m \delta_{a_j}(\Phi_r \mathbf{1}) = \sum_{j=1}^m \frac{\overline{G}(a_j + r)}{\overline{G}(a_j)}, \qquad r \geq 0.$$

Since $\overline{G}$ has density g, z has density z' given by

$$z'(r) = -\sum_{j=1}^m \frac{g(a_j + r)}{\overline{G}(a_j)} = -\sum_{j=1}^m \delta_{a_j}(\Phi_r h) = -\mu(\Phi_r h), \qquad r \geq 0.$$

Both $\overline{G}$ and g are in $\mathbb{L}^2(0,\infty)$ by Assumption III (see Remark 2.3), and hence, both z and z' also lie in $\mathbb{L}^2(0,\infty)$; thus $z \in \mathbb{H}^1(0,\infty)$.

For the second claim, let $\{\mu^k\}$ be a sequence in $\mathbb{M}_D[0,\infty)$ converging weakly to $\mu = \sum_{j=1}^m \delta_{a_j}$. Then, for large k, μ^k has a representation $\mu^k = \sum_{j=1}^m \delta_{a_j^k}$ for some $a_j^k > 0$, with $a_j^k \to a_j$ as $k \to \infty$. Therefore,

$$\begin{aligned}
\|\mathcal{T}\mu^k - \mathcal{T}\mu\|_{\mathbb{L}^2} &= \left\| \sum_{j=1}^m \frac{\overline{G}(a_j^k + \cdot)}{\overline{G}(a_j^k)} - \frac{\overline{G}(a_j + \cdot)}{\overline{G}(a_j)} \right\|_{\mathbb{L}^2} \\
&\leq \sum_{j=1}^m \left\| \frac{\overline{G}(a_j^k + \cdot)}{\overline{G}(a_j^k)} - \frac{\overline{G}(a_j + \cdot)}{\overline{G}(a_j^k)} + \frac{\overline{G}(a_j + \cdot)}{\overline{G}(a_j^k)} - \frac{\overline{G}(a_j + \cdot)}{\overline{G}(a_j)} \right\|_{\mathbb{L}^2} \\
&\leq \sum_{j=1}^m \frac{\|\overline{G}(a_j^k + \cdot) - \overline{G}(a_j - \cdot)\|_{\mathbb{L}^2}}{\overline{G}(a_j^k)} + \|\overline{G}\|_{\mathbb{L}^2} \sum_{j=1}^m \left| \frac{1}{\overline{G}(a_j^k)} - \frac{1}{\overline{G}(a_j)} \right|,
\end{aligned}$$

which converges to zero as $k \to \infty$ by the continuity of $\overline{G}$ and the continuity of the translation mapping in the $\mathbb{L}^2$ norm (see e.g. [**28**, Theorem 20.1]). Similarly,

$$\| \left(\mathcal{T}\mu^k\right)' - (\mathcal{T}\mu)' \|_{\mathbb{L}^2} \leq \sum_{j=1}^{M} \frac{\|g(a_j^k + \cdot) - g(a_j - \cdot)\|_{\mathbb{L}^2}}{\overline{G}(a_j^k)} + \|g\|_{\mathbb{L}^2} \sum_{j=1}^{m} \left| \frac{1}{\overline{G}(a_j^k)} - \frac{1}{\overline{G}(a_j)} \right|,$$

which again converges to zero as $k \to \infty$ by the same argument, using the fact that $g \in \mathbb{L}^2(0,\infty)$, as discussed above. $\square$

Using this notation, we can rewrite the quantity $Z_t^{(N)}$ defined in (2.2.4) as

$$Z_t^{(N)} = \mathcal{T}\nu_t^{(N)} = \nu_t^{(N)}(\Phi.\mathbf{1}), \tag{2.2.12}$$

and by Lemma 2.8, $Z_t^{(N)}$ lies in $\mathbb{H}^1(0,\infty)$. Consequently,

$$Y_t^{(N)} = (X_t^{(N)}, \mathcal{T}\nu_t^{(N)}). \tag{2.2.13}$$

Recall the definition (2.2.7) of $\overline{X}$ and $\overline{Y}$, and define the mapping $\widehat{\Psi} : \mathcal{V}^{(N)} \mapsto \mathbb{R} \times \mathbb{H}^1(0,\infty)$ as

$$\widehat{\Psi}(\alpha, x, \mu) \doteq \left(\frac{x - N\overline{X}}{\sqrt{N}}, \frac{\mathcal{T}(\mu) - N\overline{Z}}{\sqrt{N}} \right). \tag{2.2.14}$$

Using this notation, and by equation (2.2.13), the definition (2.2.6) of $\widehat{Y}_t^{(N)}$ can be written as

$$\widehat{Y}_t^{(N)} = \left(\frac{X_t^{(N)} - N\overline{X}}{\sqrt{N}}, \frac{\mathcal{T}(\nu_t^{(N)}) - N\overline{Z}}{\sqrt{N}} \right) = \widehat{\Psi}(V_t^{(N)}), \qquad t \geq 0. \tag{2.2.15}$$

2.2.4. Long-Time behavior of $\widehat{Y}^{(N)}$. Theorem 2.7 is a consequence of the ergodicity of the Markovian state process $V^{(N)}$, and the relation (2.2.13) between $Y^{(N)}$ and $V^{(N)}$. First, we prove that the Markov process $V^{(N)}$ is ergodic. This essentially follows from existing results in the literature. The proof in the presence of reneging is given in [**58**, Theorem 7.1], and ergodicity of another Markovian representation have been established in [**3**, Section XII.2].

PROPOSITION 2.9. Let Assumptions I and II hold. Then the Markov process $\{V_t^{(N)}, \mathcal{F}_t^{(N)}; t \geq 0\}$ is ergodic. In particular, there exists a $\mathcal{V}^{(N)}$-valued random variable $V_\infty^{(N)} \doteq (R_\infty^{(N)}, X_\infty^{(N)}, \nu_\infty^{(N)})$ such that for any initial condition $V_0^{(N)}$, $V_t^{(N)} \Rightarrow V_\infty^{(N)}$ in $\mathcal{V}$ as $t \mapsto \infty$.

PROOF. First note that by (2.1.1), the traffic intensity $\rho^{(N)} = \lambda^{(N)}/N$ satisfies $\rho^{(N)} < 1$. As discussed in Section XII of [**3**], the queueing system described in this paper can be viewed as an N server queueing system where each server has its own queue, and each job, upon arrival, is routed to the queue with the least remaining workload (i.e., the residual service time of the job in service, plus the sum of service times of all jobs in that queue). The stability of this queueing system under the condition $\rho^{(N)} < 1$ is studied in [**3**], using a different Markovian representation and a different filtration. However, some of their results are still relevant for our representation.

For the rest of the proof, we fix $N \in \mathbb{N}$, and suppress the superscript (N) for brevity. Let $\tilde{R}$ be the backward recurrence time of the arrival process, that is,

$$\tilde{R}_t \doteq t - \sup\{s < t : E_s < E_t\}; \qquad t \geq 0.$$

The process $\{\tilde{V}_t = (\tilde{R}_t, X_t, \nu_t); t \geq 0\}$ is also a càdlàg Markov process. Let $\{\sigma(k); k \geq 1\}$ be the sequence of jobs such that the system is empty right before their arrival time $T_k \doteq u_0 + \cdots + u_{\sigma(k)-1}$, that is, $\tilde{V}_{T_k-} = (0, 0, \mathbf{0})$, where $\mathbf{0}$ is the zero measure. Note that $\beta_- \doteq \sup\{x : G(x) = 0\}$ is finite because $\lim_{x\to\infty} G(x) = 1$, and $\alpha_+ \doteq \sup\{x : \overline{G}_E(x) > 0\} = \infty$ because $\tilde{G}_E$ (or equivalently, G_E) has full support on $[0, \infty)$ by Assumption I. Therefore, $\beta_- < \alpha_+$, and since $\rho < 1$, we can apply Corollary XII.2.5 of [**3**] and conclude that $\{\sigma(k); k \geq 1\}$ is an aperiodic, nonterminating (discrete-time) renewal process with finite mean renewal times. Moreover, $\tilde{G}_E$, and thus, G_E, are non-lattice by Assumption I, and another application of the fact that $\beta_- < \alpha_+$ established above implies $\mathbb{P}\{v_1 < u_1\} > 0$. Therefore, as shown in the proof of Corollary XII.2.8.(a) in [**3**, page 348], we have $m \doteq \mathbb{E}_0\{T_1\} < \infty$, where $\mathbb{E}_0$ is the expectation under the initial condition $\tilde{V}_0 = (0, 0, \mathbf{0})$. Therefore, $\tilde{V}$ is a càdlàg regenerative Markov process, with regeneration set $\{(0, 0, \mathbf{0})\}$, regenerative points T_k, and cycle lengths $T_k - T_{k-1}$

with non-lattice distribution and finite mean m. Therefore, by [**3**, Theorem VI.1.2], there exists a random variable $V_\infty = (R_\infty, X_\infty, \nu_\infty)$ such that as $t \to \infty$, $\tilde{V}_t \Rightarrow V_\infty$ in $\mathcal{V}$, and the law π_∞ of V_∞ is given by

$$\pi_\infty(f) = \frac{1}{m}\mathbb{E}_0\left\{\int_0^{T_1} f(\tilde{V}_s)ds\right\}, \qquad \forall f \in \mathbb{C}_b(\mathcal{V}).$$

Moreover, by [**3**, Theorem VII.3.2], π_∞ is a stationary distribution for the Markov process $\tilde{V}$. In particular, the law α_∞ of R_∞ is the stationary distribution for the backward recurrence time of the arrival process, and thus has the representation $\alpha_\infty(dx) = 1/\lambda \overline{G}_E(x)dx$ by [**3**, Proposition V.3.5].

Now recall from (3.2) in [**3**, Page 151] that the distributions of R_t and $\tilde{R}_t$ satisfy the relation

$$\mathbb{P}\left\{R_t > y | \tilde{R}_t = x\right\} = \frac{\overline{G}_E(x+y)}{\overline{G}_E(x)}.$$

Hence, for every bounded and continuous function f,

$$\begin{aligned}
\mathbb{E}\left[f(R_t)\right] &= \int_0^\infty f(y)\mathbb{P}\left\{R_t = dy\right\} \\
&= \int_0^\infty \int_0^\infty f(y)\frac{g_E(x+y)}{\overline{G}_E(x)}dy\, \mathbb{P}\left\{\tilde{R}_t = dx\right\} \\
&= \mathbb{E}\left[\tilde{f}(\tilde{R}_t)\right],
\end{aligned}$$

where $\tilde{f}(x) = \int_0^\infty g_E(x+y)f(y)dy/\overline{G}_E(x)$ is continuous and bounded by $\|f\|_\infty$. Therefore, for every $f \in \mathbb{C}_b[0,\infty)$, recalling the expression for α_∞, we see that as $t \to \infty$,

$$\begin{aligned}
\mathbb{E}\left[f(R_t)\right] \to \mathbb{E}\left[\tilde{f}(R_\infty)\right] &= \frac{1}{\lambda}\int_0^\infty \int_0^\infty f(y)\frac{g_E(x+y)}{\overline{G}_E(x)}dy\overline{G}_E(x)dx \\
&= \frac{1}{\lambda}\int_0^\infty f(y)\overline{G}_E(y)dy \\
&= \mathbb{E}\left[f(R_\infty)\right].
\end{aligned}$$

Therefore, $R_t \Rightarrow R_\infty$, and the law α_∞ of R_∞ is also a stationary distribution for the forward recurrence time R. We conclude that π_∞ is a stationary distribution for V, and for every initial condition V_0, V_t converges in distribution to V_∞ as $t \to \infty$. $\square$

Recall the definition (2.2.14) of $\widehat{\Psi}$, and analogous to (2.2.15), define

$$\left(\widehat{X}_{\infty}^{(N)}, \widehat{Z}_{\infty}^{(N)}\right) \doteq \left(\frac{X_{\infty}^{(N)} - N\overline{X}}{\sqrt{N}}, \frac{\mathcal{T}(\nu_{\infty}^{(N)}) - N\overline{Z}}{\sqrt{N}}\right) = \widehat{\Psi}(V_{\infty}^{(N)}).$$

We show that Theorem 2.7 holds for

$$\widehat{\pi}^{(N)} \doteq \mathscr{L}aw(\widehat{X}_{\infty}^{(N)}, \widehat{Z}_{\infty}^{(N)}) = \mathscr{L}aw\left(\widehat{\Psi}(V_{\infty}^{(N)})\right) \tag{2.2.16}$$

PROOF OF THEOREM 2.7. To prove the first assertion, fix $N \in \mathbb{N}$ and $t \geq 0$. By Assumption II.b, the function $\overline{Z}(r) = \int_r^{\infty} \overline{G}(x)dx$ lies in $\mathbb{L}^2(0,\infty)$, and is absolutely continuous with density $-\overline{G}(\cdot)$ which also lies in $\mathbb{L}^2(0,\infty)$ (see Remark 2.3). Moreover, $Z_t^{(N)} = \mathcal{T}(\nu_t^{(N)})$ (see (2.2.12)) lies in $\mathbb{H}^1(0,\infty)$ by Lemma 2.8, and hence, so does $\widehat{Z}_t^{(N)}$. Therefore, $\widehat{Y}_t^{(N)} = (\widehat{X}_t^{(N)}, \widehat{Z}_t^{(N)})$ takes values in $\mathbb{R} \times \mathbb{H}^1(0,\infty)$.

Substituting $r = 0$ in (2.2.12), we have $Z_t^{(N)}(0) = \nu_t^{(N)}(\mathbf{1})$, and hence, the non-idling condition (2.1.8) can be written as $Z_t^{(N)}(0) = N \wedge X_t^{(N)}$. Also, by definition of $\overline{Z}$ and since the distribution G has mean equal to one, we have $\overline{Z}(0) = \int_0^{\infty} \overline{G}(x)dx = 1$. Therefore,

$$\widehat{Z}_t^{(N)}(0) = \frac{X_t^{(N)} \wedge N - N}{\sqrt{N}} = \frac{(X_t^{(N)} - N) \wedge 0}{\sqrt{N}} = \widehat{X}_t^{(N)} \wedge 0 = -(\widehat{X}_t^{(N)})^{-}, \tag{2.2.17}$$

and therefore $\widehat{Y}_t^{(N)}$ lies in $\mathcal{Y}$.

To prove the second claim, recall from Proposition 2.9 that $V_t^{(N)}$ converges in distribution to $V_{\infty}^{(N)}$ as $t \to \infty$. Also, the mapping $\widehat{\Psi}$ defined in (2.2.14) is continuous by Lemma 2.8, and by the fact that $\overline{Z} \in \mathbb{H}^1(0,\infty)$. Therefore, the result follows from Proposition 2.9, the representation (2.2.15) of $\widehat{Y}_t^{(N)}$, definition (2.2.16) of $\widehat{\pi}^{(N)}$, and the continuous mapping theorem. $\square$

2.3. A Fluid Limit Theorem and Dynamics of the State Variables

In this section, we establish a fluid limit theorem for the process $Y^{(N)}$, and derive a system of equations which describe the evolution of $Y^{(N)}$ and its centered and renormalized counterpart $\widehat{Y}^{(N)}$. These equations are used later in Chapter 4.

Section 2.3.1 contains the fluid limit result for $Y^{(N)}$ (Lemma 2.10), which is established by invoking a series of results in **[61]**. Then in Section 2.3.2, we recall the dynamics of the Markovian state process $V^{(N)}$ studied in [**62**]. The main result of Section 2.3.3 is Proposition 2.14, in which we derive the equations governing the dynamics of $\widehat{Y}^{(N)}$.

2.3.1. A Fluid Limit Theorem. A fluid limit theorem is established in [**61**] for $(X^{(N)}, \nu^{(N)})$ under certain assumptions on the sequence of initial conditions. By Assumptions I and II, the load $\rho^{(N)} \doteq \lambda^N/N$ approaches to 1 as $N \to \infty$, which is referred to in [**61**] as the *critical case*. For $H = X, \nu, E, \lambda, D, K$, define

$$\overline{H}^{(N)} \doteq \frac{H^{(N)}}{N}. \tag{2.3.1}$$

Also, recall from (2.2.7) that $\overline{X} = 1$, define

$$\overline{E} \doteq \mathbf{Id} \tag{2.3.2}$$

and define the measure $\overline{\nu} \in \mathbb{M}_F[0, \infty)$ as

$$\overline{\nu}(A) \doteq \int_A \overline{G}(x)dx, \qquad \forall A \in \mathcal{B}[0, \infty). \tag{2.3.3}$$

LEMMA 2.10 (Fluid Limit). Suppose Assumptions I-III hold. Then, as $N \to \infty$,

(a) $\overline{E}^{(N)} \to \overline{E}$ almost surely in $\mathbb{D}_{\mathbb{R}}[0, \infty)$, and $\mathbb{E}[\overline{E}_t^{(N)}] \to t$ for every $t \geq 0$;

(b) if the sequence of initial conditions satisfy

$$(\overline{X}_0^{(N)}, \overline{\nu}_0^{(N)}) \Rightarrow (\overline{X}, \overline{\nu}) \text{ in } \mathbb{R} \times \mathbb{M}_F[0, \infty), \quad \text{and} \quad \mathbb{E}[\overline{X}_0^{(N)}] \to \overline{X}, \tag{2.3.4}$$

as $N \to \infty$, then $(\overline{X}_t^{(N)}, \overline{\nu}_t^{(N)}) \Rightarrow (\overline{X}, \overline{\nu})$ in $\mathbb{R} \times \mathbb{M}_F[0, \infty)$ for every $t \geq 0$.

PROOF. For part a, by the definition of $E^{(N)}$ in Assumption I we have

$$\overline{E}_t^{(N)} = \frac{E_t^{(N)}}{N} = \frac{\lambda^{(N)}}{N} \frac{\tilde{E}(\lambda^{(N)} t)}{\lambda^{(N)}}, \qquad t \geq 0.$$

Since $\lambda^{(N)} = N - \beta\sqrt{N}$, $\lambda^{(N)} \uparrow \infty$ and $\lambda^{(N)}/N \to 1$, by the functional Law of Large Numbers for renewal processes (see e.g. [**20**, Theorem 5.10]), the second term on the right-hand side of the display above converges to $\overline{E} = \mathbf{Id}$ almost surely in $\mathbb{D}_{\mathbb{R}}[0, \infty)$. Also, for every $t \geq 0$,

$$\lim_{N\to\infty} \mathbb{E}\left[\overline{E}_t^{(N)}\right] = \lim_{N\to\infty} \frac{\lambda^{(N)}}{N} \mathbb{E}\left[\frac{\tilde{E}(\lambda^{(N)} t)}{\lambda^{(N)}}\right] = t,$$

where the last equality is due to the elementary renewal theorem (see e.g. [**3**, Proposition V.1.4]).

For part b, note that Assumption 1 in [**61**] holds by part a. and the fact that initial conditions satisfy (2.3.4). On the other hand, Assumption 2 in [**61**] holds because Assumption II ensures that h is continuous. Thus, by Theorem 3.7 in [**61**], $(\overline{X}^{(N)}, \overline{\nu}^{(N)})$ converges in distribution to the unique solution of the fluid equations (3.4) - (3.7) there with initial conditions $(\overline{E}, \overline{X}, \overline{\nu})$. The result then follows from the fact that $(\overline{E}, \overline{X}, \overline{\nu})$ is a fixed point of the fluid equations (see Remark 3.8 in [**61**]). (Note that Assumption 1 in [**61**] requires the convergence in (2.3.4) to hold almost surely, and convergence result in Theorem 3.7 also holds in almost sure sense. However, as discussed in Remark 3.2 in [**61**], the result of Theorem 3.7 holds in distribution if the convergence in (2.3.4) holds in distribution.) □

COROLLARY 2.11. Under the assumptions of Lemma 2.10, $\overline{Y}^{(N)} = (\overline{X}^{(N)}, \overline{Z}^{(N)}) \Rightarrow \overline{Y} = (\overline{X}, \overline{Z})$ in $\mathbb{R} \times \mathbb{H}^1(0,\infty)$, as $N \to \infty$.

PROOF. Recall the representation (2.2.12) of $Z^{(N)}$ and the definition of $\overline{\nu}$ in 2.3.3, and note that by definition (2.2.10) of the mapping $\mathcal{T}$, $\overline{Z}$ defined in (2.2.7) satisfies $\overline{Z} = \mathcal{T}(\overline{\nu})$. The result then follows from Lemma 2.10, the continuity of the mapping $\mathcal{T}$ (Lemma 2.8), and an application of the continuous mapping theorem. □

2.3.2. Dynamics of $V^{(N)}$ and Auxiliary Processes. Dynamics of the Markovian state variable $V^{(N)}$ is studied [**61**, Section 5.1] and [**62**, Section 6]. Here, we present a summary of their definitions and result.

Let $\{D_t^{(N)}; t \geq 0\}$ be the cumulative departure process, that is, $D_t^{(N)}$ denotes the total number of departures from the system during $[0,t]$. For any bounded continuous function φ defined on $[0,\infty) \times [0,\infty)$, define the φ-weighted departure process as

$$\mathcal{Q}_\varphi^{(N)}(t) \doteq \sum_{j=j_0^{(N)}}^{K_t^{(N)}} \mathbb{1}\left(\gamma_j^{(N)} \leq t\right) \varphi(v_j, \gamma_j^{(N)}).$$

It has been shown in [**61**, Corollary 5.5] that the $\{\mathcal{F}_t^{(N)}\}$-compensator of $\mathcal{Q}_\varphi^{(N)}$ is the process

$$A_\varphi^{(N)}(t) \doteq \int_0^t \left(\int_{[0,\infty)} \varphi(x,s) h(x) \nu_s^{(N)}(dx) \right) ds, \quad t \in [0,\infty),$$

and the cádlág local $\{\mathcal{F}_t^{(N)}\}$-martingale $M_\varphi^{(N)} \doteq \mathcal{Q}_\varphi^{(N)} - A_\varphi^{(N)}$ has the predictable quadratic variation $\langle M_\varphi^{(N)} \rangle_t = A_{\varphi^2}^{(N)}(t)$. Now for every $t \in [0,\infty)$ and every Borel set $B \in \mathcal{B}[0,\infty)$, define

$$\mathcal{M}_t^{(N)}(B) \doteq M_{\mathbb{1}_B}^{(N)}(t). \tag{2.3.5}$$

It is shown in Corollary 4.3 of [**62**] that $\mathcal{M}^{(N)} = \{\mathcal{M}_t^{(N)}(B); t \in [0,\infty), B \in \mathcal{B}_0[0,\infty)\}$ is an orthogonal martingale measure with covariance functional

$$\langle \mathcal{M}^{(N)}(B), \mathcal{M}^{(N)}(\tilde{B}) \rangle_t = \int_0^t \left(\int_{B \cap \tilde{B}} h(x) \nu_s^{(N)}(dx) \right) ds.$$

The stochastic integral of a continuous and bounded function $\varphi : [0,\infty) \times [0,\infty) \mapsto \mathbb{R}$ with respect to $\mathcal{M}^{(N)}$ is denoted by $\mathcal{M}^{(N)}(\varphi)$, which is itself a martingale with quadratic variation process

$$\langle \mathcal{M}^{(N)}(\varphi) \rangle_t = \int_0^t \int_0^\infty \varphi^2(x,s) h(x) \nu_s^{(N)}(dx)\, ds, \qquad t \geq 0. \tag{2.3.6}$$

For definitions and properties of martingale measures and space-time white noise, and the corresponding stochastic integrals, we refer the reader to [**98**, Chapters 1 and 2].

Next, define another family of operators $\Psi_t, t \geq 0$, taking functions on $[0,\infty)$ to functions on $[0,\infty) \times [0,\infty)$, as follows: for every $t \geq 0$,

$$(\Psi_t f)(x,s) = f(x + (t-s)^+)\, \frac{1 - G(x + (t-s)^+)}{1 - G(x)}, \qquad (x,s) \in [0,\infty) \times [0,\infty). \tag{2.3.7}$$

In particular, Ψ_t takes continuous bounded functions to continuous bounded functions and has the property that for every $t, s \geq 0$ and function f on $[0,\infty)$, $\|\Psi_t f\|_\infty \leq \|f\|_\infty$, and

$$(\Psi_s \Phi_t f)(x,u) = (\Psi_{s+t} f)(x,u), \quad (x,u) \in [0,\infty) \times [0,s]. \tag{2.3.8}$$

Using this notation, for $f \in \mathbb{C}_b[0,\infty)$ and $t > 0$, define the stochastic convolution integral $\mathcal{H}_t^{(N)}(f)$ by

$$\mathcal{H}_t^{(N)}(f) \doteq \mathcal{M}_t^{(N)}(\Psi_t f). \tag{2.3.9}$$

Now we have introduced the required notation to describe the dynamics of N-server systems. Using equation (6.14) and definitions (5.10) (with $\widehat{\nu}^{(N)}$ replaced by $\nu^{(N)}$) and (6.11) (with $\widehat{\mathcal{K}}^{(N)}$ and $\widehat{K}^{(N)}$ replaced by $\mathcal{K}^{(N)}$ and $K^{(N)}$, respectively,) in **[62]**, for every $t \geq 0$ and every bounded, absolutely continuous function f, we have

$$\nu_t^{(N)}(f) = \nu_0^{(N)}(\Phi_t f) - \mathcal{H}_t^{(N)}(f) + \int_0^t (f\overline{G})(t-s)dK_s^{(N)}, \tag{2.3.10}$$

where invoking equations (6.3) and (6.4) of **[62]**, we see that

$$K_t^{(N)} = -\left(X_t^{(N)} - N\right)^+ + \left(X_0^{(N)} - N\right)^+ + E_t^{(N)}. \tag{2.3.11}$$

2.3.3. Equations Governing the Evolution of $\widehat{Y}^{(N)}$. To express the equations governing the evolution of $\widehat{Y}^{(N)}$, we recall the centered many-server (CMS) mapping (for the so-called critical case) introduced in **[62**, Definition 5.4**]**. Denote by $\mathbb{D}^0_{\mathbb{R}}[0,\infty)$ the space of functions $f \in \mathbb{D}_{\mathbb{R}}[0,\infty)$ with $f(0) = 0$.

DEFINITION 2.12. (**Centered Many-Server Mapping**) Given (η, x_0, ζ) in the space $\mathbb{D}^0_{\mathbb{R}}[0,\infty) \times \mathbb{R} \times \mathbb{D}_{\mathbb{R}}[0,\infty)$ with $\zeta(0) = x_0 \wedge 0$, a pair $(\kappa, x) \in \mathbb{D}^0_{\mathbb{R}}[0,\infty) \times \mathbb{D}_{\mathbb{R}}[0,\infty)$ is said to solve the centered many-server equations associated with (η, x_0, ζ) if it satisfies

$$x(t) \wedge 0 = \zeta(t) + \kappa(t) - \int_0^t g(t-s)\kappa(s)ds, \tag{2.3.12}$$

$$\kappa(t) = \eta(t) - x^+(t) + x_0^+. \tag{2.3.13}$$

for all $t \geq 0$. When such a solution exists and is unique, for every input (η, x_0, ζ), the mapping $\Lambda : \mathbb{D}^0_{\mathbb{R}}[0,\infty) \times \mathbb{R} \times \mathbb{D}_{\mathbb{R}}[0,\infty) \mapsto \mathbb{D}^0_{\mathbb{R}}[0,\infty) \times \mathbb{D}_{\mathbb{R}}[0,\infty)$ takes (η, x_0, ζ) to the solution (κ, x) is called the Centered Many-Server (CSM) Mapping. The set of (η, x_0, ζ) for which a solution exists is denoted by $\mathrm{dom}(\Lambda)$.

LEMMA 2.13. The mapping Λ is single-valued on dom(Λ), is continuous with respect to the topology of uniform convergence on compact sets on $\mathbb{D}_{\mathbb{R}}[0,\infty)$ and is measurable with respect to the Skorokhod topology on $\mathbb{D}_{\mathbb{R}}[0,\infty)$. Furthermore, the mapping $\Lambda : \mathbb{C}^0[0,\infty) \times \mathbb{R} \times \mathbb{C}_{\mathbb{R}}[0,\infty) \mapsto \mathbb{C}^0[0,\infty) \times \mathbb{C}_{\mathbb{R}}[0,\infty)$ that takes (η, x_0, ζ) to the solution (κ, x) is continuous and non-anticipative, that is, for every $T > 0$ and $(x_i, \kappa_i) = \Lambda(\eta_i, x_0, \zeta_i), i = 1, 2$, if (η_1, ζ_1) and (η_2, ζ_2) are equal on $[0, T]$, then (x_1, κ_1) coincides with (x_2, κ_2) on $[0, T]$.

PROOF. The facts that Λ is single-valued and continuous on its domain are proved in Proposition 7.3 of **[62]**, and the measurability claim is proved in Lemma 7.4 of **[62]**.

For the last claim, fix $(\eta, x_0, \zeta) \in \mathbb{C}^0[0,\infty) \times \mathbb{R} \times \mathbb{C}_{\mathbb{R}}[0,\infty)$ with $\zeta(0) = x_0 \wedge 0$, and set

$$r(t) \doteq \zeta(t) + \eta(t) - \int_0^t g(t-s)\eta(s)ds + \overline{G}(t)x_0^+, \quad t \geq 0. \tag{2.3.14}$$

Then, substituting κ from (2.3.13) into (2.3.12), it is straightforward to see that (x, κ) satisfy (2.3.12) and (2.3.13) if and only if x satisfies the Volterra equation

$$x(t) = r(t) + \int_0^t g(t-s)x^+(s)ds, \quad t \geq 0, \tag{2.3.15}$$

and κ satisfies (2.3.13). However, since $F(x) = x^+$ is Lipschitz and the assumptions on (η, x_0, ζ) imply that r is continuous and $r(0) = x_0$, Theorem 3.2.1 in **[18]** shows that there exists a unique solution $\bar{x}$ to (2.3.15). Defining $\bar{\kappa}$ as in (2.3.13), with x replaced by $\bar{x}$, it follows that $(\bar{x}, \bar{\kappa})$ is the unique solution to the equations (2.3.12)-(2.3.13) associated with (η, x_0, ζ). Let Λ denote the map that takes (η, x_0, ζ) to this unique solution.

The continuity of Λ follows from Proposition 7.3 in **[62]**. To prove the non-anticipative property, for $i = 1, 2$, define r_i by (2.3.14) with η and ζ replaced by η_i and ζ_i. We need to show that for every $t \in [0, T]$, if (η_1, ζ_1) and (η_2, ζ_2) agree on $[0, t]$, we have $r_1(t) = r_2(t)$. Subtracting equation (2.3.15) with x and r replaced by

x_i and r_i, $i = 1,2$, respectively, and defining $\Delta x = x_1 - x_2$, we have

$$\Delta x(t) = \int_0^t g(t-s)\left(x_1^+(s) - x_2^+(s)\right) ds, \qquad t \in [0,T].$$

Using the fact that the map $F(x) = x^+$ is Lipschitz with constant 1, we have

$$|\Delta x(t)| \leq \int_0^t g(t-s)|x_1^+(s) - x_2^+(s)|ds \leq \int_0^t g(t-s)|\Delta x(s)|ds \qquad t \in [0,T].$$

Since $\Delta x(0) = 0$, Gronwall's inequality implies $\Delta x(t) = 0$ for $t \in [0,T]$. Then, because κ_i satisfies (2.3.13) with η and x replaced by η_i and x_i, respectively, we also have $\kappa_1(t) = \kappa_2(t)$ for $t \in [0,T]$. $\square$

Define $\overline{K} \doteq \mathbf{Id}$, $\overline{\mathcal{M}} = 0$ and $\overline{\mathcal{H}} = 0$. Recall the definition (2.2.6) of $\widehat{H}^{(N)}$ for $H = X, Z, Y$, and define $\widehat{H}^{(N)}$ likewise for $H = E, K, \mathcal{M}, \mathcal{H}$.

PROPOSITION 2.14. Suppose Assumptions I-III hold, and the sequence of initial conditions satisfies (2.3.4). Then, for every $N \in \mathbb{N}$,

$$(\widehat{K}^{(N)}, \widehat{X}^{(N)}) = \Lambda(\widehat{E}^{(N)}, \widehat{X}_0^{(N)}, \widehat{Z}_0^{(N)} - \widehat{\mathcal{H}}^{(N)}(\mathbf{1})), \tag{2.3.16}$$

almost surely, where Λ is the CMS mapping defined in Definition 2.12, and in particular,

$$\widehat{K}_t^{(N)} = \widehat{E}_t^{(N)} - \widehat{X}_t^{(N)+} + \widehat{X}_0^{(N)+}, \qquad t \geq 0. \tag{2.3.17}$$

Also, for every $N \in \mathbb{N}$ and $t \geq 0$, $\widehat{Z}_t^{(N)}$ lies in $\mathbb{H}^1(0,\infty)$ and satisfies

(2.3.18)
$$\widehat{Z}_t^{(N)}(r) = \widehat{Z}_0^{(N)}(t+r) - \widehat{\mathcal{M}}_t^{(N)}(\Psi_{t+r}\mathbf{1}) + \widehat{K}_t^{(N)}\overline{G}(r) - \int_0^t \widehat{K}_s^{(N)} g(t-s+r)ds, \quad r \geq 0,$$

and its density $(\widehat{Z}_t^{(N)})'$ satisfies, for $r \geq 0$,

(2.3.19)
$$(\widehat{Z}_t^{(N)})'(r) = (\widehat{Z}_0^{(N)})'(t+r) + \widehat{\mathcal{M}}_t^{(N)}(\Psi_{t+r}h) - \widehat{K}_t^{(N)} g(r) - \int_0^t \widehat{K}_s^{(N)} g'(t-s+r)ds.$$

PROOF. Equation (2.3.16) is proved in Lemma 7.2 of [**62**] (the quantity $\widehat{Z}_0^{(N)}$ is denoted by $\mathcal{J}^{\widehat{\nu}_0^{(N)}}(\mathbf{1})$ in [**62**]). Note that Assumption 4 in [**62**], which is assumed in that lemma, holds under our Assumption III (see [**62**, Remark 5.2]). Equation (2.3.17) holds by definition of the mapping Λ.

Also, since $Z_t^{(N)} = \mathcal{T}(\nu_t^{(N)})$ by (2.2.12) and $\overline{Z} = \mathcal{T}(\overline{\nu})$ by Corollary 2.11, both $Z_t^{(N)}$ and $\overline{Z}$ lie in $\mathbb{H}^1(0,\infty)$ by Lemma 2.8, and hence, so does $\widehat{Z}_t^{(N)}$. Moreover, by the representation (2.2.12) of $Z^{(N)}$, substituting $f = \Phi_r \mathbf{1}$ in equation (2.3.10) and recalling that $d\overline{K}_s = ds$, and using properties (2.2.9) and (2.3.8) of $\{\Phi_s\}$ and $\{\Psi_s\}$ and definition (2.3.9) of $\mathcal{H}_t^{(N)}$, for every $t \geq 0$, we have

$$Z_t^{(N)}(r) = \nu_t^{(N)}(\Phi_r \mathbf{1}) = Z_0^{(N)}(t+r) - \mathcal{M}_t^{(N)}(\Psi_{t+r}\mathbf{1}) + \int_0^t \overline{G}(t-s+r)dK_s^{(N)}, \quad r \geq 0.$$

Also, by definition (2.2.7) of $\overline{Z}$,

$$\overline{Z}(r) = \int_r^\infty \overline{G}(s)ds = \int_{t+r}^\infty \overline{G}(s)ds + \int_r^{t+r} \overline{G}(s)ds = \overline{Z}(t+r) + \int_0^t \overline{G}(t-s+r)d\overline{K}_s.$$

Substituting $Z_t^{(N)}$ and $\overline{Z}$ from the last two displays in (2.2.6), recalling the definitions of $\widehat{\mathcal{M}}^{(N)}$ and $\widehat{K}^{(N)}$, and performing an integration by parts, (2.3.18) follows.

Moreover, by Lemma 2.8, $r \mapsto Z_t^{(N)}$ is absolutely continuous for each $t \geq 0$ with density $(Z_t^{(N)})'(r) = -\nu_t^{(N)}(\Phi_r h)$. The function h lies in $\mathbb{C}_b^1[0,\infty)$ by Assumption III, and hence, it is bounded and absolutely continuous. Substituting $f = h$ in (2.3.10), we have

$$(Z_t^{(N)})'(r) = (Z_0^{(N)})'(t+r) + \mathcal{M}_t^{(N)}(\Psi_{t+r}h) - \int_0^t g(t-s+r)dK_s^{(N)}, \quad r \geq 0.$$

Similarly, since $\overline{Z}(r) = \int_r^\infty \overline{G}(x)dx$, $\overline{Z}$ is differentiable with derivative $\overline{Z}'$, where

$$\overline{Z}'(r) = -\overline{G}(r) = -\int_r^\infty g(s)ds = \overline{Z}'(t+r) - \int_0^t g(t-s+r)d\overline{K}_s,$$

and the equation (2.3.19) follows from the last two displays. $\square$

Substituting $\widehat{K}^{(N)}$ from (2.3.17) in (2.3.18) and using integration by parts,

(2.3.20)

$$\widehat{Z}_t^{(N)}(r) = \widehat{Z}_0^{(N)}(t+r) - \widehat{\mathcal{M}}_t^{(N)}(\Psi_{t+r}\mathbf{1}) + \int_0^t \overline{G}(t+r-s)d\widehat{E}_s^{(N)} - \widehat{X}_t^{(N)+}\,\overline{G}(r)$$

(2.3.21) $$+ \int_0^t \widehat{X}_s^{(N)+} g(t+r-s)ds.$$

CHAPTER 3

Characterization of the Diffusion Limit Process

This chapter is devoted to study of an infinite-dimensional Markov process $Y = (X, Z)$ which serves as the limit of the sequence of representation processes $\{\widehat{Y}^{(N)}\}$. We characterize the components X and Z of the limit process to be the unique solution of a coupled pair of stochastic equations driven by a Brownian motion and an independent space-time white noise (see Definition 3.1). The stochastic equations that arise in the limit are non-standard; for example, boundary conditions appear in the drift term. We believe that SPDEs with this kind of boundary condition are likely to arise in the study of scaling limits and control of other parallel server networks, and so our constructions could be useful in a broader context.

The main result of this Chapter (Theorem 3.26) is to establish uniqueness of the invariant distribution of the Markov process $Y = (X, Z)$. As mentioned earlier, standard methods for establishing ergodicity of finite-dimensional Markov processes such as positive Harris recurrence are not well suited to this setting due to the infinite-dimensional nature of the state space. Instead, we adopt the asymptotic (equivalent) coupling approach developed in a series of papers by Hairer, Mattingly, Scheutzow and co-authors (see, e.g., **[30, 49, 6, 50]** and references therein) to show that (X, Z) has a unique invariant distribution. The asymptotic equivalent coupling that we construct has a somewhat different flavor from that used in previous works, and entails the analysis of a certain renewal equation.

This chapter is organized as follows. In Section 3.1, first we propose a coupled system of SPDEs, and then define an infinite-dimensional Markov process $Y = (X, Z)$, and claim that Y is the unique solution of the SPDE (Theorem 3.4).

Section 3.2 contains some technical results, including properties of certain stochastic integrals, which are required later in Section 3.3 where Theorem 3.4 is proved. Finally in Section 3.4, we establish the uniqueness of invariant distribution of the Markov process Y (Theorem 3.26).

3.1. Definition of the Diffusion Limit Process

In this section, we define an infinite-dimensional Markov process $Y = (X, Z)$, which we call the diffusion limit process, that serves as the limit of the sequence of representation processes $\{\widehat{Y}^{(N)}\}$. First, we propose a coupled system of SPDEs in Section 3.1.1. Then in Section 3.1.2, we define an infinite-dimensional Markov process $Y = (X, Z)$, and claim that Y is the unique solution of the SPDE (Theorem 3.4).

3.1.1. A Coupled System of Stochastic Partial Differential Equations. The coupled stochastic age equations are driven by a Brownian motion $B = \{B(t), t \geq 0\}$ and an independent continuous martingale measure $\mathcal{M} = \{\mathcal{M}_t(A),\ A \in \mathcal{B}[0,\infty), t \in [0,\infty)\}$ with deterministic covariance functional

$$\langle \mathcal{M}(A), \mathcal{M}(\tilde{A})\rangle_t = t \int_0^\infty \mathbb{1}_{A\cap\tilde{A}}(x)\, g(x)dx, \tag{3.1.1}$$

both defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In other words, $\mathcal{M}$ is a space-time white noise on $[0,\infty)^2$ based on the measure $gdx \otimes dt$. For definitions and properties of martingale measures and space-time white noise, we refer the reader to [**98**, Chapters 1 and 2]. For every bounded measurable function φ on $[0,\infty) \times [0,\infty)$, for $t \geq 0$, the stochastic integral of φ on $[0,\infty) \times [0,t]$ with respect to the martingale measure $\mathcal{M}$ is denoted by

$$\mathcal{M}_t(\varphi) \doteq \iint_{[0,\infty)\times[0,t]} \varphi(x,s)\mathcal{M}(dx, ds). \tag{3.1.2}$$

Also, let $\tilde{\mathcal{F}}_t \doteq \sigma(B(s), \mathcal{M}_s(A); 0 \leq s \leq t, A \in \mathcal{B}(0,\infty))$, and let the filtration $\{\mathcal{F}_t\}$ denote the augmentation of $\{\tilde{\mathcal{F}}_t\}$ with respect to $\mathbb{P}$. Recall that the stochastic integral $\{\mathcal{M}_t(\varphi), t \geq 0\}$ is an $\{\mathcal{F}_t\}$-martingale with (predictable) quadratic variation

$$\langle \mathcal{M}(\varphi)\rangle_t = \int_0^t \int_0^\infty \varphi^2(x,s) g(x) dx\, ds. \tag{3.1.3}$$

Recall the definition (2.2.8) of the family of mappings $\{\Phi_s; s \geq 0\}$.

DEFINITION 3.1 (**Coupled Stochastic Age Equations**). Let $Y_0 = (X_0, Z_0(\cdot))$ be a $\mathcal{Y}$-valued random element defined on $(\Omega, \mathcal{F}, \mathbb{P})$, independent of B and $\mathcal{M}$. A continuous $\mathcal{Y}$-valued stochastic process $Y = \{(X(t), Z(t,\cdot)); t \geq 0\}$ is said to be a solution of the *coupled stochastic age equations* with initial condition Y_0 if

(a) Y is $\{\mathcal{F}_t^{Y_0}\}$-adapted, where $\mathcal{F}_t^{Y_0} \doteq \mathcal{F}_t \vee \sigma(Y_0)$;

(b) $Y(0) = Y_0$, $\mathbb{P}$-almost surely;

(c) $\mathbb{P}$-almost surely, $\partial_r Z(\cdot,\cdot) : (0,\infty) \times (0,\infty) \mapsto \mathbb{R}$ is locally integrable and for every $t > 0$, the function

$$r \mapsto \int_0^t \partial_r Z(s,r) ds \tag{3.1.4}$$

is locally uniformly continuous. Therefore, this function can be continuously extended to $[0,\infty)$, and (with some abuse of notation) we denote its evaluation at 0 by $\int_0^t \partial_r Z(s,0) ds$.

(d) $\mathbb{P}$-almost surely, Z satisfies

(3.1.5)

$$\begin{aligned} Z(t,r) =& Z_0(r) - \overline{G}(r) Z_0(0) + \int_0^t \partial_r Z(s,r) ds - \overline{G}(r) \int_0^t \partial_r Z(s,0) ds \\ & - \mathcal{M}_t(\Phi_r \mathbf{1}) + \overline{G}(r) \mathcal{M}_t(\mathbf{1}) + \overline{G}(r) Z(t,0), \qquad \forall t, r \geq 0, \end{aligned}$$

subject to the boundary condition

$$Z(t,0) = X(t) \wedge 0, \qquad \forall t \geq 0, \tag{3.1.6}$$

where X satisfies the stochastic equation

$$X(t) = X_0 + B(t) - \beta t - \mathcal{M}_t(\mathbf{1}) + \int_0^t \partial_r Z(s,0) ds, \qquad \forall t \geq 0, \tag{3.1.7}$$

almost surely.

REMARK 3.2. Since the functions Z_0 and $Z(t,\cdot)$ lie in $\mathbb{H}^1(0,\infty)$ for every $t \geq 0$, they are locally uniformly continuous (see Lemma 2.5). Thus, they can be continuously extended to $[0,\infty)$ and the random variables $Z_0(0)$ and $Z(t,0)$ in (3.1.5) and (3.1.6) are well defined.

3.1.2. An Explicit Solution to the SPDE. We now describe the diffusion limit $Y = (X, Z)$. Recall the Definition (2.12) of the Centred Many-Server mapping Λ, and recall that B is a Brownian motion and for a fixed $\beta > 0$, define

$$E(t) \doteq B(t) - \beta t, \quad t \geq 0. \tag{3.1.8}$$

Moreover, we define a family of stochastic convolution integrals: for $t \geq 0$ and $f \in \mathbb{C}_b[0,\infty)$,

$$\mathcal{H}_t(f) \doteq \mathcal{M}_t(\Psi_t f) = \iint\limits_{[0,\infty)\times[0,t]} f(x+t-s)\,\frac{1-G(x+t-s)}{1-G(x)}\mathcal{M}(dx,ds). \tag{3.1.9}$$

Next, we introduce an auxiliary mapping, which appears in the explicit construction of the diffusion limit. For every $t \geq 0$, let Γ_t be the mapping on $\mathbb{C}_{\mathbb{R}}[0,\infty)$ defined as

$$(\Gamma_t \kappa)(r) \doteq \overline{G}(r)\kappa(t) - \int_0^t \kappa(s) g(t+r-s)ds, \qquad r \in [0,\infty), \tag{3.1.10}$$

for $\kappa \in \mathbb{C}_{\mathbb{R}}[0,\infty)$.

DEFINITION 3.3 (**Diffusion Limit**). For every random element $Y_0 = (X_0, Z_0) \in \mathcal{Y}$, the diffusion limit $Y_{Y_0} = \{Y_{Y_0}(t); t \geq 0\}$ is defined by $Y_{Y_0}(t) = (X(t), Z(t,\cdot))$, where

$$(K, X) \doteq \Lambda(E, X_0, Z_0 - \mathcal{H}(\mathbf{1})), \tag{3.1.11}$$

and for every $t, r \geq 0$,

$$Z(t,r) \doteq Z_0(t+r) - \mathcal{M}_t(\Psi_{t+r}\mathbf{1}) + \Gamma_t K(r), \tag{3.1.12}$$

where $\{\Psi_t; t \geq 0\}$ and $\{\Gamma_t; t \geq 0\}$ are defined by (2.3.7) and (3.1.10), respectively.

When the initial condition Y_0 is clear from the context, we will often omit the subscript Y_0 and just use Y to denote the diffusion limit . Also, deterministic initial conditions will be denoted by lower case: $y = (x_0, z_0) \in \mathcal{Y}$.

Here is the main result of this Section

THEOREM 3.4. Suppose Assumptions II and III hold. Then, for every random variable $Y_0 \in \mathcal{Y}$, the process Y_{Y_0} of Definition 3.3 is the unique solution to the coupled stochastic age equations with initial condition Y_0. Furthermore, $\{(Y_y(t), \mathcal{F}_t); t \geq 0, y \in \mathcal{Y}\}$ is a time-homogeneous Feller Markov family.

PROOF. The proof of Theorem 3.4 is given at the end of Section 3.3.3. $\square$

REMARK 3.5. Note that E and $\mathcal{H}(\mathbf{1})$ have continuous sample paths and $E(0) = \mathcal{H}_0(\mathbf{1}) = 0$. Also, for every $(X_0, Z_0) \in \mathcal{Y}$, Z_0 has a continuous extension on $[0, \infty)$ by Lemma 2.5.a and $Z_0(0) = X_0 \wedge 0$ by the definition of $\mathcal{Y}$. Therefore, almost surely, $(E, x_0, Z_0 - \mathcal{H}(\mathbf{1}))$ lies in the domain $\mathbb{C}^0[0, \infty) \times \mathbb{R} \times \mathbb{C}_{\mathbb{R}}[0, \infty)$ of the CMS mapping Λ. Hence, (K, X) in (3.1.11) is well defined and (by Lemma 2.13) satisfies almost surely, for $t \geq 0$,

$$X(t) \wedge 0 = Z_0(t) - \mathcal{H}_t(\mathbf{1}) + K(t) - \int_0^t g(t-s) K(s) ds, \tag{3.1.13}$$

$$K(t) = B(t) - \beta t - X^+(t) + X_0^+. \tag{3.1.14}$$

Also, almost surely, for every $t \geq 0$, $Z_0 \in \mathbb{H}^1(0, \infty)$ implies $Z_0(t + \cdot)$ lies in $\mathbb{H}^1(0, \infty)$ (see Lemma 2.5.c) the continuity of K and Lemma 3.16.a imply $\Gamma_t K(\cdot) \in \mathbb{H}^1(0, \infty)$, and Proposition 3.9 implies $r \mapsto \mathcal{M}_t(\Psi_{t+r}\mathbf{1})$ lies in $\mathbb{H}^1(0, \infty)$. Thus, $Z(t, \cdot)$ in (3.1.12) also lies in $\mathbb{H}^1(0, \infty)$ and furthermore, by Lemma 2.5.a, has an extension in $\mathbb{C}_{\mathbb{R}}[0, \infty)$, so $Z(t, r)$ in (3.1.12) is well defined for $r \in [0, \infty)$.

REMARK 3.6. For each fixed $t \in [0, \infty)$, the function $r \mapsto Z(t, r)$ can be identified with the quantity $r \mapsto \widehat{\mathcal{J}}_r^{\nu_t}(\mathbf{1})$ defined in [**62**, Section 5]. In [**62**], the process $r \mapsto \widehat{\mathcal{J}}_r^{\nu_t}(\mathbf{1})$ was viewed as a process taking values in $\mathbb{C}[0, \infty)$ and appended to the non-Markovian limit $(\widehat{X}, \widehat{\nu})$ of the sequence of Markovian state processes $(\widehat{X}^N, \widehat{\nu}^N)$, $N \in$

$\mathbb{N}$, considered in **[62]** to yield a Markov process $(\widehat{X}_t, \widehat{\nu}_t, \widehat{\mathcal{J}}^{\nu_t}(\mathbf{1}))$ on the space $\mathbb{R} \times \mathbb{H}_{-2} \times \mathbb{C}[0,\infty)$, where $\mathbb{H}_{-2}$ is a certain distribution space.

In contrast, the process $Z(t,\cdot), t \geq 0$, defined in (3.1.12) that we consider takes values in $\mathbb{H}^1(0,\infty)$ (in fact, takes values in $\mathbb{H}^1(0,\infty) \cap \mathbb{C}^1[0,\infty)$ when Z_0 also lies in that space) and then the process $(X(t), Z(t,\cdot))$ on its own turns out to be a Markov process (see Section 3.3.3). It should be emphasized that however, $(X(t), Z(t,\cdot))$ is not a Markov process on $\mathbb{R} \times \mathbb{C}[0,\infty)$, that is when Z is considered to take values in the space of continuous functions. This turns out to be an extremely useful observation. Indeed, the right choice of space eliminates the need to deal with the distribution-valued process $\{\widehat{\nu}_t\}$ of **[62]**, and provides a more tractable diffusion limit, that facilitates the study of its invariant distribution.

3.2. Preliminaries

This section contains some technical results which are required later in the proof of Theorem 3.4. First in Section 3.2.1 we establish regularity properties of various stochastic processes that arise in the analysis of the SPDE. Then in Section 3.2.2, we study properties of the family of auxiliary mappings $\{\Gamma_t; t \geq 0\}$ defined in (3.1.10).

3.2.1. Properties of the Martingale Measure.

Recall the definition given in Section 3.1.1 of the continuous martingale measure $\mathcal{M} = \{\mathcal{M}_t(A),\, A \in \mathcal{B}[0,\infty), t \in [0,\infty)\}$ and stochastic integral $\mathcal{M}_t(\varphi)$ for every $t \geq 0$ and bounded measurable function φ on $[0,\infty) \times [0,\infty)$. Also, recall the definitions (2.2.8) and (2.3.7) of the families of mapping $\{\Phi_t; t \geq 0\}$ and $\{\Psi_t; t \geq 0\}$, respectively.

3.2.1.1. *Continuity of martingale measures.* Recall that a stochastic process or random field $\{\xi_1(t); t \in \mathcal{T}\}$ with index set $\mathcal{T}$ is called a modification of another stochastic process or random field $\{\xi_2(t); t \in \mathcal{T}\}$ defined on a common probability space, if for every $t \in \mathcal{T}$, $\xi_1(t) = \xi_2(t)$, almost surely. In contrast, ξ_1 and ξ_2 are said to be indistinguishable if $\mathbb{P}\{\xi_1(t) = \xi_2(t) \text{ for all } t \in \mathcal{T}\} = 1$. The next two results, Lemma 3.7 and Proposition 3.9, show that we can choose suitably regular modifications of certain stochastic integrals.

LEMMA 3.7. Suppose Assumptions II and III hold. Then, the random fields $\{\mathcal{M}_t(\Phi_r\mathbf{1}); t, r \geq 0\}$, $\{\mathcal{M}_t(\Psi_{t+r}\mathbf{1}); t, r \geq 0\}$ and $\{\mathcal{M}_t(\Psi_{t+r}h); t, r \geq 0\}$ have modifications that are jointly continuous on $[0,\infty) \times [0,\infty)$.

PROOF. Fix $T \in (0,\infty)$, $0 \leq s \leq t \leq T$ and $0 \leq r, u \leq T$. Standard martingale estimates (e.g., apply the bound (8.5) in [**62**, Lemma 8.2] with $r = 6$ and $\bar{\nu}_s(dx) = \bar{G}(x)dx$ for every s, and note that $\bar{A}_{\phi^2}$ in [**62**] is equal to the right-hand side of (4.16) in [**62**] and $\widehat{M}$ in [**62**] coincides with the martingale measure $\mathcal{M}$ given in this paper) show that there exists a constant $\hat{c}_1 < \infty$ such that for every bounded measurable function φ on $[0,\infty) \times [0,\infty)$,

$$\mathbb{E}\left[|\mathcal{M}_t(\varphi) - \mathcal{M}_s(\varphi)|^6\right] \leq \hat{c}_1 \left(\int_0^t \int_0^\infty \varphi^2(x,v) g(x) dx dv\right)^3. \tag{3.2.1}$$

Let $\hat{c}_2 < \infty$ be such that $(a+b)^6 \leq \hat{c}_2(a^6 + b^6)$. By (2.2.8) and the bound (3.2.1) above, we have

$$\begin{aligned}
\mathbb{E}\left[|\mathcal{M}_t(\Phi_r\mathbf{1}) - \mathcal{M}_s(\Phi_u\mathbf{1})|^6\right] &\leq c\,\mathbb{E}\left[|\mathcal{M}_t(\Phi_r\mathbf{1}) - \mathcal{M}_s(\Phi_r\mathbf{1})|^6\right] + c\,\mathbb{E}[|\mathcal{M}_s(\Phi_r\mathbf{1} - \Phi_u\mathbf{1})|^6] \\
&\leq c\left(\int_s^t \int_0^\infty \frac{\overline{G}(x+r)^2}{\overline{G}(x)^2}\, g(x) dx dv\right)^3 \\
&\quad + c\left(\int_0^s \int_0^\infty \frac{(\overline{G}(x+r) - \overline{G}(x+u))^2}{\overline{G}(x)^2}\, g(x) dx dv\right)^3,
\end{aligned}$$

where $c \doteq \hat{c}_1\hat{c}_2 < \infty$. Then, by Assumption III and the mean value theorem, there exists r^* between r and u such that

$$\begin{aligned}
\mathbb{E}\left[|\mathcal{M}_t(\Phi_r\mathbf{1}) - \mathcal{M}_s(\Phi_u\mathbf{1})|^6\right] &\leq c|t-s|^3 + cT^3\left(\int_0^\infty \frac{g(x+r^*)^2}{\overline{G}(x)^2}\, g(x) dx\right)^3 |r-u|^6 \\
&\leq c|t-s|^3 + cT^3 H^6(2T)^3 |r-u|^3.
\end{aligned}$$

Therefore, by the Kolmogorov-Centsov theorem (see e.g. [**89**, Theorem I.25.2], with $n = 2$, $\alpha = 6$ and $\epsilon = 1$), the random field $\{\mathcal{M}_t(\Phi_r\mathbf{1}); t, r \geq 0\}$ has a continuous modification.

In a similar fashion, by definition (2.3.7) and another application of the bound (3.2.1), there exists r^* between $t+r$ and $s+u$ such that

$$
\begin{aligned}
&\mathbb{E}\left[|\mathcal{M}_t(\Psi_{t+r}\mathbf{1}) - \mathcal{M}_s(\Psi_{s+u}\mathbf{1})|^6\right] \\
&\quad \leq \hat{c}_2\, \mathbb{E}\left[|\mathcal{M}_t(\Psi_{t+r}\mathbf{1}) - \mathcal{M}_s(\Psi_{t+r}\mathbf{1})|^6\right] + \hat{c}_2\, \mathbb{E}\left[|\mathcal{M}_s(\Psi_{t+r}\mathbf{1} - \Psi_{s+u}\mathbf{1})|^6\right] \\
&\quad \leq\ c\left(\int_s^t \int_0^\infty \frac{\overline{G}(t+x+r-v)^2}{\overline{G}(x)^2}\, g(x)dxdv\right)^3 \\
&\qquad + c\left(\int_0^s \int_0^\infty \frac{(\overline{G}(t+x+r-v) - \overline{G}(s+x+u-v))^2}{\overline{G}(x)^2}\, g(x)dxdv\right)^3 \\
&\quad \leq\ c|t-s|^3 + c\left(\int_0^s \int_0^\infty \frac{g(x+r^*-v)^2}{\overline{G}(x)^2}\, g(x)dxdv\right)^3 (|t-s| + |r-u|)^6 \\
&\quad \leq\ c|t-s|^3 + \hat{c}_2 c T^3 H^6 (2T)^3 (|t-s|^3 + |r-u|^3) \\
&\quad \leq\ \tilde{c}(|t-s|^3 + |r-u|^3),
\end{aligned}
$$

for some $\tilde{c} < \infty$. Again, by the Kolmogorov-Centsov theorem, $\{\mathcal{M}_t(\Psi_{t+r}\mathbf{1}); t, r \geq 0\}$ has a continuous modification. Similarly, using Assumption III and (2.3.7), we obtain

$$
\begin{aligned}
\mathbb{E}\left[|\mathcal{M}_t(\Psi_{t+r}h) - \mathcal{M}_s(\Psi_{s+u}h)|^6\right] &\leq\ c|t-s|^3 + c3T^3 H_2^6 (2T)^3 |t+r-s-u|^3 \\
&\leq \tilde{c}(|t-s|^3 + |r-u|^3),
\end{aligned}
$$

for some $c, \tilde{c} < \infty$. Thus, $\{\mathcal{M}_t(\Psi_{t+r}h); t, r \geq 0\}$ has a continuous modification. □

REMARK 3.8. By $\mathcal{M}_t(\Phi_r\mathbf{1})$, $\mathcal{M}_t(\Psi_{t+r}\mathbf{1})$ and $\mathcal{M}_t(\Psi_{t+r}h)$, we always denote the jointly continuous modification. Note that, by substituting $r = 0$ in Lemma 3.7, this also implies the continuity of the stochastic processes $t \mapsto \mathcal{M}_t(\mathbf{1})$, $t \mapsto \mathcal{H}_t(\mathbf{1})$ and $t \mapsto \mathcal{H}_t(h)$.

3.2.1.2. *More Sample Path Properties.* Next, we prove another sample path property for stochastic integrals with respect to martingale measures.

PROPOSITION 3.9. Suppose Assumptions II and III hold. Then, the process $\{\mathcal{M}_t(\Psi_{t+\cdot}\mathbf{1});\ t \geq 0\}$ has a continuous $\mathbb{H}^1(0,\infty)$-valued modification. Also, almost surely, for every $t \geq 0$, the function $r \mapsto \mathcal{M}_t(\Psi_{t+r}\mathbf{1})$ has density $-\mathcal{M}_t(\Psi_{t+r}h)$ on $(0,\infty)$.

REMARK 3.10. By the continuous embedding result of Lemma 2.5.b, given a real-valued function f on $[0,\infty) \times (0,\infty)$ such that $t \mapsto f(t,\cdot)$ is continuous from $[0,\infty)$ to $\mathbb{H}^1(0,\infty)$, the mapping $(t,r) \mapsto f(t,r)$ can be extended as a jointly continuously function on $[0,\infty) \times [0,\infty)$. Therefore, the continuous modification of $\{\mathcal{M}_t(\Psi_{t+\cdot}(\mathbf{1}); t \geq 0\}$ in Proposition 3.9 is indistinguishable from the jointly continuous modification of $\{\mathcal{M}_t(\Psi_{t+r}(\mathbf{1}); t, r \geq 0\}$ in Remark 3.8.

We break the argument into several steps. We start by establishing properties of the random element $\mathcal{M}_t(\Psi_{t+\cdot}\mathbf{1})$ for fixed $t \geq 0$.

LEMMA 3.11. If Assumptions II and III holds, then the following two properties are satisfied:

(a) For every $t \geq 0$, almost surely,

$$\mathcal{M}_t(\Psi_{t+r}\mathbf{1}) = \mathcal{H}_t(\mathbf{1}) - \int_0^r \mathcal{M}_t(\Psi_{t+u}h)du, \qquad \forall r \geq 0; \tag{3.2.2}$$

(b) For every $t \geq 0$, the random functions $r \mapsto \mathcal{M}_t(\Psi_{t+r}\mathbf{1})$ and $r \mapsto \mathcal{M}_t(\Psi_{t+r}h)$ lie in $\mathbb{L}^2(0,\infty)$, almost surely.

PROOF. We start by proving property a, which is similar in spirit to, but not a direct consequence of, Lemma E.1 of **[62]**. For fixed $t > 0$, by Assumption III, the function

$$(\Psi_{t+u}h)(x,s) = \frac{g(x+t+u-s)}{\overline{G}(x)},$$

is measurable in (u,x,s), bounded and satisfies

$$\int_0^r (\Psi_{t+u}h)(x,s)du = \frac{\overline{G}(x+t-s)}{\overline{G}(x)} - \frac{\overline{G}(x+t+r-s)}{\overline{G}(x)} = \Psi_t\mathbf{1}(x,s) - \Psi_{t+r}\mathbf{1}(x,s).$$

Since $\mathcal{M}$ is an orthogonal, and therefore a worthy, martingale measure, by the stochastic Fubini theorem for martingale measures (see e.g. [**98**, Theorem 2.6]) we have

$$\int_0^r \mathcal{M}_t(\Psi_{t+u}h)du = \mathcal{M}_t\left(\int_0^r \Psi_{t+u}h\, du\right) = \mathcal{M}_t(\Psi_t\mathbf{1}) - \mathcal{M}_t(\Psi_{t+r}\mathbf{1}),$$

Since $\mathcal{H}_t(\mathbf{1}) = \mathcal{M}_t(\Psi_t\mathbf{1})$, equation (3.2.2) follows from the last display.

Now we turn to the proof of property b. For every $t, r \geq 0$, (3.1.3), (2.3.7), Assumption III and definition of the constant H in Remark 2.2 imply that

$$\begin{aligned}
\mathbb{E}[\mathcal{M}_t(\Psi_{t+r}\mathbf{1})^2] &= \int_0^t \int_0^\infty \frac{\overline{G}(x+t+r-s)^2}{\overline{G}(x)^2} g(x)dxds \\
&\leq t\int_0^\infty \frac{\overline{G}(x+r)^2}{\overline{G}(x)^2} g(x)dx \\
&\leq tH \int_r^\infty \overline{G}(x)dx.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\mathbb{E}[\mathcal{M}(\Psi_{t+r}h)^2] &= \int_0^t \int_0^\infty \frac{g(x+t+r-s)^2}{\overline{G}(x)^2} g(x)dxds \\
&\leq H^2 \int_0^t \int_0^\infty \frac{\overline{G}(x+t+r+s)^2}{\overline{G}(x)^2} g(x)dxds \\
&\leq tH^3 \int_r^\infty \overline{G}(x)dx.
\end{aligned}$$

Therefore, Fubini's theorem and the last two inequalities together show that

$$\max\left(\mathbb{E}\left[\int_0^\infty \mathcal{M}_t(\Psi_{t+r}\mathbf{1})^2 dr\right], \mathbb{E}\left[\int_0^\infty \mathcal{M}_t(\Psi_{t+r}h)^2 dr\right]\right) \leq t(H^3 \vee H)\int_0^\infty \int_r^\infty \overline{G}(x)dx\, dr,$$

which is finite by Assumption II.b (see Remark 2.3). Therefore, the $\mathbb{L}^2$ norms of both $\mathcal{M}_t(\Psi_{t+\cdot}\mathbf{1})$ and $\mathcal{M}_t(\Psi_{t+\cdot}h)$ are finite in expectation, and hence finite almost surely. $\square$

COROLLARY 3.12. Suppose Assumptions II and III hold. Then for every $t \geq 0$, almost surely, the function $\mathcal{M}_t(\Psi_{t+\cdot}\mathbf{1})$ lies in $\mathbb{H}^1(0,\infty)$ and has density $-\mathcal{M}_t(\Psi_{t+\cdot}h)$.

PROOF. Fix $t \geq 0$. It follows from part a. of Lemma 3.11 that almost surely, the function $r \mapsto \mathcal{M}_t(\Psi_{t+r}\mathbf{1})$ is absolutely continuous with density $-\mathcal{M}_t(\Psi_{t+\cdot}h)$.

Moreover, part b. shows that both the function $\mathcal{M}_t(\Psi_{t+\cdot}\mathbf{1})$ and its density, $-\mathcal{M}_t(\Psi_{t+\cdot}h)$ lie in $\mathbb{L}^2(0,\infty)$, almost surely. $\square$

Corollary 3.12 shows that $\{\mathcal{M}_t(\Psi_{t+\cdot}\mathbf{1}); t \geq 0\}$ is an $\mathbb{H}^1(0,\infty)$-valued stochastic process. Next, we show that this process has a continuous modification.

LEMMA 3.13. Suppose Assumptions II and III hold. Then the process $\{\mathcal{M}_t(\Psi_{t+\cdot}\mathbf{1});\, t \geq 0\}$ has a continuous $\mathbb{H}^1(0,\infty)$-valued modification.

PROOF. Choose $T \geq 0$ and $0 \leq s \leq t \leq T$, and define

$$\zeta_{s,t}(r) \doteq \mathcal{M}_t(\Psi_{t+r}\mathbf{1}) - \mathcal{M}_s(\Psi_{s+r}\mathbf{1}), \quad r \in (0,\infty). \tag{3.2.3}$$

Then for every $r \geq 0$,

$$\begin{aligned}
\zeta_{s,t}(r) &= \iint\limits_{[0,\infty)\times[0,t]} \Psi_{t+r}\mathbf{1}(x,v)\mathcal{M}(dx,dv) - \iint\limits_{[0,\infty)\times[0,s]} \Psi_{s+r}\mathbf{1}(x,v)\mathcal{M}(dx,dv) \\
&= \iint\limits_{[0,\infty)\times[0,s]} \left(\Psi_{t+r}\mathbf{1}(x,v) - \Psi_{s+r}\mathbf{1}(x,v)\right)\mathcal{M}(dx,dv) \\
&\quad + \iint\limits_{[0,\infty)\times(s,t]} \Psi_{t+r}\mathbf{1}(x,v)\mathcal{M}(dx,dv).
\end{aligned}$$

Since $\Psi_{t+\cdot}\mathbf{1}$ and $\Psi_{s+\cdot}\mathbf{1}$ are deterministic functions and $\mathcal{M}$ has independent increments, this shows that $\{\zeta_{s,t}(r); r \geq 0\}$ is the sum of two independent Gaussian processes. Therefore, $\{\zeta_{s,t}(r); r \geq 0\}$ is a Gaussian process with covariance function $\sigma(r,u) = \sigma_1(r,u) + \sigma_2(r,u)$, where for $r, u \geq 0$,

$$\begin{aligned}
\sigma_1(r,u) &\doteq \mathbb{E}\left[\mathcal{M}_s(\Psi_{t+r}\mathbf{1} - \Psi_{s+r}\mathbf{1})\mathcal{M}_s(\Psi_{t+u}\mathbf{1} - \Psi_{s+u}\mathbf{1})\right], \\
\sigma_2(r,u) &\doteq \mathbb{E}\left[\iint\limits_{[0,\infty)\times(s,t]} \Psi_{t+r}\mathbf{1}(x,v)\mathcal{M}(dx,dv) \iint\limits_{[0,\infty)\times(s,t]} \Psi_{t+u}\mathbf{1}(x,v)\mathcal{M}(dx,dv)\right].
\end{aligned}$$

Using the fact that for $0 \leq s, t < \infty$ and $a \geq 0$,

$$\left(\Psi_{t+a}\mathbf{1} - \Psi_{s+a}\mathbf{1}\right)(x,v) = \frac{\overline{G}(t+a+x-v) - \overline{G}(s+a+x-v)}{\overline{G}(x)},$$

by the mean value theorem for $\overline{G}$, (3.1.3), Assumption III, and the monotonicity of $\overline{G}$, we have for some $t_1, t_2 \in (s,t)$,

$$\begin{aligned}\sigma_1(r,u) &= |t-s|^2 \int_0^s \int_0^\infty \frac{g(t_1+r+x-v)g(t_2+u+x-v)}{\overline{G}(x)^2} g(x)dx\, dv \\ &\leq TH^3|t-s|^2 \int_0^\infty \frac{\overline{G}(x+r)\overline{G}(x+u)}{\overline{G}(x)}\, dx,\end{aligned}$$

and an analogous calculation shows that

$$\begin{aligned}\sigma_2(r,u) &= \int_s^t \int_0^\infty \frac{\overline{G}(t+r+x-v)\overline{G}(t+u+x-v)}{\overline{G}(x)^2} g(x)dx\, dv \\ &\leq H|t-s| \int_0^\infty \frac{\overline{G}(x+r)\overline{G}(x+u)}{\overline{G}(x)}\, dx.\end{aligned}$$

Setting $C_T \doteq (1+T^2H^2)H$, this implies that for every $r,u \geq 0$,

$$\sigma(r,u) \leq C_T|t-s| \int_0^\infty \frac{\overline{G}(x+r)\overline{G}(x+u)}{\overline{G}(x)}\, dx. \tag{3.2.4}$$

Recall that given a pair of jointly Gaussian random variables (ξ_1, ξ_2) with covariance matrix $\Sigma = [\sigma(i,j)]_{i,j=1,2}$, $\mathbb{E}\left[\xi_1^2\xi_2^2\right] = \sigma(1,1)\sigma(2,2) + \sigma(1,2)^2$. Applying this identity with $\xi_1 = \zeta_{s,t}(r)$ and $\xi_2 = \zeta_{s,t}(u)$, and using (3.2.4) and the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\mathbb{E}\left[\zeta_{s,t}(r)^2\zeta_{s,t}(u)^2\right] \leq& C_T^2|t-s|^2 \left(\int_0^\infty \frac{\overline{G}(x+r)^2}{\overline{G}(x)}\, dx\right)\left(\int_0^\infty \frac{\overline{G}(x+u)^2}{\overline{G}(x)}\, dx\right) \\ &+ C_T^2|t-s|^2 \left(\int_0^\infty \frac{\overline{G}(x+r)\overline{G}(x+u)}{\overline{G}(x)}\, dx\right)^2 \\ \leq& 2C_T^2|t-s|^2 \left(\int_r^\infty \overline{G}(x)dx\right)\left(\int_u^\infty \overline{G}(x)dx\right).\end{aligned} \tag{3.2.5}$$

Then, by Tonelli's theorem,

$$\begin{aligned}\mathbb{E}\left[\|\zeta_{s,t}(\cdot)\|_{\mathbb{L}^2}^4\right] =& \mathbb{E}\left[\left(\int_0^\infty \zeta_{s,t}^2(r)dr\right)^2\right] \\ =& \mathbb{E}\left[\int_0^\infty \int_0^\infty \zeta_{s,t}^2(r)\zeta_{s,t}^2(u)dr\, du\right] \\ =& \int_0^\infty \int_0^\infty \mathbb{E}\left[\zeta_{s,t}^2(r)\zeta_{s,t}^2(u)\right] dr\, du.\end{aligned}$$

The bound (3.2.5) then implies that

$$\mathbb{E}\left[\|\zeta_{s,t}(\cdot)\|_{\mathbb{L}^2}^4\right] =\leq 2C_T^2|t-s|^2\left(\int_0^\infty\int_r^\infty \overline{G}(x)dx\,dr\right)^2 \leq \tilde{C}_T|t-s|^2$$

for $\tilde{C}_T \doteq 2C_T^2\left(\int_0^\infty\int_r^\infty \overline{G}(x)dx\,dr\right)^2$, which is finite by Assumption II.b (see Remark 2.3). Substituting the definition (3.2.3) of $\zeta_{s,t}$ into the last inequality, it follows that

$$\mathbb{E}\left[\|\mathcal{M}_t(\Psi_{t+\cdot}\mathbf{1}) - \mathcal{M}_s(\Psi_{s+\cdot}\mathbf{1})\|_{\mathbb{L}^2}^4\right] \leq \tilde{C}_T|t-s|^2. \tag{3.2.6}$$

Now, note that by (3.2.3) and Corollary 3.12, almost surely, $\zeta_{s,t}(\cdot)$ is absolutely continuous, with density

$$\zeta'_{s,t}(r) \doteq -\mathcal{M}_t(\Psi_{t+r}h) + \mathcal{M}_s(\Psi_{s+r}h).$$

Using estimates analogous to those used above, one can show that

$$\mathbb{E}\left[\|\mathcal{M}_t(\Psi_{t+\cdot}h) - \mathcal{M}_s(\Psi_{s+\cdot}h)\|_{\mathbb{L}^2}^4\right] \leq C'_T|t-s|^2, \tag{3.2.7}$$

where now

$$C'_T = 2\left(H^2 + 2T^2HH_2^2\right)^2\left(\int_0^\infty\int_r^\infty \overline{G}(x)dx\,dr\right)^2,$$

which is finite by Assumptions II.b and III (see Remark 2.2). Combining (3.2.6) and (3.2.7), for every $s,t \leq T$, we have

$$\mathbb{E}\left[\|\mathcal{M}_t(\Psi_{t+\cdot}h) - \mathcal{M}_s(\Psi_{s+\cdot}h)\|_{\mathbb{H}^1}^4\right] \leq \tilde{C}_T|t-s|^2,$$

for a finite $\tilde{C}_T$ that depends only on T. The existence of an $\mathbb{H}^1(0,\infty)$-continuous modification of $\mathcal{M}_t(\Psi_{t+\cdot}h)$ then follows from a version of Kolmogorov's continuity criterion for stochastic processes taking values in general Polish spaces (see, e.g., Lemma 2.1 in **[91]**). □

PROOF OF PROPOSITION 3.9. Proposition 3.9 follows from Corollary 3.12 and Lemma 3.13. □

3.2.1.3. *A Fubini-Type Lemma.* Next, we establish a simple relation that is used in the next section.

LEMMA 3.14. Suppose Assumptions II and III hold. Then, almost surely

$$\mathcal{M}_t(\Psi_{t+r}\mathbf{1}) = \mathcal{M}_t(\Phi_r\mathbf{1}) - \int_0^t \mathcal{M}_s(\Psi_{s+r}h)ds, \qquad \forall t, r \geq 0. \tag{3.2.8}$$

We first prove a weaker version of this lemma.

LEMMA 3.15. Suppose Assumptions II and III hold. Then, for every $r \geq 0$, almost surely,

$$\mathcal{M}_t(\Psi_{t+r}\mathbf{1}) = \mathcal{M}_t(\Phi_r\mathbf{1}) - \int_0^t \mathcal{M}_s(\Psi_{s+r}h)ds, \qquad \forall t \geq 0. \tag{3.2.9}$$

PROOF. As shown below, this will follow as a corollary of Lemma E.1 in **[62]**, which is applicable here since Assumption 2 and Assumption 4 in **[62]** and the continuity of g can be deduced from Assumption III here. (Note that although the statement of Lemma E.1 in **[62]** imposes Assumptions 1-4 of **[62]**, Assumption 1 and Assumption 3 of **[62]** are not used in the proof of that lemma.)

Following definition (8.27) in **[62]**, for $t \geq 0$, define the operator $\Xi_t : \mathbb{C}_b([0,\infty) \times [0,t]) \mapsto \mathbb{C}_b([0,\infty) \times [0,t])$ by

$$(\Xi_t\varphi)(x,u) \doteq \int_u^t (\Psi_s\varphi(\cdot,s))(x,u)ds = \int_u^t \varphi(x+s-u,s)\frac{\overline{G}(x+s-u)}{\overline{G}(x)}ds. \tag{3.2.10}$$

Then (E.1) of Lemma E.1 of **[62]** shows that given any $\tilde{\varphi} \in \mathbb{C}_b([0,\infty) \times [0,t])$, almost surely,

$$\mathcal{M}_t(\Xi_t\tilde{\varphi}) = \int_0^t \mathcal{H}_r(\tilde{\varphi}(\cdot,r))dr, \quad t \geq 0. \tag{3.2.11}$$

Now, fix $r \geq 0$, and define

$$\varphi_r(x,u) \doteq \Phi_r h(x) = \frac{g(x+r)}{\overline{G}(x)}, \quad (x,u) \in [0,\infty) \times [0,t].$$

Note that for every $t \geq 0$, $\varphi_r \in \mathbb{C}_b([0,\infty) \times [0,t])$ due to Assumption III, and

$$\begin{aligned}
(\Xi_t \varphi_r)(x,u) &= \int_u^t \frac{g(x+s+r-u)}{\overline{G}(x)}\, ds \\
&= \frac{\overline{G}(x+r)}{\overline{G}(x)} - \frac{\overline{G}(x+t+r-u)}{\overline{G}(x)} \\
&= \Phi_r \mathbf{1}(x) - \Psi_{t+r}\mathbf{1}(x,u).
\end{aligned}$$

Applying (3.2.11) with $\tilde{\varphi} = \varphi_r = \Phi_r h$, and substituting the last identity, we see that almost surely,

$$\mathcal{M}_t(\Phi_r \mathbf{1}) - \mathcal{M}_t(\Psi_{t+r}\mathbf{1}) = \int_0^t \mathcal{H}_s\left(\Phi_r h\right) ds, \qquad \forall t \geq 0.$$

Since equation (3.1.9) for $\mathcal{H}_t$ and property (2.3.8) of the operators $\{\Psi_t; t \geq 0\}$ imply that

$$\mathcal{H}_s\left(\Phi_r h\right) = \mathcal{M}_s(\Psi_s \Phi_r h) = \mathcal{M}_s(\Psi_{s+r} h) \qquad \forall s, r \geq 0,$$

this concludes the proof of the lemma. □

Now we can prove Lemma 3.14.

PROOF OF LEMMA 3.14. It follows from Lemma 3.7 and Remark 3.8 that both sides of (3.2.9) are jointly continuous in (t,r). Therefore, Lemma 3.15 can be strengthened to obtain Lemma 3.14, namely, almost surely, for every $r, t \geq 0$, the equality in (3.2.9) is satisfied. □

Setting $r = 0$ in (3.2.8), we see that, almost surely,

$$\mathcal{H}_t(\mathbf{1}) = \mathcal{M}_t(\mathbf{1}) - \int_0^t \mathcal{H}_s(h) ds, \quad t \geq 0. \tag{3.2.12}$$

3.2.2. Properties of the Auxiliary Mapping. In this section, we study properties of the mappings $\{\Gamma_t\}$ defined in (3.1.10), summarized in the following lemma.

LEMMA 3.16. Suppose Assumptions II and III hold. Then the following assertions hold:

(a) For every $t \geq 0$ and $\kappa \in \mathbb{C}_{\mathbb{R}}[0,\infty)$, the function $\Gamma_t\kappa$ lies in $\mathbb{H}^1(0,\infty) \cap \mathbb{C}^1[0,\infty)$ and has density

$$(\Gamma_t\kappa)'(r) = -g(r)\kappa(t) - \int_0^t \kappa(s)g'(t+r-s)ds, \qquad r \in [0,\infty). \tag{3.2.13}$$

(b) For every $t \geq 0$, the mapping $\Gamma_t : \mathbb{C}_{\mathbb{R}}[0,\infty) \to \mathbb{H}^1(0,\infty)$ is continuous.

(c) For every $\kappa \in \mathbb{C}_{\mathbb{R}}[0,\infty)$, the mapping $t \mapsto \Gamma_t\kappa$ from $[0,\infty)$ to $\mathbb{H}^1(0,\infty)$ is continuous.

PROOF. We first prove property a. Fix $t \geq 0$. The finite mean assumption on G and Assumption III imply $g' \in \mathbb{L}^1(0,\infty)$. Since $\kappa \in \mathbb{C}_{\mathbb{R}}[0,\infty)$, by Fubini's theorem for $r \geq 0$, we have

$$\begin{aligned}\int_0^r \int_0^t \kappa(u)g'(s+t-u)du\, ds &= \int_0^t \kappa(u) \int_0^r g'(s+t-u)ds\, du \\ &= \int_0^t \kappa(u)g(t+r-u)du - \int_0^t \kappa(u)g(t-u)du. \end{aligned} \tag{3.2.14}$$

Substituting the equation above into the definition (3.1.10) of $\Gamma_t\kappa$, for $r \geq 0$, we have

$$(\Gamma_t\kappa)(r) = \overline{G}(r)\kappa(t) - \int_0^t \kappa(u)g(t-u)du - \int_0^r \int_0^t \kappa(u)g'(s+t-u)du\, ds.$$

Hence, $r \mapsto \Gamma_t\kappa(r)$ is absolutely continuous with density $(\Gamma_t\kappa)'$ given by (3.2.13). Also, for every $t \geq 0$, by Assumption III, the family of functions $u \mapsto \kappa(u)g'(r+t-u)$ indexed by $r \geq 0$ are uniformly bounded by the integrable function $u \mapsto H_2\|\kappa\|_t\ \overline{G}(u)$. Hence, by the dominated convergence theorem, and the continuity of g' from Assumption III, for every $r_0 \geq 0$,

$$\lim_{r\to r_0} \int_0^t \kappa(u)g'(r+t-u)du = \int_0^t \kappa(u)g'(r_0+t-u)du.$$

The above display, together with the continuity of g from Assumption II, implies that the density $(\Gamma_t\kappa)'$ in (3.2.13) is continuous on $[0,\infty)$. Therefore, $\Gamma_t\kappa \in \mathbb{C}^1[0,\infty)$.

Furthermore, from (3.1.10), we have

$$|(\Gamma_t\kappa)(r)| \leq |\kappa(t)|\overline{G}(r) + \|\kappa\|_t \int_r^{t+r} g(s)ds \leq 2\|\kappa\|_t\overline{G}(r). \tag{3.2.15}$$

Since the right-hand side of (3.2.15) is a uniformly bounded and integrable function of $r \in (0, \infty)$, it follows that for each $t \geq 0$, $\Gamma_t \kappa \in \mathbb{L}^2(0, \infty)$. Furthermore, using Assumption III and (3.2.13), we have

(3.2.16)
$$|(\Gamma_t \kappa)'(r)| \leq |\kappa(t)| g(r) + \|\kappa\|_t \int_r^{t+r} |g'(s)| ds \leq \|\kappa\|_t \left(H\overline{G}(r) + H_2 \int_r^\infty \overline{G}(u) du \right).$$

Again, $\overline{G}$ and $\int_\cdot^\infty \overline{G}(u)du$ are bounded and integrable by Assumption II.b, and therefore, $(\Gamma_t \kappa)'$ also lies in $\mathbb{L}^2(0, \infty)$. Thus, $\Gamma_t \kappa \in \mathbb{H}^1(0, \infty)$. This completes the proof of part a.

For part b, let κ and $\tilde{\kappa}$ be functions in $\mathbb{C}_{\mathbb{R}}[0, \infty)$. By linearity of the mapping Γ_t and the bounds (3.2.15) and (3.2.16), we have

$$\|\Gamma_t \kappa - \Gamma_t \tilde{\kappa}\|_{\mathbb{L}^2} \leq 2\|\overline{G}\|_{\mathbb{L}^2} \|\kappa - \tilde{\kappa}\|_t, \tag{3.2.17}$$

and

$$\|(\Gamma_t \kappa)' - (\Gamma_t \tilde{\kappa})'\|_{\mathbb{L}^2} \leq C_1 \|\kappa - \tilde{\kappa}\|_t, \tag{3.2.18}$$

with $C_1 \doteq H\|\overline{G}\|_{\mathbb{L}^2} + H_2 \| \int_\cdot^\infty \overline{G}(u)du\|_{\mathbb{L}^2}$, which is finite by Assumptions II.b and III. The assertion in part b. then follows from (3.2.17) and (3.2.18).

For part c, fix $T \geq 0$. For every $0 \leq s < t \leq T$, by definition (3.1.10) of $\{\Gamma_t; t \geq 0\}$, Minkowski's integral inequality, Assumption III and Fubini's theorem, we have

$$\begin{aligned}
\|\Gamma_t \kappa - \Gamma_s \kappa\|_{\mathbb{L}^2} &\leq \|\overline{G}\|_{\mathbb{L}^2} |\kappa(t) - \kappa(s)| + \left\| \int_s^t \kappa(u) g(\cdot + t - u) du \right\|_{\mathbb{L}^2} \\
&\quad + \left\| \int_0^s \kappa(u) |g(\cdot + t - u) - g(\cdot + s - u)| du \right\|_{\mathbb{L}^2} \\
&\leq \|\overline{G}\|_{\mathbb{L}^2} |\kappa(t) - \kappa(s)| + \|\kappa\|_T H \left\| \int_0^{t-s} \overline{G}(\cdot + u) du \right\|_{\mathbb{L}^2} \\
&\qquad + \|\kappa\|_T \left\| \int_0^s |g(\cdot + u) - g(\cdot + t - s + u)| du \right\|_{\mathbb{L}^2} \\
&\leq \|\overline{G}\|_{\mathbb{L}^2} |\kappa(t) - \kappa(s)| + \|\kappa\|_T H \|\overline{G}\|_{\mathbb{L}^2} |t - s| \qquad (3.2.19) \\
&\qquad + \|\kappa\|_T \int_0^T \|g(t - s + u + \cdot) - g(u + \cdot)\|_{\mathbb{L}^2} du.
\end{aligned}$$

The first two terms in (3.2.19) converge to zero as $|t-s| \to 0$ by the continuity of κ. Also, since Assumption III implies that g is square integrable, for every $u \in [0,T]$, the term $\|g(t-s+u+\cdot) - g(u+\cdot)\|_{\mathbb{L}^2}$ is bounded by $2\|g\|_{\mathbb{L}^2}$ and hence, the third term converges to zero as $t \to s$ by an application of the bounded convergence theorem and continuity of the translation map in the $\mathbb{L}^2$ norm.

Similarly, for every $0 \leq s < t \leq T$, by definition (3.2.13) of $(\Gamma\kappa)'$ and Assumption III (see Remark 2.2),

$$\begin{aligned}(3.2.20)\quad \|(\Gamma_t\kappa)' - (\Gamma_s\kappa)'\|_{\mathbb{L}^2} \leq \|g\|_{\mathbb{L}^2}|\kappa(t)-\kappa(s)| + \|\kappa\|_T H_2 \|\overline{G}\|_{\mathbb{L}^2}|t-s| \\ + \|\kappa\|_T \int_0^T \|g'(\cdot+u) - g'(\cdot+t-s+u)\|_{\mathbb{L}^2} du.\end{aligned}$$

Again, the first two terms on the right-hand side above converge to zero as $|t-s| \to 0$ by the continuity of κ, and the third term converges to zero as $|t-s| \to 0$ by continuity of the translation map in the $\mathbb{L}^2$ norm, boundedness of $\|g'\|_{\mathbb{L}^2}$ (see Assumption III and Remark 2.3) and the bounded convergence theorem. The $\mathbb{H}^1(0,\infty)$-continuity of $t \mapsto \Gamma_t\kappa$ follows from (3.2.19) and (3.2.20). $\square$

3.3. Properties of the Diffusion Limit Process

This section is devoted to the proof of Theorem 3.4. First in Section 3.3.1, we show that the diffusion limit process introduced in definition 3.3 is indeed a solution to the coupled SPDEs (3.1.5)-(3.1.7) (Proposition 3.20). The uniqueness of the solution to the coupled SPDE (Proposition 3.22) is then proved in Section 3.3.2. Finally, the Markov property of the diffusion limit process Y (Proposition 3.25) is established in Section 3.3.3.

3.3.1. Existence. In this section, we show that the diffusion limit defined in Section 3.1.2 is indeed a solution of the coupled stochastic age equations. Throughout this section, we fix $Y_0 = (X_0, Z_0) \in \mathcal{Y}$ and let $Y = \{(X(t), Z(t,\cdot)); t \geq 0\}$ be the diffusion limit with initial condition Y_0, as specified in Definition 3.3. We first prove that Y is a continuous $\mathcal{Y}$-valued process.

PROPOSITION 3.17. Suppose Assumptions II and III hold. Then the diffusion limit process $Y = (X, Z)$ satisfies the following properties:

(a) Almost surely, the sample paths of $\{Z(t,\cdot); t \geq 0\}$ are continuous and $\mathbb{H}^1(0,\infty)$-valued, and for every $t \geq 0$, the density $\partial_r Z(t,\cdot)$ of $Z(t,\cdot)$ satisfies for a.e. $r \in (0,\infty)$,

$$\partial_r Z(t,r) = Z_0'(t+r) + \mathcal{M}_t(\Psi_{t+r}h) - g(r)K(t) - \int_0^t K(s)g'(t+r-s)ds. \tag{3.3.1}$$

(b) Almost surely, $Z(t,0) = X(t) \wedge 0$, for all $t \geq 0$.

(c) $\{Y(t);\ t \geq 0\}$ is an almost surely continuous $\mathcal{Y}$-valued process.

PROOF. For part a., we look at each term in the definition (3.1.12) of Z separately. Since $Z_0 \in \mathbb{H}^1(0,\infty)$, the translation mapping $t \mapsto Z_0(t+\cdot)$ is continuous in $\mathbb{H}^1(0,\infty)$ by Lemma 2.5.c and for each $t > 0$, $Z_0(t+\cdot)$ has density $Z_0'(t+\cdot)$. For the second term, by Proposition 3.9, $\{\mathcal{M}_t(\Psi_{t+\cdot}\mathbf{1}); t \geq 0\}$ is a continuous $\mathbb{H}^1(0,\infty)$-valued process and almost surely, for every $t \geq 0$, the function $r \mapsto \mathcal{M}_t(\Psi_{t+r}\mathbf{1})$ has density $r \mapsto -\mathcal{M}_t(\Psi_{t+r}h)$. Finally for the third term, since the range of the CMS map Λ lies in $\mathbb{C}^0[0,\infty) \times \mathbb{C}_{\mathbb{R}}[0,\infty)$ (see Definition 2.12 and Lemma 2.13), almost surely, the sample paths of the process K defined in (3.1.11) are continuous. Therefore, by Lemma 3.16.c, $\{(\Gamma_t K); t \geq 0\}$ is a continuous $\mathbb{H}^1(0,\infty)$-valued process. Also, by (3.2.13), for every $t \geq 0$, $r \mapsto \Gamma_t K(r)$ has density $-g(r)K(t) - \int_0^t K(s)g'(t+r-s)ds$. This completes the proof of part a.

Next, for part b., substituting $r = 0$ in (3.1.12) and the definition (3.1.10) of Γ_t, and using the identity $\mathcal{H}_t(\mathbf{1}) = \mathcal{M}_t(\Psi_t\mathbf{1})$ from (3.1.9), we obtain

$$Z(t,0) = Z_0(t) - \mathcal{H}_t(\mathbf{1}) + K(t) - \int_0^t K(s)g(t-s)ds.$$

The assertion in b. then follows from equation (3.1.13).

Finally, note that by definition (3.1.11) and the fact that the range of Λ lies in $\mathbb{C}^0[0,\infty) \times \mathbb{C}_{\mathbb{R}}[0,\infty)$ (see Definition 2.12 and Lemma 2.13), $\{X(t); t \geq 0\}$ is almost surely continuous. Property c. follows from the continuity of X and parts a. and b. of this proposition. $\square$

Next, we show that Z satisfies the regularity condition required in part c of Definition 3.1 of a solution to the coupled stochastic age equations.

LEMMA 3.18. Suppose Assumptions II and III hold and let $Y = (X, Z)$ be the diffusion limit with initial condition $Y_0 = (X_0, Z_0) \in \mathcal{Y}$. Then, almost surely, $(s, r) \mapsto \partial_r Z(s, r)$ is locally integrable on $(0, \infty) \times (0, \infty)$. Moreover, for every $t \geq 0$, the function $r \mapsto \int_0^t \partial_r Z(s, r) ds$ is locally uniformly continuous, and hence can be continuously extended to $[0, \infty)$ and for $t, r \geq 0$, satisfies

(3.3.2)

$$\int_0^t \partial_r Z(s,r)ds = Z_0(t+r) - Z_0(r) + \mathcal{M}_t(\Phi_r \mathbf{1}) - \mathcal{M}_t(\Psi_{t+r}\mathbf{1}) - \int_0^t K(s)g(t+r-s)ds,$$

where $\int_0^t \partial_r Z(s, 0)$ denotes the value of the extension of $r \mapsto \int_0^t \partial_r Z(s, r)$ at $r = 0$.

PROOF. Recall from Proposition 3.17.c that almost surely for every $s \geq 0$, the density $\partial_r Z(s, \cdot)$ of $r \mapsto Z(s, r)$ exists and is given by (3.3.1). We prove the claims separately for each term on the right-hand side of (3.3.1). First, since $Z_0(\cdot) \in \mathbb{H}^1(0, \infty)$, the mapping $(s, r) \mapsto Z_0'(r + s)$ is clearly locally integrable on $(0, \infty) \times (0, \infty)$ and for every $t \geq 0$,

$$\int_0^t Z_0'(s+r)ds = Z_0(t+r) - Z_0(r). \tag{3.3.3}$$

Moreover, since $Z_0(\cdot) \in \mathbb{H}^1(0, \infty)$, as stated in Lemma 2.5.a, $Z_0(t + \cdot)$ also lies in $\mathbb{H}^1(0, \infty)$ for every $t \geq 0$ and $r \mapsto Z_0(t + r) - Z_0(r)$ is uniformly continuous on $(0, \infty)$.

Next, Lemma 3.7 implies that almost surely, $(s, r) \mapsto \mathcal{M}_s(\Psi_{s+r}h)$ is jointly continuous and hence, locally integrable on $(0, \infty) \times (0, \infty)$. Also, for every $t \geq 0$, by (3.2.8) of Lemma 3.14,

$$\int_0^t \mathcal{M}_s(\Psi_{s+r}h)ds = \mathcal{M}_t(\Phi_r \mathbf{1}) - \mathcal{M}_t(\Psi_{t+r}\mathbf{1}), \qquad r \geq 0. \tag{3.3.4}$$

By Lemma 3.7, $r \mapsto \mathcal{M}_t(\Phi_r \mathbf{1}) - \mathcal{M}_t(\Psi_{t+r}\mathbf{1})$ is continuous on $[0, \infty)$, and therefore, is locally uniformly continuous.

Finally, it follows from the continuity of K, g and g' (see Assumption III) that the mapping

$$(s,r) \mapsto g(r)K(s) + \int_0^s K(v)g'(s+r-v)dv$$

is jointly continuous and hence, locally integrable on $(0,\infty) \times (0,\infty)$. Also, for every $t \geq 0$, using Fubini's theorem,

$$\begin{aligned} \int_0^t \Big(g(r)K(s) + \int_0^s K(v)g'(s+r-v)dv \Big) ds &= g(r) \int_0^t K(s)ds \\ &\quad + \int_0^t \int_v^t K(v)g'(s+r-v)ds\,dv \\ &= \int_0^t K(s)g(t+r-s)ds. \end{aligned} \tag{3.3.5}$$

Again, by the continuity of K and g, the mapping $r \mapsto \int_0^t K(s)g(t+r-s)ds$ is continuous on $[0,\infty)$, and hence, locally uniformly continuous.

Equations (3.3.1), (3.3.3), (3.3.4) and (3.3.5), together, show that (3.3.2) holds. The local integrability of $(s,r) \mapsto \partial_r Z(s,r)$ and local uniform continuity of $r \mapsto \int_0^t \partial_r Z(s,r)\,ds$ follow from the corresponding properties of each term on the right-hand sides of (3.3.1) and (3.3.2), respectively. □

The next lemma provides an alternative characterization of the coupled stochastic age equations (3.1.5)-(3.1.7).

LEMMA 3.19. Given an initial condition $Y_0 = (X_0, Z_0(\cdot)) \in \mathcal{Y}$, let $\{Y(t) = (X(t), Z(t,\cdot)); t \geq 0\}$ be a continuous $\mathcal{Y}$-valued process that satisfies conditions a-c of Definition 3.1, and suppose (X,Z) satisfies equations (3.1.6) and (3.1.7). Then, (X,Z) satisfies equation (3.1.5) if and only if for all $t, r \geq 0$,

$$Z(t,r) = Z_0(r) + \int_0^t \partial_r Z(s,r)ds - \mathcal{M}_t(\Phi_r \mathbf{1}) + \overline{G}(r)K(t), \tag{3.3.6}$$

with $K(t) = B(t) - \beta t - X^+(t) + X_0^+$.

PROOF. Comparing (3.3.6) and (3.1.5), and recalling that $\vartheta^r = \Phi_r \mathbf{1}$, it is clear that these two equations are equivalent if and only if for every $t \geq 0$, the following identity holds:

$$K(t) = -Z_0(0) - \int_0^t \partial_r Z(s,0)ds + \mathcal{M}_t(\mathbf{1}) + Z(t,0). \tag{3.3.7}$$

On the other hand, by (3.1.6) and (3.1.7) and the definition of K given above, for $t \geq 0$, we have

$$\begin{aligned}
Z(t,0) &= X(t) - X^+(t) \\
&= X_0 + B(t) - \beta t - \mathcal{M}_t(\mathbf{1}) + \int_0^t \partial_r Z(s,0)ds - X^+(t) \\
&= K(t) - \mathcal{M}_t(\mathbf{1}) + \int_0^t \partial_r Z(s,0)ds + X_0 - X_0^+.
\end{aligned}$$

Equation (3.1.6) (for $t = 0$) also implies that $Z_0(0) = Z(0,0) = X(0) \wedge 0 = X_0 - X_0^+$. When substituted into the last display, (3.3.7) follows. □

We now show that the diffusion limit satisfies the coupled stochastic age equations.

PROPOSITION 3.20. Suppose Assumptions II and III are satisfied and given a $\mathcal{Y}$-valued random element Y_0, let $Y = \{Y(t); t \geq 0\}$ be the diffusion limit specified in Definition 3.3. Then Y is a solution of the coupled stochastic age equations with initial condition Y_0.

PROOF. By Proposition 3.17.c, Y is a continuous $\mathcal{Y}$-valued process. We show that it satisfies conditions a-d of Definition 3.1. Condition b holds because $Y(0) = Y_0$ by definition, and condition c follows from Lemma 3.18.

Next, we verify condition a of Definition 3.1 by showing that Y is $\{\mathcal{F}_t^{Y_0}\}$-adapted. Fix $t \geq 0$, and note that the stopped processes $\{E(s \wedge t); s \geq 0\}$ and $\{\mathcal{H}_{s\wedge t}(\mathbf{1}); s \geq 0\}$ agree with E and $\mathcal{H}(\mathbf{1})$, respectively, on $[0,t]$. Hence, by definition (3.1.11) of (K, X) and the non-anticipative property of Λ proved in Lemma 2.13, we have

$$\big(K(\cdot \wedge t), X(\cdot \wedge t)\big) = \Lambda\big(E(\cdot \wedge t), X_0, Z_0 - \mathcal{H}_{\cdot\wedge t}(\mathbf{1})\big).$$

By the definition of the filtration $\{\mathcal{F}_t\}$, $\{E(s \wedge t); s \in [0, \infty)\}$ and $\{\mathcal{H}_{s\wedge t}(\mathbf{1}), s \in [0, \infty)\}$ are $\mathcal{F}_t$-measurable, while X_0 and Z_0 are $\sigma(Y_0)$-measurable. Therefore, by the continuity, and hence, measurability of the mapping Λ proved in Lemma 2.13, $X(t)$ and $K(\cdot \wedge t)$ are $\mathcal{F}_t^{Y_0}$-measurable. Moreover, it is clear from the definition of the mapping Γ_t that for every κ and $r \geq 0$, the value of $(\Gamma_t \kappa)(r)$ does not depend on the values of κ outside the interval $[0, t]$, and hence, $\Gamma_t K = \Gamma_t K(\cdot \wedge t)$. On the other hand, Γ_t is continuous by Lemma 3.16.c, and hence $\Gamma_t K(\cdot \wedge t)$ is also $\mathcal{F}_t^{Y_0}$-measurable. Finally, $\mathcal{M}_t(\Psi_{t+\cdot}\mathbf{1})$ is $\mathcal{F}_t$-measurable and $Z_0(t + \cdot)$ is $\sigma(Y_0)$-measurable by definition. Therefore, $Z(t, \cdot)$ defined in (3.1.12) is $\mathcal{F}_t^{Y_0}$-measurable. Consequently, Y is $\{\mathcal{F}_t^{Y_0}\}$-adapted, and condition a follows.

We now turn to the proof of condition d. By Lemma 3.19 it suffices to show that X and Z satisfy equations (3.3.6), (3.1.6) and (3.1.7). First, substituting $\int_0^t \partial_r Z(s, r) ds$ from (3.3.2), and using the definition (3.1.10) of Γ_t, the right-hand side of (3.3.6) is equal to

$$
\begin{aligned}
Z_0(r) + Z_0(t + r) - Z_0(r) + \mathcal{M}_t(\Phi_r \mathbf{1}) - \mathcal{M}_t(\Psi_{t+r}\mathbf{1}) \\
- \int_0^t K(s) g(t + r - s) ds - \mathcal{M}_t(\Phi_r \mathbf{1}) + \overline{G}(r) K(t) \\
= Z_0(t + r) - \mathcal{M}_t(\Psi_{t+r}\mathbf{1}) + \Gamma_t K(r).
\end{aligned}
$$

By (3.1.12), this is equal to $Z(t, r)$, which proves (3.3.6). Moreover, Proposition 3.17.b shows that the relation $X(t) \wedge 0 = Z(t, 0)$ in (3.1.6) holds. Finally, combining this relation with the expression for $X^+(t)$ from (3.1.14), we have almost surely, for every $t \geq 0$,

$$
X(t) = X^+(t) + X(t) \wedge 0 = X_0^+ + B(t) - \beta t - K(t) + Z(t, 0).
$$

Together with the expression for $Z(t, 0)$ in (3.1.12) and for Γ_t in (3.1.10), both with $r = 0$, this implies

$$
X(t) = X_0^+ + B(t) - \beta t + Z_0(t) - \mathcal{H}_t(\mathbf{1}) - \int_0^t K(s) g(t - s) ds. \tag{3.3.8}
$$

whereas substituting $r = 0$ in (3.3.2), we have almost surely, for every $t \geq 0$,

$$\int_0^t \partial_r Z(s,0)ds = Z_0(t) - Z_0(0) + \mathcal{M}_t(\mathbf{1}) - \mathcal{H}_t(\mathbf{1}) - \int_0^t K(s)g(t-s)ds, \tag{3.3.9}$$

where we have used the identities $\mathcal{H}_t(\mathbf{1}) = \mathcal{M}_t(\Psi_t \mathbf{1})$ and $\Phi_0 \mathbf{1} = \mathbf{1}$. Equation (3.1.7) then follows on substituting $Z_0(t) - \mathcal{H}_t(\mathbf{1}) - \int_0^t K(s)g(t-s)ds$ from (3.3.9) in (3.3.8) and using the relation $Z_0(0) = X(0) \wedge 0$. This completes the proof. $\square$

3.3.2. Uniqueness. This section is devoted to showing that the coupled stochastic age equations have a unique solution. We first establish a result on uniqueness of the weak solution to a transport equation within a certain class of functions.

LEMMA 3.21. Suppose Assumptions II.a and III hold, and a function $F \in \mathbb{C}_{\mathbb{R}}[0,\infty)$ with $F(0) = 0$ is given. Let $\xi : [0,\infty) \times [0,\infty) \mapsto \mathbb{R}$ be a function that satisfies the following two properties:

(a) The mapping $t \mapsto \xi(t,\cdot)$ lies in $\mathbb{C}([0,\infty); \mathbb{H}^1(0,\infty))$. For each $t > 0$, let $\partial_r \xi(t,\cdot)$ denote the density of $\xi(t,\cdot)$.
(b) The map $\partial_r \xi : (0,\infty) \times (0,\infty) \mapsto \mathbb{R}$ is locally integrable.

Then ξ satisfies the equation

$$\xi(t,r) = \int_0^t \partial_r \xi(s,r)ds + \overline{G}(r)F(t), \qquad t,r \geq 0, \tag{3.3.10}$$

if and only if

$$\xi(t,r) = \Gamma_t F(r), \qquad t,r \geq 0, \tag{3.3.11}$$

with Γ_t defined as in (3.1.10).

PROOF. This result would be standard if ξ were continuously differentiable on $(0,\infty)^2$, and while there are several results for more general ξ, there appears to be no readily quotable result for the class of ξ satisfying the specific properties mentioned above. Thus, for completeness, we include a proof.

Define $\xi(t,r) = \Gamma_t F(r)$, for $t,r \geq 0$. First, we show that ξ is indeed a solution to (3.3.10). By Lemma 3.16, $t \mapsto \xi(t,\cdot) = \Gamma_t F \in \mathbb{C}([0,\infty); \mathbb{H}^1(0,\infty))$ and for every

$s \geq 0$, $\Gamma_s F$ has density

$$\partial_r \xi(t,r) = (\Gamma_s F)'(r) = -g(r)F(s) - \int_0^s F(u)g'(s+r-u)du, \qquad r \in (0,\infty).$$

Because F, g and g' are continuous (see Assumption III), the mapping $(s,r) \mapsto (\Gamma_t F)'(r)$ is measurable. Moreover, for $t,r \geq 0$,

$$\begin{aligned}
\int_0^t \partial_r \xi(s,r)ds + \overline{G}(r)F(t) &= \int_0^t (\Gamma_s F)'(r)ds + \overline{G}(r)F(t) \\
&= -g(r)\int_0^t F(s)ds - \int_0^t \int_0^s F(u)g'(r+s-u)du\, ds + \overline{G}(r)F(t) \\
&= -g(r)\int_0^t F(s)ds - \int_0^t F(u)\left(g(r+t-u) - g(r)\right)du + \overline{G}(r)F(t) \\
&= -\int_0^t F(u)g(r+t-u)du + \overline{G}(r)F(t) \\
&= \Gamma_t F(r),
\end{aligned}$$

where the application of Fubini's theorem in the second equality above is justified because g' and F are continuous and hence locally integrable. This shows that ξ satisfies (3.3.10).

Next, let $\tilde{\xi}$ be any function that satisfies properties a and b and equation (3.3.10). Then $\Delta\xi \doteq \xi - \tilde{\xi}$ also satisfies properties a and b of the lemma, as well as the following equation:

$$\Delta\xi(t,r) = \int_0^t \partial_r \Delta\xi(s,r)ds, \qquad t,r \geq 0. \tag{3.3.12}$$

In particular, note that $\Delta\xi(0,r) = 0$ for all $r \in (0,\infty)$. Moreover, since for every $r > 0$, $s \mapsto \partial_r \Delta\xi(s,r)$ is locally integrable, $t \mapsto \int_0^t \partial_r \Delta\xi(s,r)ds$ is absolutely continuous, its weak derivative coincides with its density (see [**34**, Problem 5 on p. 290]), which is equal to $\partial_r \Delta\xi(t,r)$ by (3.3.12). Therefore, for every $\varphi \in \mathbb{C}_c([0,\infty) \times (0,\infty))$, using (3.3.12) for the first equation below and Fubini's theorem in the third and fifth equalities below (which is justified because φ is continuous with compact support and $\Delta\xi$ is locally integrable), we have

$$\begin{aligned}\int_0^\infty\int_0^\infty \Delta\xi(t,r)\partial_t\varphi(t,r)dtdr &= \int_0^\infty\int_0^\infty\left(\int_0^t \partial_r\Delta\xi(s,r)ds\right)\partial_t\varphi(t,r)dtdr\\ &= -\int_0^\infty\int_0^\infty \partial_r\Delta\xi(t,r)\varphi(t,r)dtdr\\ &= -\int_0^\infty\int_0^\infty \partial_r\Delta\xi(t,r)\varphi(t,r)drdt\\ &= \int_0^\infty\int_0^\infty \Delta\xi(t,r)\partial_r\varphi(t,r)drdt\\ &= \int_0^\infty\int_0^\infty \Delta\xi(t,r)\partial_r\varphi(t,r)dtdr.\end{aligned}$$

Thus, we have shown that

$$\int_0^\infty\int_0^\infty \Delta\xi(t,r)(-\partial_t\varphi(t,r)+\partial_r\varphi(t,r))dtdr = 0. \tag{3.3.13}$$

Equality (3.3.13) and the fact that $\Delta\xi \in \mathbb{C}([0,\infty);\mathbb{L}^2)$ (because $\Delta\xi$ satisfies property a of the lemma) imply that $\Delta\xi$ is a weak solution to the transport equation (3.3.10) with $F \equiv 0$ and $\Delta\xi(0,\cdot) = 0$. By [**80**, Chapter 2, Corollary 4.1], the unique solution to the transport equation in the space $\mathbb{C}([0,\infty);\mathbb{L}^2)$ is identically zero, and thus $\Delta\xi \equiv 0$. This completes the proof of uniqueness. □

PROPOSITION 3.22. Suppose Assumptions II and III hold. Then, for every $\mathcal{Y}$-valued random element $Y_0 = (X_0, Z_0)$, there is at most one solution of the coupled stochastic age equations of Definition 3.1 with initial condition Y_0.

PROOF. For $i = 1,2$, let $Y_i = (X_i, Z_i)$ be a solution to the coupled stochastic age equations with initial condition Y_0, and define $K_i(t) \doteq B(t) - \beta t - X_i^+(t) + X_0^+$. We need to show that $Y_1 = Y_2$. Denote $\Delta H = H_1 - H_2$ for $H = Y, X, Z, K$, and note that $\Delta K(t) = X_2^+(t) - X_1^+(t), t \geq 0$. By Lemma 3.19, for $i = 1,2$, Z_i satisfies the equation (3.3.6) with K replaced by K_i, and hence ΔZ satisfies the following non-homogenous transport equation:

$$\Delta Z(t,r) = \int_0^t \partial_r Z(s,r)ds + \overline{G}(r)\Delta K(t), \quad t,r \geq 0,$$

with $\Delta Z(0,\cdot) \equiv 0$. By Definition 3.1, both Z_1, Z_2 and hence ΔZ satisfy properties a and b of Lemma 3.21. Therefore, by (3.3.11) of Lemma 3.21, with F replaced by ΔK, we have

$$\Delta Z(t,r) = \Gamma_t \Delta K(r), \qquad t,r \geq 0. \tag{3.3.14}$$

Since ΔK is continuous almost surely, an application of Lemma 3.16.a. shows that for every $t \geq 0$, $\Delta Z(t,\cdot)$ is continuously differentiable on $[0,\infty)$, with continuous density

$$\partial_r \Delta Z(s,r) = -g(r)\Delta K(s) - \int_0^s \Delta K(u) g'(r+s-v)dv, \qquad r,s \in [0,\infty).$$

In particular, the function $R(s) \doteq \partial_r \Delta Z(s,0)$ is well defined on $[0,\infty)$ and satisfies

$$R(s) = -g(0)\Delta K(s) - \int_0^s \Delta K(u) g'(s-v)dv, \tag{3.3.15}$$

and, recalling the constants H and H_2 from Remark 2.2, and the fact that $\int_0^\infty \bar{G}(x) = 1$ by Assumption II.a, satisfies

$$|\partial_r \Delta Z(s,r)| \leq (H + H_2)\|\Delta K\|_t, \qquad \forall s \in [0,t], r \geq 0. \tag{3.3.16}$$

On the other hand, for $i = 1,2$, X_i satisfies (3.1.7) with Z replaced by Z_i. Therefore,

$$\Delta X(t) = \lim_{r \to 0} \int_0^t \partial_r \Delta Z(s,r)ds = \int_0^t \partial_r \Delta Z(s,0)ds = \int_0^t R(s)ds, \tag{3.3.17}$$

where the second equality holds due to the continuity of $r \mapsto \partial_r \Delta Z(s,r)$ for every $s \geq 0$, and an application of the bounded convergence theorem, which is justified by (3.3.16). In turn, this implies

$$|\Delta K(t)| = |X_1^+(t) - X_2^+(t)| \leq |\Delta X(t)| \leq \int_0^t |R(s)|ds, \quad t \in [0,\infty), \tag{3.3.18}$$

which, when combined with (3.3.15) and the bounds in Assumption Remark 2.2, shows that for every $T < \infty$ and $t \in [0, T]$,

$$
\begin{aligned}
|R(t)| &\leq g(0)|\Delta K(t)| + \int_0^t |g'(t-s)||\Delta K(s)|ds \\
&\leq H \int_0^t |R(s)|ds + H_2 \int_0^t \left(\int_0^v |R(s)|ds \right) dv \\
&\leq (H + TH_2) \int_0^t |R(s)|ds.
\end{aligned}
$$

Since $R(0) = 0$, by Gronwall's lemma this implies $R(t) = 0$ for all $t \geq 0$. When combined with (3.3.18), this implies that $\Delta X(t) = \Delta K(t) = 0$ for all $t \geq 0$, which in turn implies $\Delta Z(t, \cdot) = 0$ due to (3.3.14). Thus, we have shown that $Y_1(t) = Y_2(t)$ for all $t \geq 0$, which proves the desired uniqueness. $\square$

3.3.3. Markov Property. We now show that the diffusion limit is a time-homogeneous Feller Markov process. For $s \geq 0$, we define the operator Θ_s as follows: for every function F on $[0, \infty)$, the shifted function $\Theta_s F$ is defined by

$$
(\Theta_s F)(t) \doteq F(s+t) - F(s), \qquad t \geq 0. \tag{3.3.19}
$$

Also, for every $f \in \mathbb{C}_b[0, \infty)$, we define the shifted convolution integral as

$$
(\Theta_s \mathcal{H})_t(f) \doteq \mathcal{M}_{s+t}(\Psi_{s+t} f) - \mathcal{M}_s(\Psi_{s+t} f), \qquad t \geq 0. \tag{3.3.20}
$$

In particular, note that using the identity (2.3.8), we have

$$
(\Theta_s \mathcal{H})_t(\Phi_r \mathbf{1}) = \mathcal{M}_{s+t}(\Psi_{s+t+r} \mathbf{1}) - \mathcal{M}_s(\Psi_{s+t+r} \mathbf{1}). \tag{3.3.21}
$$

LEMMA 3.23. Suppose Assumptions II and III hold. Then almost surely, for every $s \geq 0$, we have

$$
Z(s+t, r) = Z(s, t+r) + (\Gamma_t \Theta_s K)(r) - (\Theta_s \mathcal{H})_t(\Phi_r \mathbf{1}), \quad t, r \geq 0, \tag{3.3.22}
$$

and $(\Theta_s K, X(s + \cdot))$ solves the CMS equation associated with $(\Theta_s E, X(s), Z(s, \cdot) - (\Theta_s \mathcal{H})(\mathbf{1}))$, that is,

$$
(\Theta_s K, X(s + \cdot)) = \Lambda \left(\Theta_s E, X(s), Z(s, \cdot) - (\Theta_s \mathcal{H}).(\mathbf{1}) \right). \tag{3.3.23}
$$

PROOF. By definitions (3.1.12) and (3.1.10) of Z and Γ_t, respectively, we have

$$Z(s,t+r) = Z_0(s+t+r) - \mathcal{M}_s(\Psi_{s+t+r}\mathbf{1}) + \overline{G}(t+r)K(s) - \int_0^s K(v)g(s+t+r-v)dv. \tag{3.3.24}$$

Also, by definition (3.3.19) of $\Theta_s K$ and (3.1.10) of Γ_t, we have

$$\begin{aligned}
(\Gamma_t \Theta_s K)(r) &= \overline{G}(r)(\Theta_s K)(t) + \int_0^t (\Theta_s K)(v) g(t+r-v)dv \\
&= \overline{G}(r)\big(K(s+t) - K(s)\big) - \int_0^t \left(K(s+v) - K(s)\right) g(t+r-v)dv \\
&= \overline{G}(r)K(s+t) - \overline{G}(r)K(s) - \int_s^{t+s} K(v)g(t+s+r-v)dv \\
&\quad - K(s)\big(\overline{G}(t+r) - \overline{G}(r)\big) \\
&= \overline{G}(r)K(s+t) - \int_0^{t+s} K(v)g(t+s+r-v)dv - \overline{G}(t+r)K(s) \\
&\quad + \int_0^s K(v)g(t+s+r-v)dv \\
&= (\Gamma_{s+t}K)(r) - (\Gamma_s K)(t+r). \qquad (3.3.25)
\end{aligned}$$

Substituting Z from (3.1.12) (with t and r replaced by s and $t+r$, respectively) and using equations (3.3.21) and (3.3.25), the right-hand side of (3.3.22) is equal to

$$\begin{aligned}
&Z_0(s+t+r) - \mathcal{M}_s(\Psi_{s+t+r}\mathbf{1}) + \Gamma_s K(t+r) + (\Gamma_{s+t}K)(r) - (\Gamma_s K)(t+r) \\
&\quad - \mathcal{M}_{s+t}(\Psi_{s+t+r}\mathbf{1}) + \mathcal{M}_s(\Psi_{s+t+r}\mathbf{1}) \\
&\qquad = Z_0(s+t+r) + (\Gamma_{s+t}K)(r) - \mathcal{M}_{s+t}(\Psi_{s+t+r}\mathbf{1}) \\
&\qquad = Z(s+t,r),
\end{aligned}$$

where the last equality once again uses (3.1.12). This proves (3.3.22).

To prove (3.3.23), subtract equation (3.1.14) with t replaced by s from the same equation with t replaced by $t+s$, and using (3.1.8), we get

$$\begin{aligned}
(\Theta_s K)(t) = K(s+t) - K(s) &= E(s+t) - E(s) - X^+(s+t) + X^+(s), \\
&= (\Theta_s E)(t) - X^+(s+t) + X^+(s). \qquad (3.3.26)
\end{aligned}$$

Additionally, by Proposition 3.17.b with t replaced by $s+t$,

$$X(s+t) \wedge 0 = X(s+t) \wedge 0 = Z(s+t,0).$$

Substituting $Z(s+t,0)$ from (3.3.22), using definition (3.1.10) of Γ_t and the identity $\Phi_0 \mathbf{1} = \mathbf{1}$, we obtain

$$X(s+t) \wedge 0 = Z(s,t) - (\Theta_s \mathcal{H})_t(\mathbf{1}) + (\Theta_s K)(t) - \int_0^t (\Theta_s K)(v) g(t-v) dv. \tag{3.3.27}$$

Equation (3.3.23) then follows from (3.3.26), (3.3.27) and Definition 2.12 of the CMS mapping Λ (also see Lemma 2.13). □

LEMMA 3.24. Suppose Assumptions II and III hold. Then, for every $s \geq 0$, the processes $\{\Theta_s E(t); t \geq 0\}$ and $\{(\Theta_s \mathcal{H})_t(\mathbf{1}); t \geq 0\}$ are independent of $\mathcal{F}_s$, and have the same distribution as the processes E and $\mathcal{H}(\mathbf{1})$, respectively. Moreover, for every $s, t \geq 0$, the $\mathbb{H}^1(0,\infty)$-valued random element $(\Theta_s \mathcal{H})_t(\Phi.\mathbf{1})$ is independent of $\mathcal{F}_s$ and has the same distribution as $\mathcal{H}_t(\Phi.\mathbf{1})$.

PROOF. Fix $s \geq 0$. By definition (3.1.8) of E and (3.3.19) of $\Theta_s E$, we have

$$\Theta_s E(t) = E(s+t) - E(s) = B(s+t) - B(s) - \beta t, \qquad t \geq 0.$$

Since Brownian motion B has independent stationary increments and is independent of $\mathcal{M}$, $B(s+t) - B(s)$ is independent of $\mathcal{F}_s$ for all $t \geq 0$, and $\{B(s+t) - B(s); t \geq 0\}$ is itself a Brownian motion. Hence, the claim for $(\Theta_s E)$ follows.

Next, define the martingale measure $\tilde{\mathcal{M}} = \{\tilde{\mathcal{M}}_t(A); t \geq 0, A \in \mathcal{B}[0,\infty)\}$ as follows: $\tilde{\mathcal{M}}_t(A) \doteq \mathcal{M}_{s+t}(A) - \mathcal{M}_s(A)$ for $t \geq 0$. Since the martingale measure $\mathcal{M}$ is a white noise independent of B, $\tilde{\mathcal{M}}$ is again a white noise with the same distribution as $\mathcal{M}$, and independent of $\mathcal{F}_s$. Also, for any continuous f, using (3.3.20) we have

$$
\begin{aligned}
(\Theta_s \mathcal{H})_t(f) &= \mathcal{M}_{s+t}(\Psi_{s+t} f) - \mathcal{M}_s(\Psi_{s+t} f) \\
&= \int_s^{s+t} \int_0^\infty \frac{\overline{G}(t+s-v+x)}{\overline{G}(x)} f(x) \mathcal{M}(dv, dx) \\
&= \int_0^t \int_0^\infty \frac{\overline{G}(t-v+x)}{\overline{G}(x)} f(x) \tilde{\mathcal{M}}(dv, dx) \\
&= \tilde{\mathcal{M}}_t(\Psi_t f).
\end{aligned}
$$

Substituting $f = \mathbf{1}$ and $f = \Phi_r \mathbf{1}, r \geq 0$, we conclude that the processes $\{(\Theta_s \mathcal{H})_t(\mathbf{1}); t \geq 0\}$ and $(\Theta_s \mathcal{H})_t(\Phi.\mathbf{1})$ are independent of $\mathcal{F}_s$, and have the same distribution as $\{\mathcal{H}_t(\mathbf{1}); t \geq 0\}$ and $\mathcal{H}_t(\Phi.\mathbf{1})$, respectively. □

PROPOSITION 3.25. Suppose Assumptions II and III hold, and for $y \in \mathcal{Y}$, let $Y_y(t)$ be as defined in Definition 3.3. Then $\{Y_y(t), \mathcal{F}_t; t \geq 0\}$ is a time-homogeneous Feller Markov family on the state space $\mathcal{Y}$.

PROOF. Recall from Proposition 3.9 that for any $s, t \geq 0$, $\mathcal{M}_{s+t}(\Psi_{s+t+\cdot}\mathbf{1})$ and $\mathcal{M}_s(\Psi_{s+t+\cdot}\mathbf{1})$ both lie in $\mathbb{H}^1(0,\infty)$ almost surely, and hence, so does $(\Theta_s \mathcal{H})_t(\Phi.\mathbf{1})$, using equation (3.3.21). We claim that for any initial condition $y = (x_0, z_0) \in \mathcal{Y}$ and associated diffusion limit $Y_y = (X, Z)$, for every $t \geq 0$, there exists a continuous mapping $\Pi_t : \mathbb{R} \times \mathbb{H}^1(0,\infty) \times \mathbb{C}_{\mathbb{R}}[0,\infty)^2 \times \mathbb{H}^1(0,\infty) \mapsto \mathbb{R} \times \mathbb{H}^1(0,\infty)$ such that for every $s \geq 0$,

(3.3.28)

$$\big(X(s+t), Z(s+t,\cdot)\big) = \Pi_t\big(X(s), Z(s,\cdot), \Theta_s E(\cdot), (\Theta_s \mathcal{H}).(\mathbf{1}), (\Theta_s \mathcal{H})_t(\Phi.\mathbf{1})\big).$$

To see why the claim is true, for notational convenience first set $\mathbb{D} \doteq \mathbb{R} \times \mathbb{H}^1(0,\infty) \times \mathbb{C}^0_R[0,\infty)^2$ and note that by Lemma 2.13, the CMS mapping $\Lambda : \mathbb{C}^0_R[0,\infty) \times \mathbb{R} \times \mathbb{C}_{\mathbb{R}}[0,\infty) \mapsto \mathbb{C}^0_R[0,\infty) \times \mathbb{C}_{\mathbb{R}}[0,\infty)$ is continuous. Together with the representation for $(X(s+\cdot), \Theta_s K(\cdot))$ in terms of Λ given in (3.3.23) of Lemma 3.23, the fact that the embedding from $\mathbb{H}^1(0,\infty)$ to $\mathbb{C}_{\mathbb{R}}[0,\infty)$ is continuous (see Lemma 2.5.a) and the continuity of the evaluation map $f \in \mathbb{C}_{\mathbb{R}}[0,\infty) \mapsto f(t) \in \mathbb{R}$, it follows that the mapping from $\mathbb{D}$ to $\mathbb{C}^0_R[0,\infty) \times \mathbb{C}_{\mathbb{R}}[0,\infty)$ that takes $(X(s), Z(s,\cdot), \Theta_s E(\cdot), (\Theta_s \mathcal{H}).(\mathbf{1}))$

to $(\Theta_s K(\cdot), X(s+t))$ is continuous (and only depends on t). Together with the continuity of Γ_t established in Lemma 3.16.b, this also implies that the map that takes $(X(s), Z(s,\cdot), \Theta_s E(\cdot), (\Theta_s \mathcal{H}).(\mathbf{1}))$ in $\mathbb{D}$ to $\Gamma_t \Theta_s K$ in $\mathbb{H}^1(0,\infty)$ is continuous. Furthermore, since the shift $f(\cdot) \mapsto f(t+\cdot)$ is a continuous mapping from $\mathbb{H}^1(0,\infty)$ to itself, the mapping $Z(s,\cdot) \mapsto Z(s,t+\cdot)$ from $\mathbb{H}^1(0,\infty)$ to itself is continuous. Since the representation (3.3.22) expresses $Z(s+t,\cdot)$ as a sum of $\Gamma_t\Theta_s K(\cdot)$, $Z(s,t+\cdot)$ and $(\Theta_s\mathcal{H})_t(\Phi.\mathbf{1})$, it follows that $Z(s+t,\cdot)$ is also obtained as a continuous mapping (depending only on t) of the arguments of Π_t in (3.3.28). This completes the proof of the claim.

We now show that the Feller property follows from the claim above. This can be deduced from the proof of Theorem 5(2) in Section 9.5 of **[62]**, but since in our case the Markov process is homogeneous, the proof is simpler and so we include it below. Let $Y = (X, Z)$ be the diffusion limit starting at a deterministic initial value $y_0 \in \mathcal{Y}$. Since y_0 is deterministic, $\mathcal{F}_s^{y_0} = \mathcal{F}_s \vee \sigma(y_0) = \mathcal{F}_s$ and hence, by condition 1. of Definition 3.1 and Proposition 3.20, $(X(s), Z(s,\cdot))$ is $\mathcal{F}_s$-measurable. Also, $\Theta_s E(\cdot)$, $(\Theta_s\mathcal{H}).(\mathbf{1})$ and $(\Theta_s\mathcal{H})_t(\Phi.\mathbf{1})$ are independent of $\mathcal{F}_s$ by Lemma 3.24. Therefore, for every bounded and measurable functional $F : \mathcal{Y} \mapsto \mathbb{R}$, using the claim (3.3.28) above we have

$$
\begin{aligned}
\mathbb{E}\left[F(Y(s+t))|\mathcal{F}_s\right] &= \mathbb{E}\left[F\left(\Pi_t\big(X(s), Z(s,\cdot), \Theta_s E(\cdot), (\Theta_s\mathcal{H}).(\mathbf{1}), (\Theta_s\mathcal{H})_t(\Phi.\mathbf{1})\big)\right)\Big|\mathcal{F}_s\right] \\
&= \mathbb{E}\left[F\left(\Pi_t\big(X(s), Z(s,\cdot), \Theta_s E(\cdot), (\Theta_s\mathcal{H}).(\mathbf{1}), (\Theta_s\mathcal{H})_t(\Phi.\mathbf{1})\big)\right)\Big|Y(s)\right] \\
&= \mathbb{E}\left[F(Y(s+t))|Y(s)\right],
\end{aligned}
$$

which shows that $\{Y_t, \mathcal{F}_t\}$ is a Markov process. In fact, it is easy to see that $\{Y_y, \mathcal{F}_t; y \in \mathcal{Y}\}$ is a time-homogeneous Markov family. Indeed, by another use of Lemma 3.24, for every $y = (x, z) \in \mathcal{Y}$,

$$
\begin{aligned}
&\mathbb{E}\left[F(Y(s+t))|Y(s)=y\right] \\
&\qquad = \mathbb{E}\left[F\left(\Pi_t\big(X(s), Z(s,\cdot), \Theta_s E(\cdot), (\Theta_s\mathcal{H}).(\mathbf{1}), (\Theta_s\mathcal{H})_t(\Phi.\mathbf{1})\big)\right)\Big| Y(s)=y\right] \\
&\qquad = \mathbb{E}\left[F\left(\Pi_t\big(x, z, E, \mathcal{H}(\mathbf{1}), \mathcal{H}_t(\Phi.\mathbf{1})\big)\right)\right] \\
&\qquad = \mathbb{E}\left[F(Y_y(t))\right]. \qquad (3.3.29)
\end{aligned}
$$

Lastly, for the Feller property, first recall from Proposition 3.17.b that Y has continuous sample paths, almost surely. In addition, by the continuity of the mapping Π_t and (3.3.29), the mapping $y \mapsto \mathbb{E}\left[F(Y_y)(t)\right]$ is continuous for every bounded and continuous functional F. Therefore, $\{Y_y, \mathcal{F}_t; y \in \mathcal{Y}\}$ is a Feller family (see **[13**, Definition 16.20]). □

PROOF OF THEOREM 3.4. Existence of a solution follows from Proposition 3.20 while uniqueness follows from Proposition 3.22, and the Feller Markov property is proved in Proposition 3.25. □

3.4. Uniqueness of the Invariant Distribution

In this section, we prove the main result of this chapter: the uniqueness of invariant distribution of the limit process (Theorem 3.26).

Let $\{\mathcal{P}_t\} = \{\mathcal{P}_t; t \geq 0\}$ denote the transition semigroup of the time-homogeneous Feller Markov family of solutions Y to the coupled stochastic age equations on the state space $\mathcal{Y}$. Recall that a probability measure μ on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ is said to be an invariant distribution for the semigroup $\{\mathcal{P}_t\}$ if and only if

$$\mathcal{P}_t\mu = \mu, \qquad t \geq 0. \qquad (3.4.1)$$

Our main result establishes uniqueness of the invariant distribution of this semigroup.

THEOREM 3.26. Suppose Assumptions II and III are satisfied. Then the transition semigroup $\{\mathcal{P}_t\}$ on $\mathcal{Y}$ has at most one invariant distribution.

PROOF. Theorem 3.26 is proved in Section 3.4.6. $\square$

REMARK 3.27. We show later in Section 4.2.3 that every subsequential limit of the sequence of stationary distributions $\{\widehat{\pi}^{(N)}\}$ is in fact an invariant distribution of the semigroup $\{\mathcal{P}_t\}$, which proves the existence of an invariant distribution (see Corollary 4.23).

As mentioned in the introduction, to prove uniqueness of the invariant distribution, we will adopt the so-called *asymptotic equivalent coupling* method, which is particularly well suited for infinite-dimensional Markov processes. In Section 3.4.1, we first describe this method in the generality required for our problem, following the exposition in [**50**]. At the end of Section 3.4.1, we discuss the main steps involved in applying this framework to our problem, which are carried out in the subsequent sections.

3.4.1. Asymptotic Coupling: The General Framework. Let $\mathcal{X}$ be a Polish space with a compatible metric $d(\cdot,\cdot)$, equipped with the Borel σ-algebra $\mathcal{B}(\mathcal{X})$. As usual, let $\mathcal{X}^{\mathbb{R}_+}$ denote the space of $\mathcal{X}$-valued functions on $[0,\infty)$. Recall that a subset $C \subset \mathcal{X}^{\mathbb{R}_+}$ is called a cylinder set if there exists $n \in \mathbb{N}$, $t_1, ..., t_n \in [0,\infty)$ and Borel sets $B_1, ..., B_n \in \mathcal{B}(\mathcal{X})$ such that

$$C = \{f \in \mathcal{X}^{\mathbb{R}_+}\,;\ (f(t_1), ..., f(t_n)) \in (B_1, ..., B_n)\}.$$

We endow $\mathcal{X}^{\mathbb{R}_+}$ with the Kolmogorov σ-algebra $\mathcal{B}(\mathcal{X})^{\mathbb{R}_+}$, which is the σ-algebra generated by all cylinder sets. Let $\mathbb{M}_1(\mathcal{X}^{\mathbb{R}_+})$ and $\mathbb{M}_1(\mathcal{X}^{\mathbb{R}_+} \times \mathcal{X}^{\mathbb{R}_+})$ denote the spaces of probability measures on $(\mathcal{X}^{\mathbb{R}_+}, \mathcal{B}(\mathcal{X})^{\mathbb{R}_+})$ and $(\mathcal{X}^{\mathbb{R}_+} \times \mathcal{X}^{\mathbb{R}_+}, \mathcal{B}(\mathcal{X})^{\mathbb{R}_+} \otimes \mathcal{B}(\mathcal{X})^{\mathbb{R}_+})$. For every $m_1, m_2 \in \mathbb{M}_1(\mathcal{X}^{\mathbb{R}_+})$, recall that a coupling of m_1 and m_2 is a probability measure $\Upsilon \in \mathbb{M}_1(\mathcal{X}^{\mathbb{R}_+} \times \mathcal{X}^{\mathbb{R}_+})$ whose first and second marginals, respectively, are m_1 and m_2, that is, $\Pi^{(i)}_{\#}\Upsilon = m_i$ for $i = 1, 2$, where, $\Pi^{(i)}$ is the ith coordinate projection map, and $\Pi^{(i)}_{\#}\Upsilon$ is the push-forward of the measure Υ under $\Pi^{(i)}$. Define $\mathcal{C}(m_1, m_2) \subset \mathbb{M}_1(\mathcal{X}^{\mathbb{R}_+} \times \mathcal{X}^{\mathbb{R}_+})$ to be the set of all couplings of m_1 and m_2. One can extend $\mathcal{C}(m_1, m_2)$ by relaxing the definition of a coupling to define the space of *absolutely continuous couplings* as follows:

$$\tilde{\mathcal{C}}(m_1, m_2) \doteq \{\Upsilon \in \mathbb{M}_1(\mathcal{X}^{\mathbb{R}_+} \times \mathcal{X}^{\mathbb{R}_+});\ \Pi^{(i)}_{\#}\Upsilon \ll m_i, i = 1, 2\}. \tag{3.4.2}$$

If Υ in $\tilde{\mathcal{C}}(m_1, m_2)$ satisfies $\Pi^{(i)}_{\#}\Upsilon \sim m_i$, $i = 1, 2$, Υ will be referred to as an *equivalent coupling* of m_1 and m_2. In contrast to a coupling, the corresponding marginals of an absolutely continuous (or equivalent) coupling Υ need only be absolutely continuous with respect to (respectively, equivalent to), and not necessarily equal to, m_1 and m_2, respectively.

Let $\{\mathcal{P}_t\} = \{\mathcal{P}_t;\, t \geq 0\}$ be a semigroup of Markov kernels on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Recall that a probability measure μ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ is called an invariant distribution for the semigroup $\{\mathcal{P}_t\}$ if (3.4.1) holds for every $t \geq 0$. For every probability measure μ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, let $\mathcal{P}_{[0,\infty)}\mu$ denote the distribution of the Markov process with initial distribution μ and transition semigroup $\{\mathcal{P}_t\}$ on the path space $(\mathcal{X}^{\mathbb{R}_+}, \mathcal{B}(\mathcal{X})^{\mathbb{R}_+})$. Let $\mathcal{D}$ be the set of pairs of paths that meet at infinity:

$$\mathcal{D} \doteq \left\{(x, y) \in \mathcal{X}^{\mathbb{R}_+} \times \mathcal{X}^{\mathbb{R}_+} :\ \lim_{t\to\infty} d(x(t), y(t)) = 0\right\}, \tag{3.4.3}$$

and note that $\mathcal{D} \in \mathcal{B}(\mathcal{X})^{\mathbb{R}_+} \otimes \mathcal{B}(\mathcal{X})^{\mathbb{R}_+}$.

We now state a continuous version of a result established in [**50**, Corollary 2.2], which provides a sufficient condition for uniqueness of the invariant measure for the Markov semigroup $\{\mathcal{P}_t\}$.

PROPOSITION 3.28. Assume there exists a measurable set $A \in \mathcal{B}(\mathcal{X})$ and a mapping $\Upsilon : A \times A \ni (y, \tilde{y}) \mapsto \Upsilon_{y,\tilde{y}} \in \tilde{\mathcal{C}}(\mathcal{P}_{[0,\infty)}\delta_y, \mathcal{P}_{[0,\infty)}\delta_{\tilde{y}})$ with the following properties:

- (I) $\mu(A) > 0$ for any invariant probability measure μ of $\{\mathcal{P}_t\}$.
- (II) For every measurable set $B \in \mathcal{B}(\mathcal{X})^{\mathbb{R}_+} \otimes \mathcal{B}(\mathcal{X})^{\mathbb{R}_+}$, $(y, \tilde{y}) \mapsto \Upsilon_{y,\tilde{y}}(B)$ is measurable.
- (III) For every $y, \tilde{y} \in A$, $\Upsilon_{y,\tilde{y}}(\mathcal{D}) > 0$.

Then $\{\mathcal{P}_t\}$ has at most one invariant probability measure.

PROOF. The proof of Proposition 3.28 follows from essentially the same arguments as those used in [50, Theorem 1.1. and Corollary 2.2] in [50], except that the space $\mathcal{X}^\infty$ and discrete shift operator θ should be replaced, respectively, by $\mathcal{X}^{\mathbb{R}_+}$ and the continuous shift operator $\theta_s, s > 0$, and the continuous version of Birkhoff's ergodic theorem (see, e.g., [68, Theorem 1 of Section 1.2]) should be invoked to obtain the analog of (2.1) in [50]. A full proof is given in Appendix 3.A. $\square$

REMARK 3.29. As observed in Remark 1.2 of [50], for the purpose of Proposition 3.28, in the definition (3.4.2) of $\tilde{\mathcal{C}}$, one can without loss of generality replace absolute continuity by (the apparently stronger condition of) equivalence. This is because if there is an absolutely continuous coupling Υ that satisfies conditions (II) and (III) of Proposition 3.28, then the measure $\frac{1}{2}(\Upsilon + \mathcal{P}_{[0,\infty)}\mu_1 \otimes \mathcal{P}_{[0,\infty)}\mu_2)$ is an equivalent coupling that also satisfies the same conditions. Thus, we refer to the approach to establishing uniqueness of the invariant distribution by invoking Proposition 3.28 as the *asymptotic equivalent coupling* approach.

In Sections 3.4.2-3.4.6, we apply the asymptotic equivalent coupling framework of Proposition 3.28, with the state space $\mathcal{Y}$ of our diffusion limit playing the role of the generic Polish space $\mathcal{X}$, in order to establish uniqueness of the invariant distribution of the $\mathcal{Y}$-valued Markov process defined in Section 3.1.2. Let A be the measurable subset of $\mathcal{Y}$ defined as

$$A \doteq \{(x,z) \in \mathcal{Y} : x \geq 0\}. \tag{3.4.4}$$

First, in Section 3.4.2, for each pair $y, \tilde{y} \in A$, we construct a pair of stochastic processes $(Y_y, \tilde{Y}_{\tilde{y}})$ on a common probability space and define the mapping

$$\Upsilon_{y,\tilde{y}} : (y,\tilde{y}) \in A \times A \mapsto \mathcal{L}aw(Y_y, \tilde{Y}_{\tilde{y}}) \in \mathbb{M}_1(\mathcal{Y}^{\mathbb{R}_+} \times \mathcal{Y}^{\mathbb{R}_+}), \tag{3.4.5}$$

where $\mathcal{L}aw(Y_y, \tilde{Y}_{\tilde{y}})$ denotes the joint distribution of $(Y_y, Y_{\tilde{y}})$. The required measurability properties are proved in Section 3.4.3. Next, in Section 3.4.4, we show that $\Upsilon_{y,\tilde{y}}(\mathcal{D}) > 0$, and in Section 3.4.5 we show that the law of Y_y is $\mathcal{P}_{[0,\infty)}\delta_y$, and that the law of $\tilde{Y}_{\tilde{y}}$ is equivalent to $\mathcal{P}_{[0,\infty)}\delta_{\tilde{y}}$, thus establishing that Y defines an asymptotic equivalent coupling. Finally, these results are combined in Section 3.4.6 to complete the proof of Theorem 3.26.

REMARK 3.30. It is worthwhile to clarify why, in Proposition 3.28, we formulate a continuous-time version of the asymptotic coupling theorem, rather than deduce the result by simply applying the original discrete-time version established in [**50**, Corollary 2.2] to the discrete skeleton of the continuous-time process (i.e., the Markov chain obtained by sampling at integer times). To show uniqueness of the invariant measure of the continuous-time Markov process, is clearly suffices to show uniqueness of the invariant measure of its discrete skeleton (that is, the Markov chain obtained by sampling the continuous process at integer times). In turn, by Corollary 2.2 of [**50**], for uniqueness of the invariant measure of the discrete skeleton, it suffices to verify the three conditions of that corollary, which are the natural discrete analogs of properties (I), (II) and (III) of Proposition 3.28. Now, the continuous version of properties (II) and (III) immediately imply the discrete version because any asymptotic equivalent coupling of the continuous-time process on some set A induces a corresponding asymptotic equivalent coupling of the discrete skeleton. Thus, if $A = \mathcal{X}$ then the discrete-time version can be directly invoked because in this case the first condition of Corollary 2.2 of [**50**], which states that $\mu(A) > 0$ for every invariant measure μ of the discrete skeleton, is trivially satisfied. However, when A is a strict subset of $\mathcal{X}$ then the property that $\mu(A) > 0$ for all invariant measures μ of the continuous-time process need not imply that $\mu(A) > 0$ for all invariant measures of the discrete skeleton, since the latter could in general be strictly larger. Moreover, in some situations (as turns out to be the case in our application), it may be relatively easy to show the former, but nontrivial to show the latter. In such cases, it is more convenient to directly apply the continuous-time version of the result, as formulated in Proposition 3.28.

3.4.2. Construction of a Candidate Coupling. Fix two initial conditions $y = (x_0, z_0)$ and $\tilde{y} = (\tilde{x}_0, \tilde{z}_0)$ in the set A defined in (3.4.4), and let $Y(t) = (X(t), Z(t))$ be the diffusion limit of Definition 3.3 with initial condition y. Define the stochastic process

$$R(t) \doteq z_0'(t) + \mathcal{H}_t(h) - g(0)K(t) - \int_0^t K(s)g'(t-s)ds. \tag{3.4.6}$$

Combining (3.2.12) and (3.3.2), the latter with $r = 0$, we have

$$\int_0^t R(s)ds = \int_0^t \partial_r Z(s,0)ds. \tag{3.4.7}$$

We now construct another process $\tilde{Y}$, which starts from $\tilde{y}$, on the same probability space. Fix $\lambda > 0$. Given X defined above, it is easy to see that $\mathbb{P}$-almost surely, the linear integral equation

$$\tilde{X}(t) = \tilde{x}_0 - x_0 - \lambda \int_0^t \tilde{X}(s)ds + X(t) + \lambda \int_0^t X(s)ds, \tag{3.4.8}$$

has a unique continuous solution $\tilde{X}$, which has the from

$$\tilde{X}(t) = \tilde{x}_0 e^{-\lambda t} + F(t) - \lambda \int_0^t e^{-\lambda(t-s)}F(s)ds, \qquad t \geq 0, \tag{3.4.9}$$

where $F(t) \doteq X(t) - x_0 + \lambda \int_0^t X(s)ds$. Since Y satisfies the coupled stochastic age equations, by Proposition 3.20, X satisfies (3.1.7), which when combined with (3.4.7) and (3.4.8), shows that for $t \geq 0$,

(3.4.10)
$$\tilde{X}(t) = \tilde{x}_0 + B(t) - \beta t - \mathcal{M}_t(\mathbf{1}) - \lambda \int_0^t \tilde{X}(s)ds + \lambda \int_0^t X(s)ds + \int_0^t R(s)ds.$$

Also, for $t \geq 0$, define

$$\overline{K}(t) \doteq B(t) - \beta t - \left(\tilde{X}^+(t) - \tilde{x}_0^+\right) + \int_0^t \left(R(s) + \lambda X(s) - \lambda \tilde{X}(s)\right)ds. \tag{3.4.11}$$

Clearly, $\overline{K}$ is also continuous, almost surely. We now introduce a process $\tilde{R}$ that, as shown in Corollary 3.32 below, can be characterized as the unique (almost surely locally integrable) solution of the following renewal equation:

(3.4.12)

$$\tilde{R}(t) = \tilde{z}_0'(t) - g(0)\overline{K}(t) - \int_0^t \overline{K}(s)g'(t-s)ds + \mathcal{H}_t(h) + \int_0^t g(t-s)\tilde{R}(s)ds.$$

In Proposition 3.31, we first collect some general results on solutions of the renewal equation. Part a of the proposition is used below in the proof of Corollary 3.32, part b is used in Section 3.4.4 to establish an asymptotic convergence property of the diffusion limit , and part c is used in Section 3.4.3 to establish measurability properties of the candidate asymptotic coupling. Recall the definition of the convolution operator $*$ from Section 1.4.

PROPOSITION 3.31. Let G be a cumulative probability distribution on $[0,\infty)$ with a continuous density g.

(a) If $f \in \mathbb{L}^1_{\text{loc}}(0,\infty)$, the following renewal equation

$$\varphi = f + g * \varphi \tag{3.4.13}$$

has a unique locally integrable solution φ_*.

(b) If G has a finite second moment, the function f lies in $\mathbb{L}^2(0,\infty)$ and its integral $\mathcal{I}_f(t) \doteq \int_0^t f(s)ds, t \geq 0$, lies in $\mathbb{L}^2(0,\infty)$, then the solution φ_* of the renewal equation (3.4.13) also lies in $\mathbb{L}^2(0,\infty)$. Moreover, there exist deterministic constants $c_1, c_2 < \infty$ such that

$$\|\varphi_*\|_{\mathbb{L}^2} \leq c_1\|f\|_{\mathbb{L}^2} + c_2\|\mathcal{I}_f\|_{\mathbb{L}^2}. \tag{3.4.14}$$

(c) If $f \in \mathbb{C}_{\mathbb{R}}[0,\infty)$, then the solution φ_* of the renewal equation (3.4.13) also lies in $\mathbb{C}_{\mathbb{R}}[0,\infty)$ and the mapping that takes f to $\varphi_* \in \mathbb{C}_{\mathbb{R}}[0,\infty)$ is continuous.

PROOF. While existence and uniqueness of solutions to renewal equations under various conditions have been extensively studied (see, e.g., [**3**, Theorem 2.4, p. 146]), the particular version stated above appears not to be readily available in the literature. Hence, we provide the proof.

For part a., let $U = G^{*n}$, where G^{*n} denotes the n-fold convolution of G, be the renewal function associated to the distribution function G. Since G has density g, by [**3**, Proposition 2.7, Section V], U has density $u = U * g$, which satisfies the equation $u = g + g * u$. Moreover, since g is continuous (and hence locally bounded), u is also locally bounded. Define the function $\varphi_* \doteq f + u * f$. Then the local integrability of f and local boundedness of u imply the local integrability of φ_*, and, using the distributive and associative properties of the convolution operation, we have

$$g * \varphi_* = g * (f + u * f) = g * f + (g * u) * f = g * f + (u - g) * f = u * f = \varphi_* - f.$$

Therefore, φ_* is a solution of the equation $\varphi = f * \varphi$ given in (3.4.13).

To show that φ_* is the unique solution to (3.4.13) that lies in $\mathbb{L}^1_{\text{loc}}(0,\infty)$, let $\varphi_i \in \mathbb{L}^1_{\text{loc}}(0,\infty)$, $i = 1, 2$, be two solutions to (3.4.13). Then $\varphi = \varphi_1 - \varphi_2$ is a solution to the equation $\varphi = g * \varphi$. For $\varepsilon > 0$, let $\eta_\varepsilon(x) \doteq \mathbb{1}_{(0,\varepsilon)}(x)/\varepsilon$, and note that the function $\varphi_\varepsilon \doteq \varphi * \eta_\varepsilon$ satisfies

$$\varphi_\varepsilon = \varphi * \eta_\varepsilon = g * \varphi * \eta_\varepsilon = g * \varphi_\varepsilon.$$

Also, for every $T < \infty$,

$$|\varphi_\varepsilon(t)| \leq \frac{1}{\varepsilon} \int_0^T |\varphi(x)| dx < \infty, \qquad t \in [0, T],$$

where the finiteness holds since $\varphi \in \mathbb{L}^1_{\text{loc}}(0,\infty)$. Hence, φ_ε is locally bounded and satisfies the renewal equation $\varphi_\varepsilon = g * \varphi_\varepsilon$. However, by [**3**, Theorem 2.4, p. 146], this implies $\varphi_\varepsilon \equiv 0$. Since this holds for every $\varepsilon > 0$, this implies $\varphi \equiv 0$.

To see why part b. holds, first note that $f \in \mathbb{L}^2(0,\infty)$ implies $f \in \mathbb{L}^1_{\text{loc}}(0,\infty)$. Therefore, $\varphi_* = f + u * f$ is a solution of (3.4.13) from part a, and, moreover, it satisfies

$$|\varphi_*(t)| \leq |f(t)| + |u * f(t)| \leq |f(t)| + \left| \int_0^t f(t-s)(u(s) - 1) ds \right| + \left| \int_0^t f(s) ds \right|,$$

and hence, recalling the notation $\mathcal{I}_f(t) = \int_0^t f(s) ds$,

$$\|\varphi_*\|_{\mathbb{L}^2} \leq \|f\|_{\mathbb{L}^2} + \|f * (u - 1)\|_{\mathbb{L}^2} + \|\mathcal{I}_f\|_{\mathbb{L}^2}. \tag{3.4.15}$$

With some abuse of notation, we also let U denote the renewal measure associated with the distribution G, and let l^+ denote Lebesgue measure on $(0,\infty)$. Then, since $u-1$ is the density of the signed measure $U-l^+$ with respect to l^+ on $(0,\infty)$, we have

$$\int_0^\infty |u(s)-1|ds = ||U-l^+||_{TV},$$

where $\|\cdot\|_{TV}$ is the total variation norm. Since G has a finite second moment (and has mean 1), it follows from [**73**, eqn. (6.10), p. 86] (with $\lambda=1$) that $||U-l^+||_{TV}$ is finite and hence, $u-1\in\mathbb{L}^1(0,\infty)$. By Young's inequality, we then have

$$||f*(u-1)||_{\mathbb{L}^2} \leq ||u-1||_{\mathbb{L}^1}||f||_{\mathbb{L}^2}.$$

Substituting this inequality into (3.4.15), we obtain the bound (3.4.14) with $c_1 = 1+\|u-1||_{\mathbb{L}^1}<\infty$ and $c_2=1$.

Finally, to see c, note that since $f\in\mathbb{C}_{\mathbb{R}}[0,\infty)$ and u is bounded on finite intervals, $u*f$ also lies in $\mathbb{C}_{\mathbb{R}}[0,\infty)$, and hence, so does $\varphi_*=f+u*f$. Moreover, for every $f^1,f^2\in\mathbb{C}_{\mathbb{R}}[0,\infty)$ and corresponding solutions φ_*^1,φ_*^2, we have

$$\|\varphi_*^1-\varphi_*^2\|_T \leq (1+U(T))\,\|f^1-f^2\|_T, \qquad \forall T\geq 0,$$

and the continuity claim follows. □

COROLLARY 3.32. There exists a unique locally integrable process $\tilde{R}$ that satisfies equation (3.4.12).

PROOF. Almost surely, due to the continuity of $\overline{K}$ defined in (3.4.11), the continuity of $t\mapsto\mathcal{H}_t$ (see Remark 3.8), the fact that $\tilde{z}_0'\in\mathbb{L}^2(0,\infty)$, and the local integrability of g', the function

$$t\mapsto \tilde{z}_0'(t)-g(0)\overline{K}(t)-\int_0^t \overline{K}(s)g'(t-s)ds+\mathcal{H}_t(h)$$

lies in $\mathbb{L}^1_{\text{loc}}(0,\infty)$. The corollary then follows from Proposition 3.31.a. □

Next, for $t\geq 0$, define

$$\tilde{K}(t)\doteq\overline{K}(t)-\int_0^t\tilde{R}(s)ds. \qquad (3.4.16)$$

Substituting $\overline{K}(t)$ from above into (3.4.12) and simplifying terms, we obtain

$$\tilde{R}(t) = \tilde{z}_0'(t) + \mathcal{H}_t(h) - g(0)\tilde{K}(t) - \int_0^t \tilde{K}(u)g'(t-u)du, \tag{3.4.17}$$

which we observe is analogous to (3.4.6). Moreover, recall the definition (3.1.10) of the family of mappings $\{\Gamma_t; t \geq 0\}$, and in a fashion analogous to the definition of Z in (3.1.12), define

$$\tilde{Z}(t,r) \doteq \tilde{z}_0(t+r) - \mathcal{M}_t(\Psi_{t+r}\mathbf{1}) + \left(\Gamma_t\tilde{K}\right)(r), \qquad t, r \geq 0. \tag{3.4.18}$$

Since $\tilde{K}$ is almost surely continuous, it follows from (the proof of) Proposition 3.17.a that $\tilde{Z}$ is also a continuous $\mathbb{H}^1(0,\infty)$-valued process. Finally, set $\tilde{Y}(t) \doteq (\tilde{X}(t), \tilde{Z}(t))$, $t \geq 0$, and define $\Upsilon_{y,\tilde{y}}$ to be the joint distribution of $(Y, \tilde{Y})$.

3.4.3. Proof of the Measurability Property. In this section we prove the measurability condition (II) of Proposition 3.28 for the family of candidate equivalent couplings $\{\Upsilon_{y,\tilde{y}}, (y,\tilde{y}) \in \mathcal{Y} \times \mathcal{Y}\}$ that we constructed in the last section.

LEMMA 3.33. Suppose Assumptions II and III hold, and let A be the measurable subset of $\mathcal{Y}$ defined in (3.4.4). Then, for every $B \in \mathcal{B}(\mathcal{X})^{\mathbb{R}_+} \otimes \mathcal{B}(\mathcal{X})^{\mathbb{R}_+}$, the mapping $(y,\tilde{y}) \mapsto \Upsilon_{y,\tilde{y}}(B)$ from $A \times A$ to $\mathbb{R}$ is measurable.

PROOF. We claim that for every $t \geq 0$, almost surely, the mapping $(y,\tilde{y}) \mapsto (Y_y(t), \tilde{Y}_{\tilde{y}}(t))$ is continuous from $A \times A$ to $\mathcal{Y} \times \mathcal{Y}$. We first show that assuming the claim, the assertion of the lemma holds, and then we prove the claim.

Assuming that the claim holds, for every $k \geq 1$, $t_1, \ldots, t_k \geq 0$ and functions $\varphi_1, ..., \varphi_k \in \mathbb{C}_b(\mathcal{Y} \times \mathcal{Y}; \mathbb{R})$, by invoking the bounded convergence theorem we conclude that the mapping

$$(y,\tilde{y}) \mapsto \mathbb{E}\left[\varphi_1\left(Y_y(t_1), \tilde{Y}_{\tilde{y}}(t_1)\right) ... \varphi_k\left(Y_y(t_k), \tilde{Y}_{\tilde{y}}(t_k)\right)\right].$$

is continuous, and hence, measurable. The assertion in the lemma then follows from the definition of the Kolmogorov σ-algebra and a standard monotone class argument.

We now turn to the proof of the claim. Fix $t \geq 0$. We start by showing that the mapping from $y = (x_0, z_0) \in \mathcal{Y}$ to $Y_y(t)$ is continuous, almost surely. Fix $\omega \in \tilde{\Omega}$, where Ω^0 is a subset of Ω with $\mathbb{P}\{\Omega^0\} = 1$, on which Proposition 3.17 holds and the Brownian motion $\{B(t); t \geq 0\}$ has continuous paths. The map from $y = (x_0, z_0) \in \mathcal{Y}$ to $(E, x_0, z_0 - \mathcal{H}(\mathbf{1})) \in \mathbb{C}^0[0,\infty) \times \mathbb{R} \times \mathbb{C}_{\mathbb{R}}[0,\infty)$ is continuous due to Lemma 2.5.b. Since $(K, X) = \Lambda(E, x_0, z_0 - \mathcal{H}(\mathbf{1}))$ by definition (3.1.11), and the CMS mapping Λ is continuous by Lemma 2.13, it follows that the mapping $y \in \mathcal{Y} \mapsto (K, X) \in \mathbb{C}^0[0,\infty) \times \mathbb{C}_{\mathbb{R}}[0,\infty)$ is continuous. In particular, for fixed $t > 0$, the mapping $y \mapsto X(t)$ is continuous and, by the definition (3.1.12) of $Z(t,\cdot)$, continuity of the translation map (Lemma 2.5.c) and continuity of the mapping Γ_t (Lemma 3.16.b), the mapping $y \mapsto Z(t,\cdot) \in \mathbb{H}^1[0,\infty)$ is also continuous. Thus, we have shown that the mapping $y \in \mathcal{Y} \mapsto Y_y(t) \in \mathcal{Y}$ is continuous for every $\omega \in \tilde{\Omega}$ and $t \geq 0$.

Next, we prove continuity of the mapping $(y, \tilde{y}) = ((x_0, z_0), (\tilde{x}_0, \tilde{z}_0)) \in A \times A \mapsto \tilde{Y}_{\tilde{y}}(t) \in \mathcal{Y}$. It is clear from the equation (3.4.9) and the form of F specified below (3.4.9) that the map $(x_0, \tilde{x}_0, X) \in \mathbb{R}^2 \times \mathbb{C}_{\mathbb{R}}[0,\infty) \mapsto \tilde{X} \in \mathbb{C}_{\mathbb{R}}[0,\infty)$ is continuous. Together with the continuity of $y \mapsto X$ proved above, this implies that the map $(y, \tilde{y}) \in \mathcal{Y}^2 \mapsto \tilde{X} \in \mathbb{C}_{\mathbb{R}}[0,\infty)$ is continuous. Next, recall that by the definition of A, if $y = (x_0, z_0) \in A$ then $z_0(0) = x_0 \wedge 0 = 0$. Using this identity, equation (3.4.7) and equation (3.3.2) with $r = 0$, we have

$$\mathcal{I}_R(t) \doteq \int_0^t R(s)ds = z_0(t) + \mathcal{M}_t(\mathbf{1}) - \mathcal{M}_t(\Psi_t \mathbf{1}) - \int_0^t K(u)g(t-u)du.$$

When combined with the continuity of the mappings $z_0 \in \mathbb{H}^1(0,\infty) \mapsto z_0 \in \mathbb{C}_{\mathbb{R}}[0,\infty)$ (which holds by Lemma 2.5.b) and $y \mapsto K$ (established above), this implies that the mapping $(y, \tilde{y}) \in A \times A \mapsto \mathcal{I}_R \in \mathbb{C}_{\mathbb{R}}[0,\infty)$ is continuous. In turn, due to the continuity of the mappings that take $(y, \tilde{y})$ to X and $\tilde{X}$ established above, this implies that the map from $(y, \tilde{y}) \in A \times A$ to $\overline{K} \in \mathbb{C}_{\mathbb{R}}[0,\infty)$ defined in (3.4.11) is also continuous. Next, observe from (3.4.12) that $\tilde{R}$ satisfies the renewal equation

$\tilde{R} = \tilde{F} + g * \tilde{R}$ where

$$\tilde{F}(t) \doteq \tilde{z}_0'(t) - g(0)\overline{K}(t) - \int_0^t \overline{K}(s)g'(t-s)ds + \mathcal{H}_t(h), \qquad t \geq 0.$$

Therefore, $\mathcal{I}_{\tilde{R}}$ defined as $\mathcal{I}_{\tilde{R}}(t) = \int_0^t \tilde{R}(s)ds$, $t \geq 0$, satisfies the renewal equation $\mathcal{I}_{\tilde{R}} = \mathcal{I}_{\tilde{F}} + g * \mathcal{I}_{\tilde{R}}$, where, using the identity $\tilde{z}_0(0) = \tilde{x}_0 \wedge 0 = 0$ (which holds because $\tilde{y} \in A$), we obtain

$$\mathcal{I}_{\tilde{F}}(t) \doteq \int_0^t \tilde{F}(s)ds = \tilde{z}_0(t) - g * \overline{K}(t) + \int_0^t \mathcal{H}_s(h)ds, \qquad t \geq 0.$$

Therefore, $\mathcal{I}_{\tilde{F}}$ lies in $\mathbb{C}^0[0,\infty)$ and the map from $(y,\tilde{y})$ to $\mathcal{I}_{\tilde{F}}$ is continuous. An application of Proposition 3.31.c then shows that $\mathcal{I}_{\tilde{R}}$ also lies in $\mathbb{C}_{\mathbb{R}}[0,\infty)$, and that the map from $\mathcal{I}_{\tilde{F}}$ to $\mathcal{I}_{\tilde{R}}$ and hence, from $(y,\tilde{y})$ to $\mathcal{I}_{\tilde{R}}$, is continuous. Consequently, $\tilde{K} \in \mathbb{C}_{\mathbb{R}}[0,\infty)$ defined in (3.4.16) is also obtained as a continuous map of $(y,\tilde{y})$. Finally, by definition (3.4.18) of $\tilde{Z}(t,\cdot)$, continuity of the translation map (Lemma 2.5.c) and the continuity of the mapping Γ_t (Lemma 3.16.b), $\tilde{Z}(t,\cdot)$ can also be expressed as a continuous mapping of $(y,\tilde{y})$. Therefore, the mapping $(y,\tilde{y}) \in A \times A \mapsto \tilde{Y}_{\tilde{y}}(t) \in \mathcal{Y}$ is also continuous for every $\omega \in \Omega^0$. This completes the proof of the claim, and hence, the proof of the lemma. □

3.4.4. Asymptotic Convergence. In this section we prove condition (III) of Proposition 3.28, that is, the asymptotic convergence of the processes Y and $\tilde{Y}$ whenever their respective initial conditions $y = (x_0, z_0)$ and $\tilde{y} = (\tilde{x}_0, \tilde{z}_0)$ lie in the set A. Specifically, given the processes defined in Section 3.4.2, define $\Delta H \doteq H - \tilde{H}$ for $H = Y, y, Z, z_0, K, X, X^+, X^-, x_0$, and R. Our goal is to show that almost surely,

(3.4.19) $$\Delta Y(t) \to 0 \text{ in } \mathcal{Y}, \qquad \text{as } t \to \infty.$$

This will follow from (3.4.21) and Lemma 3.36 below.

Subtracting X from both sides of equation (3.4.8), we see that ΔX satisfies the integral equation

(3.4.20) $$\Delta X(t) = \Delta x_0 - \lambda \int_0^t \Delta X(s)ds,$$

whose solution is given by

$$\Delta X(t) = \Delta x_0 e^{-\lambda t}. \tag{3.4.21}$$

Clearly, $\Delta X(t)$ converges to zero with probability one as $t \to \infty$.

We now turn to the proof of the asymptotic convergence of ΔZ, which consists of two main steps. In the first step (see Lemma 3.34) we use the fact that ΔR satisfies a certain renewal equation to show that it is square integrable. In the second step (Lemma 3.36) we combine the square integrability of ΔR with other estimates to show that $\Delta Z(t,\cdot)$ converges to zero in $\mathbb{H}^1(0,\infty)$. First, note that on substituting the expression for $\overline{K}$ from (3.4.11) into the definition of $\tilde{K}$ in (3.4.16), subtracting this from the equation for K in (3.1.14), and recalling the definition of E from (3.1.8), we obtain

$$\Delta K(t) = -\Delta X^+(t) + \Delta x_0^+ - \int_0^t \Delta R(s)ds - \lambda \int_0^t \Delta X(s)ds. \tag{3.4.22}$$

When combined with (3.4.20) and the fact that $x_0, \tilde{x}_0 \geq 0$, (3.4.22) further simplifies to

$$\Delta K(t) = -\Delta X^-(t) - \int_0^t \Delta R(s)ds. \tag{3.4.23}$$

LEMMA 3.34. Suppose Assumptions II and III hold. Then $t \mapsto \Delta R(t) \in \mathbb{L}^2(0,\infty)$, almost surely. Moreover, there exist deterministic constants $\bar{C}_1, \bar{C}_2 < \infty$ such that

$$\|\Delta R\|_{\mathbb{L}^2} \leq \bar{C}_1 \|\Delta z_0\|_{\mathbb{H}^1} + \bar{C}_2 |\Delta x_0|. \tag{3.4.24}$$

PROOF. Subtracting (3.4.17) from (3.4.6), we obtain

$$\Delta R(t) = \Delta z_0'(t) - g(0)\Delta K(t) - \int_0^t \Delta K(s) g'(t-s)ds. \tag{3.4.25}$$

Substituting ΔK from (3.4.23) into (3.4.25) and using Fubini's theorem, we see that ΔR satisfies the renewal equation

$$\Delta R = \bar{F} + g * \Delta R, \tag{3.4.26}$$

with

$$\bar{F}(t) \doteq \Delta z_0'(t) + g(0)\Delta X^-(t) + (\Delta X^- * g')(t), \qquad t \geq 0. \tag{3.4.27}$$

We now claim that there exists $C < \infty$ such that

$$\max\left(\|\bar{F}\|_{\mathbb{L}^2}, \|\mathcal{I}_{\bar{F}}\|_{\mathbb{L}^2}\right) \leq \|\Delta z_0\|_{\mathbb{H}^1} + C|\Delta x_0| < \infty, \tag{3.4.28}$$

where $\mathcal{I}_{\bar{F}}(t) \doteq \int_0^t \bar{F}(s)\,ds$. We first show how the proof of the lemma follows from the claim. Indeed, since ΔR satisfies the renewal equation (3.4.26), G has a finite second moment by Assumption II.b, and $\bar{F}$ and $\mathcal{I}_{\bar{F}}$ lie in $\mathbb{L}^2(0,\infty)$ by (3.4.28), Proposition 3.31.b implies that there exist deterministic constants such that

$$\|\Delta R\|_{\mathbb{L}^2} \leq c_1\|\bar{F}\|_{\mathbb{L}^2} + c_2\|\mathcal{I}_{\bar{F}}\|_{\mathbb{L}^2}.$$

When combined with (3.4.28), this implies that ΔR lies in $\mathbb{L}^2(0,\infty)$ and satisfies (3.4.24) with $\bar{C}_1 \doteq c_1 + c_2$ and $\bar{C}_2 \doteq (c_1+c_2)C$, for $i = 1, 2$. It only remains to prove the claim (3.4.28). First, note that for every $t \geq 0$, using the fact that $\Delta z_0(0) = -\Delta x_0^- = 0$ because $y, \tilde{y} \in A$ implies $x_0^- = \tilde{x}_0^- = 0$, we have

$$\begin{aligned}
\mathcal{I}_{\bar{F}}(t) &= \int_0^t \Delta z_0'(s)ds + g(0)\int_0^t \Delta X^-(s)ds + \int_0^t \Delta X^- * g'(s)ds \\
&= \Delta z_0(t) + g(0)\int_0^t \Delta X^-(s)ds + \int_0^t \Delta X^-(s)\big(g(t-s) - g(0)\big)ds \\
&= \Delta z_0(t) + \Delta X^- * g(t). \qquad (3.4.29)
\end{aligned}$$

Moreover, by Young's inequality, Assumptions III and II.a, and (3.4.21), we have

$$\|\Delta X^- * g'\|_{\mathbb{L}^2} \leq \|g'\|_{\mathbb{L}^1}\,\|\Delta X^-\|_{\mathbb{L}^2} \leq H_2\|\Delta X^-\|_{\mathbb{L}^2} < \infty. \tag{3.4.30}$$

Together with (3.4.27), (3.4.29) and Minkowski's integral inequality, this implies that

$$\begin{aligned}
\|\bar{F}\|_{\mathbb{L}^2} &\leq \|\Delta z_0\|_{\mathbb{H}^1} + (g(0) + H_2)\|\Delta X^-\|_{\mathbb{L}^2}, \\
\|\mathcal{I}_{\bar{F}}\|_{\mathbb{L}^2} &\leq \|\Delta z_0\|_{\mathbb{H}^1} + H_2\|\Delta X^-\|_{\mathbb{L}^2}.
\end{aligned}$$

Since $\|\Delta X^-\|_{\mathbb{L}^2} \leq \frac{1}{\sqrt{2\lambda}}|\Delta x_0|$ due to (3.4.21), we conclude that (3.4.28) holds with $C \doteq (g(0) + H_2)/\sqrt{2\lambda}$. This completes the proof of the lemma. □

In Lemma 3.36 below, we use (3.4.24) to prove the asymptotic convergence of $\Delta Z(t,\cdot)$. The proof of Lemma 3.36 makes use of the following elementary result on the convolution operator.

LEMMA 3.35. Let $v \in \mathbb{L}^1(0,\infty)$, and suppose $w \in \mathbb{L}^1_{\text{loc}}(0,\infty)$ is bounded and $w(t) \to 0$ as $t \to \infty$. Then $\varphi \doteq v * w$ satisfies $\lim_{t\to\infty} \varphi(t) = 0$.

PROOF. Fix $\varepsilon > 0$ and choose $t_0 < \infty$ such that $\sup_{t \geq t_0} |w(t)| \leq \varepsilon$. Then for $t \geq t_0$,

$$\phi(t) = \int_0^{t_0} w(s)v(t-s)ds + \int_{t_0}^{t} w(s)v(t-s)ds,$$

which implies that

$$|\phi(t)| \leq ||w||_\infty \int_{t-t_0}^{t} |v(s)|ds + \varepsilon \|v\|_{\mathbb{L}^1}.$$

Sending $t \to \infty$ in the last display, and noting that $\int_{t-t_0}^{t} |v(s)|ds \to 0$ because $v \in \mathbb{L}^1(0,\infty)$, it follows that $\limsup_{t\to\infty} |\phi(t)| \leq \varepsilon\|v\|_{\mathbb{L}^1}$. Since $\varepsilon > 0$ is arbitrary, the lemma follows on sending $\varepsilon \downarrow 0$. $\square$

LEMMA 3.36. Under Assumptions II and III, almost surely, $\Delta Z(t,\cdot) \to 0$ in $\mathbb{H}^1(0,\infty)$ as $t \to \infty$.

PROOF. Subtracting (3.4.18) from (3.1.12), we see that

$$\Delta Z(t,r) = \Delta z_0(t+r) + (\Gamma_t \Delta K)(r).$$

Using (3.1.10) and (3.4.23) to expand the right-hand side above, we have

$$\Delta Z(t,r) = \Delta z_0(t+r) - \overline{G}(r)\Delta X^-(t) + \zeta_t(r) + \xi_t(r), \tag{3.4.31}$$

where $\zeta_t(r) \doteq \int_0^t \Delta X^-(s) g(t+r-s)ds$ and $\xi_t(r) \doteq -\int_0^t \Delta R(s)\overline{G}(t+r-s)ds$. To prove the lemma, it suffices to show that almost surely, the $\mathbb{H}^1(0,\infty)$ norm of each of the terms on the right-hand side of (3.4.31) goes to zero as $t \to \infty$. For the first term, we have

$$\|\Delta z_0(t+\cdot)\|_{\mathbb{H}^1} = \|\Delta z_0(t+\cdot)\|_{\mathbb{L}^2} + \|\Delta z_0'(t+\cdot)\|_{\mathbb{L}^2},$$

which converges to zero as $t \to \infty$ by (2.2.2) of Lemma 2.5. For the second term, by (3.4.21) we have

$$\|\overline{G}(\cdot)\Delta X^-(t)\|_{\mathbb{H}^1} = \|\overline{G}\|_{\mathbb{H}^1}|\Delta X^-(t)| \leq \|\overline{G}\|_{\mathbb{H}^1}|\Delta x_0|e^{-\lambda t},$$

which converges to zero because $\|\overline{G}\|_{\mathbb{H}^1}$ is finite by Assumptions II and III (see Remark 2.3).

Next, since ΔX^- is continuous and g' is bounded and continuous by Assumption III, by the bounded convergence theorem, ζ_t is absolutely continuous with density

$$\zeta_t'(r) = \int_0^t \Delta X^-(s)g'(t+r-s)ds = \int_0^t \Delta X^-(t-s)g'(r+s)ds, \quad r \in (0,\infty).$$

Therefore, applying Hölder's inequality and Tonelli's theorem, we see that

$$\begin{aligned}
\|\zeta_t\|_{\mathbb{L}^2}^2 &= \int_0^\infty \left(\int_0^t \Delta X^-(t-s)g(r+s)ds\right)^2 dr \\
&\leq \int_0^\infty \left(\int_0^t \Delta X^-(t-s)^2 g(r+s)ds\right)\left(\int_0^t g(r+s)ds\right) dr \\
&\leq \int_0^t \Delta X^-(t-s)^2 \int_0^\infty g(r+s)dr\, ds \\
&= \left((\Delta X^-)^2 * \overline{G}\right)(t),
\end{aligned}$$

and, likewise, we have

$$\begin{aligned}
\|\zeta_t'\|_{\mathbb{L}^2}^2 &\leq \int_0^\infty \left(\int_0^t \Delta X^-(t-s)|g'(r+s)|ds\right)^2 dr \\
&\leq H_2^2 \int_0^\infty \left(\int_0^t \Delta X^-(t-s)\overline{G}(r+s)ds\right)^2 dr \\
&\leq H_2^2 \int_0^t \Delta X^-(t-s)^2 \int_0^\infty \overline{G}(r+s)dr\, ds \\
&= H_2^2 \left((\Delta X^-)^2 * \int_\cdot^\infty \overline{G}(r)dr\right)(t).
\end{aligned}$$

Now, $(\Delta X^-)^2$ is integrable by (3.4.21) and both $\overline{G}$ and $\int_\cdot^\infty \overline{G}(s)ds$ are continuous, bounded by 1, lie in $\mathbb{L}^1(0,\infty)$ and vanish at infinity. Thus, it follows from Lemma 3.35 that as $t \to \infty$, $\|\zeta_t\|_{\mathbb{L}^2} \to 0$ and $\|\zeta_t'\|_{\mathbb{L}^2} \to 0$, and consequently, $\|\zeta_t\|_{\mathbb{H}^1} \to 0$.

Similarly, since ΔR is continuous and $\overline{G}$ has a continuous and bounded density g due to Assumption III, by the bounded convergence theorem, $\tilde{\zeta}$ has density

$$\tilde{\zeta}_t'(r) = \int_0^t \Delta R(s) g(r+t-s) ds.$$

An exactly analogous analysis then shows that

$$\|\tilde{\zeta}_t\|_{\mathbb{L}^2}^2 \leq \left(\Delta R^2 * \overline{G}\right)(t),$$

and

$$\|\tilde{\zeta}_t'\|_{\mathbb{L}^2}^2 \leq H\left(\Delta R^2 * \int_\cdot^\infty \overline{G}(r)dr\right)(t).$$

Again, since ΔR^2 is integrable by Lemma 3.34 and both $\bar{G}$ and $\int_\cdot^\infty \overline{G}(s)ds$ are integrable, bounded by 1 and vanish at infinity, another application of Lemma 3.35 shows that as $t \to \infty$, $\|\zeta_t\|_{\mathbb{L}^2} \to 0$ and $\|\zeta_t'\|_{\mathbb{L}^2} \to 0$, and consequently, $\|\zeta_t\|_{\mathbb{H}^1} \to 0$. This completes the proof. □

3.4.5. Marginal Distributions. In this section, we continue to use the ΔH notation introduced in Section 3.4.4. The main result of this section is as follows:

PROPOSITION 3.37. If Assumptions II and III hold, then Y has distribution $\mathcal{P}_{[0,\infty)}\delta_y$, and the distribution of $\tilde{Y}$ on $\mathcal{Y}^{\mathbb{R}_+}$ is equivalent to $\mathcal{P}_{[0,\infty)}\delta_{\tilde{y}}$.

PROOF. Y has distribution $\mathcal{P}_{[0,\infty)}\delta_y$ by definition. To show that the distribution of $\tilde{Y}$ is equivalent to $\mathcal{P}_{[0,\infty)}\delta_{\tilde{y}}$, we first show that $\tilde{Y}$ satisfies the same equations as Y, but with B replaced by some process $\tilde{B}$, and then invoke Girsanov's theorem to prove that $\tilde{B}$ is a Brownian motion on the entire time interval $[0,\infty)$ under another probability measure $\tilde{\mathbb{P}}$ that is equivalent to $\mathbb{P}$. Let $\tilde{B}_t \doteq B_t - \int_0^t m(s)ds$, where m is defined by

$$m(s) \doteq -R(s) + \tilde{R}(s) - \lambda\big(X(s) + \tilde{X}(s)\big) = -\Delta R(s) - \lambda \Delta X(s), \tag{3.4.32}$$

and set $\tilde{E} \doteq \tilde{B}(t) - \beta t$. We will first show that

$$(\tilde{K}, \tilde{X}) = \Lambda(\tilde{E}, \tilde{x}_0, \tilde{z}_0 - \mathcal{H}_t(\mathbf{1})), \tag{3.4.33}$$

where Λ is the CMS mapping introduced in Definition 2.12. To prove (3.4.33), first note that the expression (3.4.10) for $\tilde{X}$ can be rewritten as

$$\tilde{X}(t) = \tilde{x}_0 + \tilde{E}(t) - \mathcal{M}_t(\mathbf{1}) + \int_0^t \tilde{R}(s)ds. \tag{3.4.34}$$

By equation (3.4.17) for $\tilde{R}$, relation (3.2.12), and Fubini's theorem, we have

$$\begin{aligned}\int_0^t \tilde{R}(s)ds &= \int_0^t \tilde{z}_0'(s)ds + \int_0^t \mathcal{H}_s(h)ds - g(0)\int_0^t \tilde{K}(s)ds \\ &\quad - \int_0^t \left(\int_0^s \tilde{K}(v)g'(s-v)dv\right)ds \\ &= \tilde{z}_0(t) - \tilde{z}_0(0) - \mathcal{H}_t(\mathbf{1}) + \mathcal{M}_t(\mathbf{1}) - \int_0^t \tilde{K}(s)g(t-s)ds.\end{aligned}$$

Together with (3.4.34), this implies

$$\tilde{X}(t) = \tilde{x}_0 + \tilde{E}(t) + \tilde{V}(t) - \tilde{V}(0) - \tilde{K}(t), \tag{3.4.35}$$

where

$$\tilde{V}(t) \doteq \tilde{z}_0(t) - \mathcal{H}_t(\mathbf{1}) + \tilde{K}(t) - \int_0^t \tilde{K}(s)g(t-s)ds. \tag{3.4.36}$$

Also, by definitions (3.4.11) and (3.4.16) of $\bar{K}$ and $\tilde{K}$ and equation (3.4.32), we have

$$\tilde{K}(t) = \bar{K}(t) - \int_0^t \tilde{R}(s)ds = \tilde{E}(t) - \tilde{X}^+(t) + \tilde{x}_0^+. \tag{3.4.37}$$

Substituting $\tilde{K}$ from (3.4.37) in (3.4.35) and noting from (3.4.36) and the fact that $(\tilde{x}_0, \tilde{z}_0) \in \mathcal{Y}$, $\tilde{V}(0) = \tilde{z}_0(0) = -\tilde{x}_0^-$, we conclude that

$$\tilde{V}(t) = \tilde{X}(t) \wedge 0 = -\tilde{X}(t)^-. \tag{3.4.38}$$

Comparing the equation obtained on substituting $\tilde{V}$ from (3.4.38) into (3.4.37) with the CMS equation (3.1.13), and comparing (3.4.37) with the CMS equation (3.1.14), it follows that (3.4.33) holds. Furthermore, $\tilde{Z}$ defined in (3.4.18) has the same form as the expression (3.1.12) for Z, but with K replaced by $\tilde{K}$. In summary, we have shown that $\tilde{Y}$ is defined in the same way as the diffusion limit $Y_{\tilde{y}}$ in Definition 3.3, except that B is replaced by $\tilde{B}$. Therefore, to complete the proof, it suffices to show that there exists a new probability measure $\tilde{\mathbb{P}}$ on $(\Omega, \mathcal{F})$ that is equivalent to $\mathbb{P}$ and under $\tilde{\mathbb{P}}$, $\tilde{B}$ is a Brownian motion independent of $\mathcal{M}$.

Define the process $\{N_t; t \geq 0\}$ as

$$N_t \doteq \exp\left(\int_0^t m(s)dB_s - \frac{1}{2}\int_0^t m(s)^2 ds\right), \qquad t \geq 0, \tag{3.4.39}$$

where $m = -(\Delta R + \lambda \Delta X)$ is as above. Since N is a uniformly integrable exponential martingale by Lemma 3.38 below, by Doob's convergence theorem N_t converges almost surely as $t \to \infty$ to an integrable random variable N_∞, which is almost surely positive. Define a probability measure $\tilde{\mathbb{P}}$ on $(\Omega, \mathcal{F})$ by:

$$\tilde{\mathbb{P}}(A) = \mathbb{E}\left[\mathbb{1}_A N_\infty\right], \qquad A \in \mathcal{F}, \tag{3.4.40}$$

where $\mathbb{E}$ denotes expectation with respect to $\mathbb{P}$. Since η_∞ is almost surely positive, $\tilde{\mathbb{P}}$ is equivalent to $\mathbb{P}$. Also, by Girsanov's theorem (see, e.g., [**90**, Theorem (38.5), Chapter IV]), $\{\tilde{B}(t) = B(t) - \int_0^t m(s)ds, t \geq 0\}$ is a Brownian motion under $\tilde{\mathbb{P}}$. Moreover, for every $A_1, A_2 \in \mathcal{B}[0,\infty)$, since under $\mathbb{P}$, $\mathcal{M}(A_i)$, $i = 1,2$, is a martingale independent of B, $\langle \mathcal{M}(A_i), B\rangle_t \equiv 0$, and by [**59**, Proposition 5.4, Chapter 3] (note that since N is a martingale, for every $T > 0$ and $A \in \mathcal{F}_T$, $\tilde{\mathbb{P}}(A) = \mathbb{E}\left[\mathbb{1}_A N_T\right]$, and hence, our definition of $\tilde{\mathbb{P}}$ is compatible with its definition (5.4) in [**59**]) under $\tilde{\mathbb{P}}$, $\mathcal{M}(A_i)$ is a martingale, $\langle \mathcal{M}(A_1), \mathcal{M}(A_2)\rangle_t = t\int_0^\infty \mathbb{1}_{A_1 \cap A_2}(x)\, g(x)dx$ and $\langle \tilde{B}, \mathcal{M}(A_i)\rangle \equiv 0$ for $i = 1,2$. Therefore, under $\tilde{\mathbb{P}}$, $\mathcal{M}$ is a martingale measure with covariance function given in (3.1.1) and is independent of $\tilde{B}$. This completes the proof. □

LEMMA 3.38. Suppose Assumptions II and III are satisfied. Then, the process $\{N_t; t \geq 0\}$ defined in (3.4.39) is a uniformly integrable martingale.

PROOF. Invoking Proposition 2 and Theorem 6 in [**64**], it suffices to show that the local martingale $M(t) = \int_0^t m(s)dB_s$ is of bounded mean oscillation (BMO), that is, there exists a constant $\bar{C} < \infty$ such that for all $\mathcal{F}_t$-stopping times T,

$$\mathbb{E}[\langle M\rangle_\infty - \langle M\rangle_T | \mathcal{F}_T] \leq \bar{C}.$$

Since $\langle M\rangle_t = \int_0^t m(s)^2 ds \geq 0$ for all $t \geq 0$, M is BMO if $\langle M\rangle_\infty \leq \bar{C}$ a.s. for some $\bar{C} < \infty$. Note that

$$\langle M\rangle_\infty = \int_0^\infty m^2(s)ds = ||m||^2_{\mathbb{L}^2},$$

and by (3.4.32), (3.4.24) and (3.4.21), there exist deterministic constants $\bar{C}_1, \bar{C}_2 < \infty$ such that

$$\|m\|^2_{\mathbb{L}^2} \leq 2\|\Delta R\|^2_{\mathbb{L}^2} + 2\lambda^2\|\Delta X\|^2_{\mathbb{L}^2} \leq \bar{C} \doteq 4\bar{C}_1^2\|z_0\|^2_{\mathbb{H}^1} + (4\bar{C}_2^2 + \lambda)|\Delta x_0|^2 < \infty.$$

This completes the proof. □

3.4.6. Proof of Theorem 3.26.

PROOF OF THEOREM 3.26. Proposition 3.25 shows that the diffusion limit Y is a time-homogeneous Feller Markov process on the Polish space $\mathcal{Y}$. Let $\mathcal{P} = \{\mathcal{P}_t, t \geq 0\}$ be the associated semigroup of Markov kernels on $\mathcal{Y}$. In order to show that the diffusion limit has at most one invariant distribution, it suffices to show that the candidate coupling $\{\Upsilon_{y,\tilde{y}}, (y,\tilde{y}) \in A^2\}$ constructed in Section 3.4.2 satisfies the conditions of Proposition 3.28. Recall that $A = \{(x,z) \in \mathcal{Y}\,; x \geq 0\}$. For every $y, \tilde{y} \in A$, it follows from Lemma 3.33 that the mapping $(y,\tilde{y}) \mapsto \Upsilon_{y,\tilde{y}}(B)$ is measurable for every $B \in \mathcal{B}(\mathcal{X})^{\mathbb{R}_+} \otimes \mathcal{B}(\mathcal{X})^{\mathbb{R}_+}$, from Proposition 3.37 that $\Upsilon_{y,\tilde{y}} \in \tilde{\mathcal{C}}(\mathcal{P}_{[0,\infty)}\delta_y, \mathcal{P}_{[0,\infty)}\delta_{\tilde{y}})$, and from (3.4.19) that that $\Upsilon_{y,\tilde{y}}(\mathcal{D}) = 1$. So, to complete the proof of the theorem, it suffices to show that the subset A satisfies the first condition of Proposition 3.28.

Let μ be an invariant distribution of $\mathcal{P}$. Assume to the contrary that $\mu(A) = 0$. Let $\mathbb{P}_\mu$ be the conditional probability measure given $Y(0) \sim \mu$. Since μ is invariant for $\{\mathcal{P}_t\}$, for every $t \geq 0$, $Y(t) \sim \mu$ under $\mathbb{P}_\mu$, and therefore, $\mathbb{P}_\mu\{Y(t) \in A\} = \mathbb{P}_\mu\{X(t) \geq 0\} = 0$. Equivalently, we have $\mathbb{P}_\mu\{X(t) < 0\} = 1$ for all $t \geq 0$. Since X is continuous $\mathbb{P}_\mu$-almost surely, this implies that

$$\mathbb{P}_\mu\{X(t) \leq 0 \text{ for every } t \geq 0\} = 1. \tag{3.4.41}$$

By (3.1.13), (3.1.14) and (3.4.41) we have, $\mathbb{P}_\mu$-almost surely, for every $t \geq 0$, $K(t) = B(t) - \beta t - X^+(t) + X^+(0) = B(t) - \beta t$, and

$$
\begin{aligned}
X(t) = X(t) \wedge 0 = Z_0(t) + B(t) - \beta t - \mathcal{H}_t(\mathbf{1}) - \int_0^t g(t-s)(B(s) - \beta s) ds \\
= Z_0(t) - \beta t + \beta \int_0^t s g(t-s) ds + \int_0^t \overline{G}(t-s) dB_s - \mathcal{H}_t(\mathbf{1}).
\end{aligned}
$$

For every $t > 0$, $\int_0^t \overline{G}(t-s) dB_s$ and $-\mathcal{H}_t(\mathbf{1})$ are two independent finite variance Gaussian random variables that are independent of $Z_0(\cdot)$, and $c_t \doteq \beta t - \beta \int_0^t s g(t-s) ds$ is a finite deterministic constant. Hence, for any $z_0 \in \mathbb{H}^1(0, \infty)$ and $t > 0$,

$$
\mathbb{P}\left\{ \int_0^t \overline{G}(t-s) dB_s - \mathcal{H}_t(\mathbf{1}) > -z_0(t) + c_t \right\} > 0.
$$

Therefore,

$$
\mathbb{P}_\mu\{X(t) > 0\} = \int_{y=(x_0,z_0)\in\mathcal{Y}} \mathbb{P}\left\{ \int_0^t \overline{G}(t-s) dB_s - \mathcal{H}_t(\mathbf{1}) > -z_0(t) + c_t \right\} \mu(dy) > 0,
$$

which contradicts (3.4.41). Therefore, $\mu(A) > 0$, and the proof is complete. $\square$

3.A. A Continuous Version of the Asymptotic Coupling Theorem

In this section, for completeness, we prove the continuous version of Corollary 2.2 of [**50**], as stated in Proposition 3.28. The proof is a straightforward adaptation of the proof of [**50**, Theorem 1.1 and Corollary 2.2] to the continuous time setting (see also [**30**, Theorem 2] and [**79**, Lemma 8.5], where a continuous version is used). In what follows, recall that for any semi-group of measurable operators $\phi^s : (\mathcal{X}^{\mathbb{R}_+}, \mathcal{B}(\mathcal{X})^{\mathbb{R}_+}) \mapsto (\mathcal{X}^{\mathbb{R}_+}, \mathcal{B}(\mathcal{X})^{\mathbb{R}_+})$, $s \geq 0$, on the probability space $(\mathcal{X}^{\mathbb{R}_+}, \mathcal{B}(\mathcal{X})^{\mathbb{R}_+})$, a probability measure μ on $(\mathcal{X}^{\mathbb{R}_+}, \mathcal{B}(\mathcal{X})^{\mathbb{R}_+})$ is said to be an invariant measure of $\{\phi^s\}$ if $\phi^s_{\#}\mu = \mu$ for all $s \geq 0$. Also, a set $A \in \mathcal{B}(\mathcal{X})^{\mathbb{R}_+}$ is said to be an invariant set of $\{\phi^s\}$ if $(\phi^s)^{-1}(A) = A$ for all $s \geq 0$. Finally, an invariant measure μ of $\{\phi^s\}$ is said to be ergodic if $\mu(A) \in \{0, 1\}$ for every invariant set A of $\{\phi^s\}$.

PROOF OF PROPOSITION 3.28. Define the family $\{\Theta_s; s \geq 0\}$ of shift operators Θ_s as $\Theta_s x(\cdot) \doteq x(s+\cdot), x \in \mathcal{X}^{\mathbb{R}_+}$, and note that it forms a semigroup of measurable operators from $(\mathcal{X}^{\mathbb{R}_+}, \mathcal{B}(\mathcal{X})^{\mathbb{R}_+})$ to itself. Let μ_1 and μ_2 be two invariant distributions of $\{\mathcal{P}_t\}$. Given the ergodic decomposition of invariant measures, we can assume without loss of generality that μ_1 and μ_2 are both ergodic invariant distributions of $\{\mathcal{P}_t\}$, and hence $m_i \doteq \mathcal{P}_{[0,\infty)}\mu_i$, $i = 1, 2$, are ergodic invariant distributions for the semigroup $\{\Theta_s\}$.

First, we extend the definition of the measurable map Υ in the statement of the proposition from $A \times A$ to the whole space $\mathcal{X} \times \mathcal{X}$ by setting $\Upsilon_{y,\tilde{y}} = \mathcal{P}_{[0,\infty)}\delta_y \times \mathcal{P}_{[0,\infty)}\delta_{\tilde{y}}$ for $(y,\tilde{y}) \notin A \times A$. Next, let $\bar{\Upsilon} \in \mathbb{M}_1(\mathcal{X}^{\mathbb{R}_+} \times \mathcal{X}^{\mathbb{R}_+})$ be the measure given by

$$\bar{\Upsilon}(B) \doteq \int_{\mathcal{X}\times\mathcal{X}} \Upsilon_{y,\tilde{y}}(B)\mu_1(dy)\mu_2(d\tilde{y}), \qquad B \in \mathcal{B}(\mathcal{X})^{\mathbb{R}_+} \otimes \mathcal{B}(\mathcal{X})^{\mathbb{R}_+}.$$

Note that by construction, $\bar{\Upsilon} \in \tilde{\mathcal{C}}(m_1, m_2)$. Also, since by conditions (I) and (III) of the proposition, $\bar{\Upsilon}_{y,\tilde{y}}(\mathcal{D}) > 0$ for all $(y,\tilde{y}) \in A \times A$ and $\mu_1(A), \mu_2(A) > 0$, it follows that $\bar{\Upsilon}(\mathcal{D}) > 0$.

Now, fix any bounded, Lipschitz function ϕ on $\mathcal{X}$, and define $\tilde{\phi} : \mathcal{X}^{\mathbb{R}_+} \mapsto \mathbb{R}$ by $\tilde{\phi}(x) \doteq \phi(x(0)), x \in \mathcal{X}^{\mathbb{R}_+}$. By (the continuous-time version of) Birkhoff's ergodic theorem (see, e.g., [**68**, Theorem 1 of Section 1.2]), for $i = 1, 2$, there exist sets $B_i^\phi \in \mathcal{B}(\mathcal{X})^{\mathbb{R}_+}$ with $m_i(B_i^\phi) = 1$ such that for every $x_i \in B_i^\phi$,

(3.A.1)
$$\lim_{t\to\infty} \frac{1}{t}\int_0^t \phi(x_i(s))ds = \lim_{t\to\infty} \frac{1}{t}\int_0^t \tilde{\phi}(\Theta_s x_i)ds = \int_{\mathcal{X}^{\mathbb{R}_+}} \tilde{\phi}(x_i)m_i(dx) = \int_{\mathcal{X}} \phi(z)\mu_i(dz).$$

Now, since $\bar{\Upsilon} \in \tilde{\mathcal{C}}(m_1, m_2)$, for $i = 1, 2$, the marginal $\Pi_{\#}^{(i)}\bar{\Upsilon}$ is absolutely continuous with respect to m_i, which in turn implies $\Pi_{\#}^{(i)}\bar{\Upsilon}(B_i^\phi) = 1$ because $m_i(B_i^\phi) = 1$. Thus, we have $\bar{\Upsilon}(B_1^\phi \times B_2^\phi) = 1$. Also, since $\bar{\Upsilon}(\mathcal{D}) > 0$, defining $\bar{\mathcal{D}} \doteq \mathcal{D} \cap (B_1^\phi \times B_2^\phi)$, we have $\bar{\Upsilon}(\bar{\mathcal{D}}) > 0$, and in particular, $\bar{\mathcal{D}}$ is not empty. Take any $(x_1, x_2) \in \bar{\mathcal{D}}$. Then

$(x_1, x_2) \in \mathcal{D}$ and by (3.A.1), we conclude that

$$\begin{aligned}\left|\int_{\mathcal{X}} \phi(z)\mu_1(dz) - \int_{\mathcal{X}} \phi(z)\mu_2(dz)\right| &= \lim_{t\to\infty} \frac{1}{t}\left|\int_0^t (\phi(x_1(s)) - \phi(x_2(s)))\, ds\right| \\ &\le \lim_{t\to\infty} \frac{C_\phi}{t}\int_0^t d((x_1(s)), x_2(s))ds \\ &= 0,\end{aligned}$$

where C_ϕ is the Lipschitz constant of ϕ and the last equality follows from the definition (3.4.3) of $\mathcal{D}$. Therefore, $\int_{\mathcal{X}} \phi(z)\mu_1(dz) = \int_{\mathcal{X}} \phi(z)\mu_2(dz)$ for every bounded Lipschitz function ϕ, and hence, $\mu_1 = \mu_2$. $\square$

3.B. On the Generator of the Limit Markov Process

In this section, we formally discuss the problem of identification of the generator of the limit Markov process Y, as the first step towards the characterization of the unique invariant distribution π. Some of the results here *not rigorous* yet; a complete treatment of this problem is the subject of future work.

Let $\mathbb{F}$ be the collection of functions $\tilde{f} : \mathbb{R} \times \mathbb{H}^1(0,\infty) \mapsto \mathbb{R}$ for which there exist $n \ge 0, r_1, ..., r_n \in [0,\infty)$ and $f \in \mathbb{C}_c^\infty(\mathbb{R} \times \mathbb{R}^n; \mathbb{R})$ such that

$$\tilde{f}(x,z) = f(x, z(r_1), ..., z(r_n)), \qquad x \in \mathbb{R}, z \in \mathbb{H}^1(0,\infty).$$

Fix $\tilde{f} \in \mathbb{F}$, with corresponding $n \ge 0$, $r_1, ..., r_n$ and f. For $z \in \mathbb{H}^1(0,\infty)$, denote $z^n = (z(r_1), ..., z(r_n))$ and let f_x and f_j be the partial derivative of $f(x, z_1, ..., z_n)$ with respect to x and z_j, respectively . By Ito's formula,

$$\begin{aligned}d\tilde{f}(X_t, Z_t) =& f_x(X_t, Z_t^n)dX_t + \sum_{j=1}^n f_j(X_t, Z_t^n)dZ_t(r_j) \\ &+ \frac{1}{2} f_{xx}(X_t, Z_t^n)d\langle X\rangle_t \\ &+ \sum_{j=1}^n f_{xj}(X_t, Z_t^n)d\langle X, Z(r_j)\rangle_t \\ &+ \frac{1}{2}\sum_{i,j=1}^n f_{ij}(X_t, Z_t^n)d\langle Z(r_i), Z(r_j)\rangle_t.\end{aligned}$$

It can be shown that when the initial condition $y_0 = (x_0, z_0)$ of the SPDE of Definition (3.1) is such that $z_0 \in \mathbb{H}^1(0,\infty) \cap \mathbb{C}^1[0,\infty)$, then the solution $Y^{y_0} = (X, Z)$ satisfies $Z_t \in \mathbb{H}^1(0,\infty) \cap \mathbb{C}^1[0,\infty)$, almost surely, for every $t \geq 0$. For such initial conditions, we have

$$dX_t = \big(Z_t'(0) - \beta\big)dt + dB_t - d\mathcal{M}_t(\mathbf{1}), \tag{3.B.1}$$

$$dZ_t(r_j) = Z_t'(r_j)dt - d\mathcal{M}_t(\Phi_{r_j}\mathbf{1}) + \bar{G}(r_j)dK_t, \tag{3.B.2}$$

Also, by Theorem 4 in **[62]**,

$$dK_t = \mathbb{1}_{\{X(t)>0\}}\{d\mathcal{M}_t(\mathbf{1}) - Z_t'(0)dt\} + \mathbb{1}_{\{X(t)\leq 0\}}\{dB_t - \beta dt\} + \frac{1}{2}dL_0^X(t),$$

where L_0^X is the local time of X at 0, and hence,

$$\begin{aligned} dZ_t(r_j) =& \mathbb{1}_{\{X(t)>0\}}\Big\{\big(Z_t'(r_j) - \bar{G}(r_j)Z_t'(0)\big)dt - d\mathcal{M}(\Phi_{r_j}\mathbf{1}) + \bar{G}(r_j)d\mathcal{M}_t(\mathbf{1})\Big\} \\ &+ \mathbb{1}_{\{X(t)\leq 0\}}\Big\{\big(Z_t'(r_j) - \bar{G}(r_j)\beta\big)\,dt - d\mathcal{M}_t(\Phi_{r_j}\mathbf{1}) + \bar{G}(r_j)dB_t\Big\} \\ &+ \frac{1}{2}\bar{G}(r_j)dL_0^X(t). \end{aligned}$$

Using the terminology of **[62]**, we first consider simpler stochastic process, which we call super-critical and sub-critical processes, that appear as the diffusion limits of the state representation $\widehat{Y}^{(N)}$, but in the so-called super-critical and sub-critical asymptotic regimes, respectively.

We will frequently use the following equality on the covariance function of the martingale measure term

$$\langle \mathcal{M}(\Phi_{r_i}\mathbf{1}), \mathcal{M}(\Phi_{r_j}\mathbf{1})\rangle_t = F(r_i, r_j)t,$$

where

$$F(r_i, r_j) \doteq \int_0^\infty \frac{\bar{G}(x+r_i)\bar{G}(x+r_j)}{\bar{G}(x)^2}g(x)dx. \tag{3.B.3}$$

We first compute the generator for sub- and super- critical cases, and then show that the generator for critical case equals to that of sub-critical case on the set $\{(x,z) \in \mathbb{R} \times \mathbb{H}^1(0,\infty); x < 0\}$ (which we call sub-critical region) and equals to the generator of super-critical case on the set $\{(x,z) \in \mathbb{R} \times \mathbb{H}^1(0,\infty); x > 0\}$ (which we call super-critical region.)

3.B.1. Generator of the Sub-Critical Process. First consider the sub-critical process $Y = (X, Z)$ that satisfies (3.B.1) with $dK_t = dB_t - \beta dt$. Hence,

$$dZ_t(r_j) = \left(Z_t'(r_j) - \bar{G}(r_j)\beta\right) dt - d\mathcal{M}_t(\Phi_{r_j}\mathbf{1}) + \bar{G}(r_j) dB_t.$$

Therefore,

$$\begin{aligned} d\langle X\rangle_t &= 2dt, \\ d\langle X, Z(r_j)\rangle_t &= \left(F(0,r_j) + \bar{G}(r_j)\right) dt = c(0,j)dt \\ d\langle Z(r_i), Z(r_j)\rangle_t &= \left(F(r_i,r_j) + \bar{G}(r_i)\bar{G}(r_j)\right) dt = c(i,j)dt, \end{aligned}$$

where

$$c(i,j) = F(r_i,r_j) + \bar{G}(r_i)\bar{G}(r_j), \tag{3.B.4}$$

and $r_0 = 0$ by convention. Therefore, for every $(x,z) \in \mathbb{R} \times \mathbb{H}^1(0,\infty)$ with $z(0) = x$, when $(X_0, Z_0) = (x,z)$ we have

$$\begin{aligned} \mathbb{E}\left[\tilde{f}(X_t, Z_t)\right] - \tilde{f}(x,z) = &\int_0^t \mathbb{E}\left[f_x(X_s, Z_s^n)\left(Z_s'(0) - \beta\right)\right] ds \\ &+ \sum_{j=1}^n \int_0^t \mathbb{E}\left[f_j(X_s, Z_s^n)\left(Z_s'(r_j) - \bar{G}(r_j)\beta\right)\right] ds \\ &+ \int_0^t \mathbb{E}\left[f_{xx}(X_s, Z_s^n)\right] ds \\ &+ \sum_{j=1}^n c(0,r_j) \int_0^t \mathbb{E}\left[f_{xj}(X_s, Z_s^n)\right] ds \\ &+ \frac{1}{2}\sum_{i,j=1}^n c(i,j) \int_0^t \mathbb{E}\left[f_{ij}(X_s, Z_s^n)\right] ds, \end{aligned}$$

By continuity of X and Z and since f is continuous and bounded,

(3.B.5)
$$\begin{aligned}\mathcal{L}_{-}\tilde{f}(x,z) = \lim_{t\to\infty}&\frac{\mathbb{E}\left[f(X_s, Z_t^n)\right] - f(x,z^n)}{t}\\ =&f_x(x,z^n)\{z'(0)-\beta\} + \sum_{j=1}^{n} f_j(x,z^n)\{z'(r_j) - \bar{G}(r_j)\beta\}\\ &+ f_{xx}(x,z^n) + \sum_{j=1}^{n} c(0,j) f_{xj}(x,z^n) + \frac{1}{2}\sum_{i,j=1}^{n} c(i,j) f_{i,j}(x,z^n).\end{aligned}$$

where $\mathcal{L}_{-}$ is the generator for sub-critical case.

3.B.2. generator for super-critical case. Next, consider the super-critical process $Y = (X, Z)$ that satisfies (3.B.1) with $dK_t = d\mathcal{M}_t(\mathbf{1}) - Z_t'(0)dt$. Hence,

$$dZ_t(r_j) = \big(Z_t'(r_j) - \bar{G}(r_j)Z_t'(0)\big)dt - d\mathcal{M}(\Phi_{r_j}\mathbf{1}) + \bar{G}(r_j)d\mathcal{M}_t(\mathbf{1})$$

Therefore,

$$\begin{aligned}d\langle X\rangle_t =&2dt,\\ d\langle X, Z(r_j)\rangle_t =&a(j)dt,\\ d\langle Z(r_i), Z(r_j)\rangle_t =&b(i,j)dt,\end{aligned}$$

where

(3.B.6)
$$a(j) = F(r_j, 0) - \bar{G}(r_j),$$

and

(3.B.7)
$$b(i,j) = F(r_i, r_j) - \bar{G}(r_i)F(r_j,0) - \bar{G}(r_j)F(r_i,0) + \bar{G}(r_i)\bar{G}(r_j).$$

Therefore, for every $(x,z) \in \mathbb{R} \times \mathbb{H}^1(0,\infty)$ with $z(0)=0$, when $(X_0, Z_0) = (x,z)$ we have

$$
\begin{aligned}
\mathbb{E}\left[f(X_t, Z_t^n)\right] - f(x, z^n) = &\int_0^t \mathbb{E}\left[f_x(X_s, Z_s^n)\big(Z_s'(0) - \beta\big)\right] ds \\
&+ \sum_{j=1}^n \int_0^t \mathbb{E}\left[f_j(X_s, Z_s^n)\big(Z_s'(r_j) - \bar{G}(r_j) Z_s'(0)\big)\right] ds \\
&+ \int_0^t \mathbb{E}\left[f_{xx}(X_s, Z_s^n)\right] ds \\
&+ \sum_{j=1}^n a(j) \int_0^t \mathbb{E}\left[f_{xj}(X_s, Z_s^n)\right] ds \\
&+ \frac{1}{2} \sum_{i,j=1}^n b(i,j) \int_0^t \mathbb{E}\left[f_{ij}(X_s, Z_s^n)\right] ds,
\end{aligned}
$$

By continuity of X and Z and since f is continuous and bounded,

$$
\begin{aligned}
\mathcal{L}_+ \tilde{f}(x,z) = &\lim_{t\to\infty} \frac{\mathbb{E}\left[f(X_t, Z_t^n)\right] - f(x,z^n)}{t} \\
= &f_x(x,z^n)\{z'(0) - \beta\} + \sum_{j=1}^n f_j(x,z^n)\{z'(r_j) - \bar{G}(r_j) z'(0)\} \\
&+ f_{xx}(x,z^n) + \sum_{j=1}^n a(j) f_{xj}(x,z^n) + \frac{1}{2}\sum_{i,j=1}^n b(i,j) f_{i,j}(x,z^n).
\end{aligned} \tag{3.B.8}
$$

where $\mathcal{L}_+$ is the generator for super-critical case.

3.B.3. Generator of the Limit Process Y. Combining the results of the last two sections, we can characterize the generator of the limit process Y. One can show that for every $(X_0, Z_0) = (x,z)$ and $x \neq 0$, for every smooth function f with compact support,

$$
\lim_{t\to\infty} \frac{1}{t} \mathbb{E}\left[\int_0^t f(X_s, Z_s^n) dL_0^X(s)\right] = 0,
$$

Therefore, by continuity of (X,Z) and continuity and boundedness of f, similar to the sub- and super-critical cases,

$$
\mathcal{L}\tilde{f}(x,z) = \lim_{t\to\infty} \frac{\mathbb{E}\left[\tilde{f}(X_t, Z_t)\right] - \tilde{f}(x,z)}{t} = \begin{cases} \mathcal{L}_+ \tilde{f}(x,z) & \text{if } x > 0 \\ \mathcal{L}_- \tilde{f}(x,z) & \text{if } x < 0 \end{cases} \tag{3.B.9}
$$

This shows that the generator for critical case equals to that of sub-critical case on the sub-critical region $\{x < 0\}$ and equals to the generator of super-critical case on the super-critical region $\{x > 0\}$.

3.B.4. An Equation for Invariant Distribution. We are looking for a distribution $\boldsymbol{\pi}$ on $\mathbb{R} \times \mathbb{H}^1(0,\infty)$ such that for every $\tilde{f} \in \mathbb{F}$,

$$\int \mathcal{L}\tilde{f}(x,z)\boldsymbol{\pi}(dx,dz) = 0 \tag{3.B.10}$$

Recall that $\tilde{f}(x,z) = f(x, z(r_1), ..., z(r_n))$ for some $n \geq 0, r_1, ..., r_n \in [0,\infty)$ and $f \in \mathbb{C}_c^\infty(\mathbb{R} \times \mathbb{R}^n; \mathbb{R})$. By equations (3.B.5), (3.B.8) and (3.B.9), for $x \neq 0$, $\mathcal{L}\tilde{f}(x,z)$ only depends on the vector $(x, z^n, z(0), z'^{\,n})$. Let the probability measure $\tilde{\boldsymbol{\pi}}$ be the push-forward of $\boldsymbol{\pi}$ under the projection

$$p : \mathbb{R} \times \mathbb{H}^1(0,\infty) \mapsto \mathbb{R}^{2n+2}; \quad p(x,z) = (x, z^n, z'(0), z'^{\,n}),$$

and assume that $\tilde{\boldsymbol{\pi}}$ has a density ϕ. If the support of f lies in the super-critical region $\{(x,z^n); x > 0\}$, then by equality (3.B.9),

$$\begin{aligned}\int \mathcal{L}\tilde{f}(x,z)\boldsymbol{\pi}(dx,dz) &= \int \mathcal{L}_+\tilde{f}(x,z)\phi(x,z^n,z_0,z'^{\,n})dxdz^n dz_0' dz'^{\,n} \\ &= -\int f(x,z)\mathcal{L}_+^*\phi(x,z^n,z_0,z'^{\,n})dxdz^n dz_0' dz'^{\,n}\end{aligned}$$

where $\mathcal{L}_+^*$ is the "dual" operator of $\mathcal{L}_+$. Therefore, the equation (3.B.10) implies

$$\int_{z_0, z'^{\,n}} \mathcal{L}_+^*\phi(x,z^n,z_0,z'^{\,n})dz_0' dz'^{\,n} = 0 \qquad \forall z^n \in \mathbb{R}^n, x > 0. \tag{3.B.11}$$

Similarly, if the support of f lies in the sub-critical region $\{(x,z^n), x < 0\}$, then the equation (3.B.10) implies

$$\int_{z_0, z'^{\,n}} \mathcal{L}_-^*\phi(x,z^n,z_0,z'^{\,n})dz_0' dz'^{\,n} = 0 \qquad \forall z^n \in \mathbb{R}^n, x < 0. \tag{3.B.12}$$

where $\mathcal{L}_-^*$ is the dual operator of $\mathcal{L}_-$. Now for general $\tilde{f} \in \mathbb{F}$,

$$
\begin{aligned}
\int \mathcal{L}\tilde{f}(x,z)\boldsymbol{\pi}(dx,dz) &= \int_{\{x>0\}} \mathcal{L}\tilde{f}(x,z)\boldsymbol{\pi}(dx,dz) + \int_{\{x<0\}} \mathcal{L}\tilde{f}(x,z)\boldsymbol{\pi}(dx,dz) \\
&= -\int_{\{x>0\}} f(x,z)\mathcal{L}_+^*\phi(x,z^n,z_0,z'^n)dxdz^ndz_0'dz'^n \\
&\quad -\int_{\{x<0\}} f(x,z)\mathcal{L}_-^*\phi(x,z^n,z_0,z'^n)dxdz^ndz_0'dz'^n \\
&\quad + \textbf{boundary terms on super-critical side} \\
&\quad + \textbf{boundary terms on sub-critical side}
\end{aligned}
$$

Since the first two terms on the right-hand side above are equal to zero, equation (3.B.10) implies

(3.B.13) $$\textbf{boundary terms on sub-critical side} = \textbf{boundary terms on sub-critical side}$$

Solving these equations gives a characterization of the stationary distribution π of Y. This would hopefully be the addressed in future works.

CHAPTER 4

Central Limit Theorem for the Sequence of Stationary Distributions

Finally in this chapter, we show that in the Halfin-Whitt asymptotic regime, the limit of the corresponding sequence of stationary distributions $\widehat{\pi}^{(N)}$ of the representation process $\{\widehat{Y}^{(N)}\}$ has a limit, and the limit is the unique invariant distribution of limit diffusion process $Y = (X, Z)$. This resolves the question of convergence of the scaled and centered steady-state distributions of $GI/GI/N$ queues in the Halfin-Whitt regime posed in **[51]** for a large class of service distributions. Here is the main result.

THEOREM 4.1. Suppose Assumptions I-IV hold. Then there exists a measure π on $\mathcal{Y}$ such that $\widehat{\pi}^{(N)} \Rightarrow \pi$ in $\mathcal{Y}$ as $\mathbb{N} \to \infty$. Furthermore, π is the unique invariant distribution of a $\mathcal{Y}$-valued Feller Markov family $\{Y^y_t; t \geq 0, y \in \mathcal{Y}\}$.

PROOF. See Section 4.2.3. □

First in Section 4.1, we prove the tightness of the sequence of stationary processes $\{\widehat{\pi}^{(N)}\}$ of the representation process $\widehat{Y}^{(N)}$. Although the tightness for the positive tail of $\widehat{X}^{(N)}$-marginal, $\widehat{X}^{(N)+}_\infty$, of $\widehat{\pi}^{(N)}$ is proved in **[37]**, we need in addition to establish tightness of the $\widehat{Z}^{(N)}$-marginal in its function space $\mathbb{H}^1(0, \infty)$. This requires a fair amount of computations (Section 4.1.3). We also have to strengthen the tightness result in **[37]** and instead show the uniform boundedness of $\widehat{X}^{(N)+}_\infty$ in $\mathbb{L}^1$ (Section 4.1.2). Finally in Section 4.2, we prove the convergence of $\{\widehat{\pi}^{(N)}\}$ by showing that every subsequential limit of $\{\widehat{\pi}^{(N)}\}$ is indeed the unique invariant distribution of the diffusion limit process Y.

4.1. Tightness of Stationary Distributions

Recall from Theorem 2.7 that for every $N \in \mathbb{N}$ and every initial configuration of the queueing system, $\widehat{Y}_t^{(N)}$ converges in distribution to $\widehat{\pi}^{(N)}$ as $t \to \infty$. In this section, we prove that the sequence of steady-state distributions $\{\widehat{\pi}^{(N)}\}_{N\in\mathbb{N}}$ is tight. Tightness of the $\widehat{X}^{(N)}$-marginal is proved in [**37**], but we also need to show tightness of $\widehat{Z}^{(N)}$-marginal in the right space. This result is stated in the next theorem.

THEOREM 4.2. Suppose Assumptions I-IV hold. Then, the sequence $\{\widehat{\pi}^{(N)}\}_{N\in\mathbb{N}}$ is tight in $\mathcal{Y}$.

PROOF. See Section 4.1.4. $\square$

By Remark 4.3 below, to prove Theorem 4.2, it is enough to show that for some specifically chosen initial condition, the family $\{\widehat{Y}_t^{(N)} = (\widehat{X}_t^{(N)}, \widehat{Z}_t^{(N)}); t \geq 0, N \in \mathbb{N}\}$ is tight in $\mathcal{Y}$. We choose a specific initial condition also considered in [**37**], that is convenient for establishing tightness of associated family $\{\widehat{Y}_t^{(N)}; t \geq 0, N \in \mathbb{N}\}$. This initial condition is described in Section 4.1.1. In Section 4.1.2, we recall a bound on the queue length of the N-server queue, obtained in [**37**] under the aforementioned specific initial condition, and establish a uniform $\mathbb{L}^1$ bound for the scaled queue length process. Our bound is stronger than the tightness result established in [**37**, Theorem 1], and is required to establish the tightness of the family $\{\widehat{Z}_t^{(N)}; t \geq 0, N \in \mathbb{N}\}$ in Section 4.1.3. The proof of Theorem 4.2 is concluded in Section 4.1.4.

REMARK 4.3. Suppose a family $\{\xi_t^{(N)}; t \geq 0, N \in \mathbb{N}\}$ of random variables taking values in a Polish space $\mathcal{X}$ is tight, and for every $N \in \mathbb{N}$, $\xi_t^{(N)} \Rightarrow \xi_\infty^{(N)}$ in $\mathcal{X}$ as $t \to \infty$. Then, for every $\epsilon > 0$, there exists a compact subset $K_\epsilon \subset \mathcal{X}$ such that

$$\sup_{N\in\mathbb{N},\, t\geq 0} \mathbb{P}\left\{\xi_t^{(N)} \in K_\epsilon^c\right\} < \epsilon.$$

Since K_ϵ^c is open, by the portmanteau theorem for weak convergence, for every $N \in \mathbb{N}$ we have

$$\mathbb{P}\left\{\xi_\infty^{(N)} \in K_\epsilon^c\right\} \le \limsup_{t\to\infty} \mathbb{P}\left\{\xi_t^{(N)} \in K_\epsilon^c\right\} \le \sup_{t\ge 0} \mathbb{P}\left\{\xi_t^{(N)} \in K_\epsilon^c\right\},$$

and hence

$$\sup_{N\in\mathbb{N}} \mathbb{P}\left\{\xi_\infty^{(N)} \in K_\epsilon^c\right\} \le \sup_{N\in\mathbb{N},\, t\ge 0} \mathbb{P}\left\{\xi_t^{(N)} \in K_\epsilon^c\right\} < \epsilon.$$

Therefore, the sequence $\{\xi_\infty^{(N)}\}_{N\in\mathbb{N}}$ is also tight in $\mathcal{X}$.

4.1.1. A Special Initial Condition. Recall from equation (2.2.15) that the process $\widehat{Y}^{(N)}$ is a function of the Markovian state space $V^{(N)}$. We now describe the specific initial condition for $V^{(N)}$, that is also used in [**37**], which facilitates establishing tightness of the family $\{\widehat{Y}_t^{(N)}; t \ge 0, N \in \mathbb{N}\}$ (see the discussion above).

Recall from Assumption I that the arrival process $E^{(N)}$ to the N-server system is a renewal process with inter-arrival distribution $G_E^{(N)}$ and delay $u_0^{(N)}$. Define $R_*^{(N)}$ to be a random variable with distribution

$$\mathbb{P}\left\{R_*^{(N)} \in A\right\} = \int_A \overline{G}_E^{(N)}(x)dx, \qquad \forall A \in \mathcal{B}[0,\infty).$$

When the delay $u_0^{(N)}$, and hence the initial forward recurrence time $R_0^{(N)}$, has the same distribution as $R_*^{(N)}$, the arrival process $E^{(N)}$ becomes a stationary renewal process (see [**3**, Theorem V.3.3]). Next, define $X_*^{(N)} \doteq N$. Also, let $\{a_j^*; j \in \mathbb{N}\}$ be a sequence of i.i.d. random variables with common distribution $\mathbb{P}\{a_j^* \in A\} = \int_A \overline{G}(x)dx$ for all $A \in \mathcal{B}[0,\infty)$, and define

$$\nu_*^{(N)} \doteq \sum_{j=1}^{N} \delta_{a_j^*}. \tag{4.1.1}$$

Finally, define

$$\pi_*^{(N)} \doteq \mathscr{L}aw(R_*^{(N)}, X_*^{(N)}, \nu_*^{(N)}). \tag{4.1.2}$$

To summarize, when the initial condition $V_0^{(N)}$ is distributed as $\pi_*^{(N)}$, there are initially N jobs in system (and hence by the non-idling assumption, all jobs are in service), with independent initial ages distributed according to the so-called residual distribution of G (i.e. with density $\overline{G}(x)dx$), and the arrival process is stationary.

LEMMA 4.4. If $(X_0^{(N)}, \nu_0^{(N)}) = (X_*^{(N)}, \nu_*^{(N)})$ for every N, then the sequence $\{(X_0^{(N)}, \nu_0^{(N)})\}$ satisfies (2.3.4).

PROOF. Since $\overline{X}_*^{(N)} = 1$ for all N, $\overline{X}_*^{(N)} \to 1$ both almost surely and in expectation. Also, for every continuous bounded function f, $\overline{\nu}_*^{(N)}(f) = \sum_{j=1}^{N} f(a_j^*)/N$. Since $\{a_j^*\}_{j\in\mathbb{Z}}$ are i.i.d. with p.d.f. $\overline{G}$, by the strong law of large numbers and definition (2.3.3) of $\overline{\nu}$, we have

$$\overline{\nu}_*^{(N)}(f) \to \mathbb{E}\left[f(a_1^*)\right] = \int_0^\infty f(x)\overline{G}(x)dx = \overline{\nu}(f), \qquad a.s.$$

Therefore, $\overline{\nu}^{(N)} \to \overline{\nu}$ in $\mathbb{M}_F[0,\infty)$, almost surely, and hence, in distribution. □

4.1.2. Bounds on the Queue Length. Now we recall the bound on the queue length $(X_t^{(N)})^+$, that was obtained in **[37]**. For every $N \in \mathbb{N}$, define the random variable

$$\widehat{Q}^{(N)} \doteq \frac{1}{\sqrt{N}} \sup_{t\geq 0} \left(A^{(N)}(t) - \sum_{i=1}^{N} D_i(t) \right), \tag{4.1.3}$$

where $A^{(N)}$ is an equilibrium renewal process with renewal time distribution $G_E^{(N)}$, and $D_i; i = 1, ..., N$, are i.i.d. equilibrium renewal processes independent of $A^{(N)}$, with renewal distribution G. Also, recall that a random variable X is said to be stochastically dominated by another random variables Y, denoted $X \overset{(d)}{\leq} Y$, if $\mathbb{P}\{X > x\} \leq \mathbb{P}\{Y > x\}$ for every $x \in \mathbb{R}$.

PROPOSITION 4.5. Suppose Assumptions I-II hold, and $V_0^{(N)}$ is distributed as $\pi_*^{(N)}$. Then, for every $t \geq 0$,

$$\widehat{X}_t^{(N)+} \overset{(d)}{\leq} \widehat{Q}^{(N)}. \tag{4.1.4}$$

PROOF. Define

$$\tilde{Q}^{(N)}(t) \doteq \frac{1}{\sqrt{N}} \sup_{0\leq s\leq t} \left(A^{(N)}(s) - \sum_{i=1}^{N} D_i(s) \right), \qquad t \geq 0.$$

Since $\tilde{Q}^{(N)}(t) \leq \widehat{Q}^{(N)}$ for all $t \geq 0$, it suffices to show that for every $t \geq 0$, $\widehat{X}_t^{(N)+} \overset{(d)}{\leq} \tilde{Q}^{(N)}(t)$. However, the latter inequality would follow from Theorem 3 in [**37**], once we verify the assumptions H-W and T_0 in [**37**, Section 2.1]. Recalling these assumptions in our notation, H-W requires that $\mathbb{P}\{v_j = 0\} = \mathbb{P}\{\tilde{u}_n = 0\} = 0$, and that $\mathbb{E}\left[v_j\right] = \mathbb{E}\left[\tilde{u}_n\right] = 1$, where recall that $v_j, j \geq 1$ are the service times, and $\tilde{u}_n, n \geq 1$ are the inter-arrival times of the process $\tilde{E}$, and inter-arrival times of the arrival process $E^{(N)}$ must satisfy $u_n^{(N)} = \tilde{u}_n/\lambda^{(N)}$, with $\lambda^{(N)}$ defined in (2.1.1). This follows from our Assumptions I and II.a. Next, condition T_0 of [**37**], again in our notation, requires the following.

(i) There exists $\epsilon > 0$ such that $\mathbb{E}[\tilde{u}_1^{(2+\epsilon)}], \mathbb{E}[v_1^{(2+\epsilon)}] < \infty$.

(ii) v_j has a non-zero variance.

(iii) $\lim_{t\to 0} t^{-1}G(t) < \infty$.

(iv) For every $N \in \mathbb{N}$, the number of jobs in system $X_t^{(N)}$ (which is denote by $Q^N(t)$ in [**37**]), satisfies $X_t^{(N)} \Rightarrow X_\infty^{(N)}$ as $t \to \infty$.

First note that (i) holds because the inter-arrival times and service times have finite $(2+\epsilon)$ moments by Assumptions I and II.b, respectively. Next, since the service time distribution has a density g, it has a non-zero variance and also satisfies the condition (iii); see the discussion below assumption T_0 on page 4 of [**37**]. Finally, under Assumptions I and II, (iv) holds by Proposition 2.9. This concludes the proof.

□

The tightness of the sequence $\{\widehat{Q}^{(N)}; N \in \mathbb{N}\}$ is proved in [**37**, Theorem 1]. Here, we prove a stronger result which we use to show tightness of $\{\widehat{Z}_t^{(N)}; t \geq 0, N \in \mathbb{N}\}$, namely that the sequence $\{\widehat{Q}^{(N)}; N \in \mathbb{N}\}$ is uniformly bounded in $\mathbb{L}^1$.

PROPOSITION 4.6. Suppose Assumptions I and II hold, and $V_0^{(N)}$ is distributed as $\pi_*^{(N)}$. Then, there exists a constant $C_Q < \infty$ such that

$$\sup_{N\in\mathbb{N}} \mathbb{E}\left[\widehat{Q}^{(N)}\right] \leq C_Q. \tag{4.1.5}$$

PROOF. First, note that one can construct a pure renewal process $\tilde{A}^{(N)}$ with the same renewal distribution as $A^{(N)}$, such that $A^{(N)}(t) \leq \tilde{A}^{(N)}(t) + 1$ for all $t \geq 0$. Therefore, it is sufficient to prove the proposition for the case when $A^{(N)}$ is a pure renewal process, which we assume for the rest of the proof.

Set $\tau_0^{(N)} = 0$, and let $\tau_n^{(N)}$ and $\theta_n^{(N)}$, $n \geq 1$, be, respectively, the epoch times and renewal times of the process $A^{(N)}$. Since the process $A^{(N)} - \sum_{i=1}^N D_i$ is decreasing on $[\tau_n^{(N)}, \tau_{n+1}^{(N)})$ for all $n \geq 0$, $\widehat{Q}^{(N)}$ in (4.1.3) can be written as $\widehat{Q}^{(N)} = Q^{(N)}/\sqrt{N}$, with

$$Q^{(N)} = \sup_{k \geq 0} \left(k - \sum_{i=1}^N D_i(\tau_k^{(N)}) \right) = \sup_{k \geq 0} \left(S_k^{(N)} \right), \tag{4.1.6}$$

where

$$S_k^{(N)} \doteq \sum_{j=1}^k \zeta_j^{(N)}, \qquad \text{with} \qquad \zeta_j^{(N)} \doteq 1 - \sum_{i=1}^N (D_i(\tau_j^{(N)}) - D_i(\tau_{j-1}^{(N)})).$$

By the stationarity of D_i and the independence of $A^{(N)}$ and D_i, $i = 1, \ldots, N$, $\zeta_j^{(N)}$, $j \geq 1$, are identically distributed, with common mean

$$\epsilon^{(N)} \doteq \mathbb{E}\left[\zeta_1^{(N)}\right] = 1 - N\mathbb{E}\left[D_1(\theta_1^{(N)})\right] = 1 - \frac{N}{\lambda^{(N)}} = \frac{-\beta}{\sqrt{N} - \beta},$$

where the last equation uses (2.1.1). By equation (12) in **[37]**, there exists $p > 2$ and a constant $C_1 < \infty$, such that for every $k \in \mathbb{N}$ and $x \geq 0$,

$$\mathbb{P}\left\{ \max_{1 \leq j \leq k} (S_j^{(N)} - j\epsilon^{(N)}) > x \right\} \leq C_1 k^{\frac{p}{2}} x^{-p}. \tag{4.1.7}$$

Note that in **[37]**, $S_j^{(N)}$ is denoted by $W_{N,j}$, $\epsilon^{(N)}$ is denoted by a_N, p is replaced by r, and C_1 is replaced by $K_r(C_1+1)^{r/2}$; see the definitions of $W_{n,j}$ and a_n on page 17 of **[37]**.

Since $|\epsilon^{(N)}|\sqrt{N} \to \beta$ as $N \to \infty$, (4.1.5) is equivalent to showing $\mathbb{E}[|\epsilon^{(N)}|Q^{(N)}] \leq C$ for a constant $C < \infty$. Analogous to the argument used in **[95]**, for every $x \geq 0$

and large enough N (i.e. $N > \beta^2$),

$$\begin{aligned}
\mathbb{P}\left\{|\epsilon^{(N)}|Q^{(N)} > x\right\} &= \mathbb{P}\left\{S_j^{(N)} > \frac{x}{|\epsilon^{(N)}|} \text{ for some } j \geq 0\right\} \\
&= \mathbb{P}\left\{(S_j^{(N)} - j\epsilon^{(N)}) > \frac{x}{|\epsilon^{(N)}|} + j|\epsilon^{(N)}| \text{ for some } j \geq 0\right\} \\
&\leq \sum_{k=0}^{\infty} \mathbb{P}\left\{(S_j^{(N)} - j\epsilon^{(N)}) > \frac{x}{|\epsilon^{(N)}|} + j|\epsilon^{(N)}| \text{ for some } 2^k \leq j < 2^{k+1}\right\} \\
&\leq \sum_{k=0}^{\infty} \mathbb{P}\left\{\max_{0\leq j\leq 2^{k+1}} (S_j^{(N)} - j\epsilon^{(N)}) > \frac{x}{|\epsilon^{(N)}|} + 2^k|\epsilon^{(N)}|\right\} \\
&\leq C_1 \sum_{k=0}^{\infty} (2^{k+1})^{\frac{p}{2}} \left(\frac{x}{|\epsilon^{(N)}|} + 2^k|\epsilon^{(N)}|\right)^{-p} \\
&= C_2 \sum_{k=0}^{\infty} \left(\frac{x}{|\epsilon^{(N)}|\sqrt{2^k}} + |\epsilon^{(N)}|\sqrt{2^k}\right)^{-p}
\end{aligned}$$

with $C_2 \doteq \sqrt{2^p}C_1$, where the last inequality follows from (4.1.7). Now, since for every non-negative random variable X

$$\mathbb{E}[X] = \int_0^{\infty} \mathbb{P}\{X > x\}\, dx \leq 1 + \int_1^{\infty} \mathbb{P}\{X > x\}\, dx,$$

we have

$$\mathbb{E}\left[|\epsilon^{(N)}|Q^{(N)}\right] \leq 1 + C_2 \sum_{k=0}^{\infty} \int_1^{\infty} \left(\frac{x}{|\epsilon^{(N)}|\sqrt{2^k}} + |\epsilon^{(N)}|\sqrt{2^k}\right)^{-p} dx.$$

Performing the change of variable

$$u = \frac{x}{|\epsilon^{(N)}|\sqrt{2^k}} + |\epsilon^{(N)}|\sqrt{2^k},$$

setting

$$u_0 \doteq \frac{1}{|\epsilon^{(N)}|\sqrt{2^k}} + |\epsilon^{(N)}|\sqrt{2^k}, \quad \text{and} \quad C_3 \doteq \frac{C_2}{p-1},$$

and recalling that $p > 1$, we have, for arbitrary $n \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}\left[|\epsilon^{(N)}|Q^{(N)}\right] &\leq 1 + C_2 \sum_{k=0}^{\infty} |\epsilon^{(N)}|\sqrt{2^k} \int_{u_0}^{\infty} u^{-p} du \\
&= 1 + C_3 \sum_{k=0}^{\infty} |\epsilon^{(N)}|\sqrt{2^k} \left(\frac{1}{|\epsilon^{(N)}|\sqrt{2^k}} + |\epsilon^{(N)}|\sqrt{2^k} \right)^{-p+1} \\
&\leq 1 + C_3 \sum_{k=0}^{n} |\epsilon^{(N)}|\sqrt{2^k} \left(\frac{1}{|\epsilon^{(N)}|\sqrt{2^k}} \right)^{-p+1} \\
&\quad + C_3 \sum_{k=n+1}^{\infty} |\epsilon^{(N)}|\sqrt{2^k} \left(|\epsilon^{(N)}|\sqrt{2^k} \right)^{-p+1} \\
&\leq 1 + C_3 \sum_{k=0}^{n} \left(|\epsilon^{(N)}|\sqrt{2^k} \right)^{p} + C_3 \sum_{k=n}^{\infty} \left(|\epsilon^{(N)}|\sqrt{2^k} \right)^{-p+2}.
\end{aligned}$$

By standard calculations for geometric sums, we have

$$\sum_{k=0}^{n} \left(|\epsilon^{(N)}|\sqrt{2^k} \right)^{p} = |\epsilon^{(N)}|^p \sum_{k=0}^{n} (2^{\frac{p}{2}})^k = |\epsilon^{(N)}|^p \frac{(2^{\frac{p}{2}})^n - 1}{2^{\frac{p}{2}} - 1} \leq (|\epsilon^{(N)}|\sqrt{2^n})^p,$$

and (using the fact that $p > 2$),

$$\begin{aligned}
\sum_{k=n}^{\infty} \left(|\epsilon^{(N)}|\sqrt{2^k} \right)^{-p+2} &= |\epsilon^{(N)}|^{-p+2} \sum_{k=n}^{\infty} (2^{-\frac{p-2}{2}})^k \\
&= |\epsilon^{(N)}|^{-p+2} \frac{(2^{-\frac{p-2}{2}})^n}{1 - 2^{-\frac{p-2}{2}}} \\
&\leq \frac{1}{1 - 2^{-\frac{p-2}{2}}} (|\epsilon^{(N)}|\sqrt{2^n})^{-p+2}.
\end{aligned}$$

Combining the last three displays, for $C_4 \doteq (1 - 2^{-(p-2)/2})^{-1} C_3$, we have

$$\mathbb{E}\left[|\epsilon^{(N)}|Q^{(N)}\right] \leq 1 + C_4 \left((|\epsilon^{(N)}|\sqrt{2^n})^p + (|\epsilon^{(N)}|\sqrt{2^n})^{-p+2} \right). \tag{4.1.8}$$

For fixed N, the above inequality holds for every $n \in \mathbb{N}$. We now find an appropriate $n = n^*(N)$ for each N to get the desirable bound. Define $z(N) \doteq \log_2(|\epsilon^{(N)}|^{-2}) > 0$, so that $|\epsilon^{(N)}|\sqrt{2^{z(N)}} = 1$, and let $n^*(N) = \lceil z(N) \rceil$, so that $z(N) \leq n^*(N) \leq z(N) + 1$, to obtain

$$\begin{aligned}
(|\epsilon^{(N)}|\sqrt{2^{n^*(N)}})^p + (|\epsilon^{(N)}|\sqrt{2^{n^*(N)}})^{-p+2} &\leq (|\epsilon^{(N)}|\sqrt{2^{z(N)+1}})^p + (|\epsilon^{(N)}|\sqrt{2^{z(N)}})^{-p+2} \\
&\leq 1 + 2^{p/2}. \qquad (4.1.9)
\end{aligned}$$

Combining (4.1.8) and (4.1.9), we have

$$\begin{aligned}\mathbb{E}\left[|\epsilon^{(N)}|Q^{(N)}\right] &\leq 1 + C_4\left((|\epsilon^{(N)}|\sqrt{2^{n^*(N)}})^p + (|\epsilon^{(N)}|\sqrt{2^{n^*(N)}})^{-p+2}\right) \\ &\leq 1 + C_4(1 + 2^{p/2}),\end{aligned}$$

which completes the proof. □

The following corollary follows on combining the last two propositions.

COROLLARY 4.7. Suppose Assumptions I and II holds, and $V_0^{(N)}$ is distributed as $\pi_*^{(N)}$. Then, for every $t \geq 0$,

$$\sup_{N\in\mathbb{N},t\geq 0} \mathbb{E}\left[\widehat{X}_t^{(N)+}\right] \leq C_Q. \tag{4.1.10}$$

4.1.3. Tightness of $\widehat{Z}^{(N)}$. In this section, we establish the following result.

PROPOSITION 4.8. Suppose Assumptions I- IV hold, and $V_0^{(N)}$ is distributed as $\pi_*^{(N)}$. Then, the family $\{\widehat{Z}_t^{(N)}; t \geq 0, N \in \mathbb{N}\}$ is tight in $\mathbb{H}^1(0,\infty)$.

The rest of the section is devoted to the proof of this proposition. We first describe criteria for tightness of $\mathbb{H}^1(0,\infty)$-valued processes in Section 4.1.3.1, an then obtain estimates in Sections 4.1.3.2 and 4.1.3.3 to verify these criteria for the family $\{\widehat{Z}_t^{(N)}; t \geq 0, N \in \mathbb{N}\}$. As shown at the end of Section 4.1.3.3, the proof of Proposition 4.8 then follows as an immediate consequence.

4.1.3.1. *Tightness Criteria for $\mathbb{H}^1(0,\infty)$-valued Processes.* We start with criteria for tightness in $\mathbb{H}^1(0,\infty)$.

PROPOSITION 4.9. Suppose a family $\{\zeta_\alpha(r); \alpha \in \mathcal{A}, r \in (0,\infty)\}$ of random variables satisfies the followings.

(a) For every $L > 0$,

$$\lim_{\lambda\to\infty} \sup_{\alpha\in\mathcal{A}} \mathbb{P}\left\{\|\zeta_\alpha\|_{\mathbb{L}^2(0,L)} > \lambda\right\} = 0, \tag{4.1.11}$$

and for every $\delta > 0$,

$$\lim_{L\to\infty} \sup_{\alpha\in\mathcal{A}} \mathbb{P}\left\{\|\zeta_\alpha\|_{\mathbb{L}^2(L,\infty)} > \delta\right\} = 0. \tag{4.1.12}$$

(b) For every $\alpha \in \mathcal{A}$, almost surely, $\zeta_\alpha(\cdot)$ has a density $\zeta'_\alpha(\cdot)$ on $(0,\infty)$ such that for every $L > 0$,

$$\lim_{\lambda\to\infty} \sup_{\alpha\in\mathcal{A}} \mathbb{P}\left\{\|\zeta'_\alpha\|_{\mathbb{L}^2(0,L)} > \lambda\right\} = 0, \tag{4.1.13}$$

and for every $\delta > 0$,

$$\lim_{L\to\infty} \sup_{\alpha\in\mathcal{A}} \mathbb{P}\left\{\|\zeta'_\alpha\|_{\mathbb{L}^2(L,\infty)} > \delta\right\} = 0. \tag{4.1.14}$$

(c) For every $\alpha \in \mathcal{A}$, almost surely, $\zeta'_\alpha(\cdot)$ has itself a density $\zeta''_\alpha(\cdot)$ on $(0,\infty)$ such that for every $L > 0$,

$$\lim_{\lambda\to\infty} \sup_{\alpha\in\mathcal{A}} \mathbb{P}\left\{\|\zeta''_\alpha\|_{\mathbb{L}^2(0,L)} > \lambda\right\} = 0. \tag{4.1.15}$$

Then, $\{\zeta_\alpha(\cdot); \alpha \in \mathcal{A}\}$ it is tight in $\mathbb{H}^1(0,\infty)$.

We differ the proof of Proposition 4.9 to Appendix 4.A.2. The following Lemma provides a sufficient condition for the criteria of Proposition 4.9 which is easier to verify in some cases.

LEMMA 4.10. For a given family $\{\zeta_\alpha(r); \alpha \in \mathcal{A}, r \in (0,\infty)\}$ of random variables, assume for every $\alpha \in \mathcal{A}$, almost surely, the densities ζ'_α and ζ''_α exist, and there exists integrable functions $R_1, R_2 \in \mathbb{L}^1(0,\infty)$, and a locally integrable function $R_3 \in \mathbb{L}^1_{\text{loc}}(0,\infty)$ such that for every $\alpha \in \mathcal{A}$ and $r \geq 0$,

$$\mathbb{E}\left[|\zeta_\alpha(r)|^2\right] \leq R_1(r), \qquad \mathbb{E}\left[|\zeta'_\alpha(r)|^2\right] \leq R_2(r), \qquad \mathbb{E}\left[\mathbb{E}|\zeta''_\alpha(r)|^2\right] \leq R_3(r).$$

Then, the family $\{\zeta_\alpha(\cdot); \alpha \in \mathcal{A}\}$ is tight in $\mathbb{H}^1(0,\infty)$.

PROOF. Fix $L > 0$. Using Markov's inequality and Tonelli's theorem, for every $\alpha \in \mathcal{A}$ and $\lambda \geq 0$, we have

$$\mathbb{P}\left\{\|\zeta_\alpha\|_{\mathbb{L}^2(0,L)} > \lambda\right\} \leq \frac{1}{\lambda^2}\mathbb{E}\left[\|\zeta_\alpha\|^2_{\mathbb{L}^2(0,L)}\right] = \frac{1}{\lambda^2}\int_0^L \mathbb{E}\left[|\zeta_\alpha(r)|^2\right] dr \leq \frac{1}{\lambda^2}\int_0^L R_1(r)dr.$$

The right-hand side does not depend on α, and is finite because R_1 is locally integrable. Thus, taking the supremum over $\alpha \in \mathcal{A}$, and then the limit $\lambda \to \infty$ on both sides of the above inequality, (4.1.11) follows. The proofs of properties (4.1.13) and (4.1.15) are exactly analogous.

Similarly, for every $\delta > 0$,

$$\mathbb{P}\left\{\|\zeta_\alpha\|_{\mathbb{L}^2(L,\infty)} > \delta\right\} \leq \frac{1}{\delta^2}\int_L^\infty \mathbb{E}\left[|\zeta_\alpha(r)|^2\right] dr \leq \frac{1}{\delta^2}\int_L^\infty R_1(r)dr. \tag{4.1.16}$$

Since R_1 is integrable, the right-hand side converges to zero as $L \to \infty$. Taking the supremum over $\alpha \in \mathcal{A}$, and then the limit $L \to \infty$ on both sides above, (4.1.13) follows. The proof of property (4.1.12) is exactly analogous. □

4.1.3.2. *A Bound on the Martingale Measure Stochastic Integrals.* In this section we deal with the component $\widehat{\mathcal{M}}_t^{(N)}(\Psi_{t+\cdot}\mathbf{1})$ of $\widehat{X}_t^{(N)}$. The next lemma derives estimates on the covariance function (2.3.6) of $\mathcal{M}^{(N)}$.

LEMMA 4.11. Suppose Assumptions I-III hold and $V_0^{(N)}$ is distributed as $\pi_*^{(N)}$. Then, for every $t \geq 0$ and nonnegative, absolutely continuous, bounded function f, we have

$$\mathbb{E}\left[\overline{\nu}_t^{(N)}(f)\right] \leq \int_0^\infty (f\overline{G})(x)dx + C_Q\int_0^t |(f\overline{G})'(s)|ds. \tag{4.1.17}$$

In addition, if Assumption IV.a holds, there exists a bounded function R_M on $[0,\infty)$, which is also integrable on $(0,\infty)$, such that for every $t \geq 0$ and $N \in \mathbb{N}$,

$$\mathbb{E}\left[\int_0^t \overline{\nu}_{t-s}^{(N)}\left(\frac{\overline{G}^2(\cdot + s + r)}{\overline{G}^2(\cdot)}\frac{g(\cdot)}{\overline{G}(\cdot)}\right)ds\right] \leq R_M(r), \quad r \geq 0. \tag{4.1.18}$$

PROOF. For the first claim, fix f and $t \geq 0$. Dividing both sides of (2.3.10) by N, and taking expectations, we have

$$\mathbb{E}\left[\overline{\nu}_t^{(N)}(f)\right] = \mathbb{E}\left[\overline{\nu}_0^{(N)}(\Phi_t f)\right] - \mathbb{E}\left[\overline{\mathcal{H}}_t^{(N)}(f)\right] + \mathbb{E}\left[\int_0^t (f\overline{G})(t-s)d\overline{K}_s^{(N)}\right]. \tag{4.1.19}$$

Since $\nu_0^{(N)}$ has the same distribution as $\nu_*^{(N)} = \sum_{i=1}^N \delta_{a_j^*}$, where $\{a_j^*, j \geq 0\}$ are i.i.d with p.d.f. $\overline{G}$, and by definition (2.2.8) of Φ_t, we have

$$\begin{aligned}
\mathbb{E}\left[\overline{\nu}_0^{(N)}(\Phi_t f)\right] &= \frac{1}{N}\mathbb{E}\left[\sum_{j=1}^N \frac{f(a_j^*+t)\overline{G}(a_j^*+t)}{\overline{G}(a_j^*)}\right] \\
&= \mathbb{E}\left[\frac{f(a_1^*+t)\overline{G}(a_1^*+t)}{\overline{G}(a_1^*)}\right] \\
&= \int_0^\infty (f\overline{G})(x+t)dx. \qquad (4.1.20)
\end{aligned}$$

Next, by definition (2.3.9) of $\mathcal{H}^{(N)}$ and properties of the stochastic integral with respect to the martingale measure $\mathcal{M}^{(N)}$, we have

$$\mathbb{E}[\mathcal{H}_t^{(N)}(f)] = 0. \qquad (4.1.21)$$

Finally, substituting $K^{(N)}$ from (2.3.11), integrating by parts, and using the fact that $X_0^{(N)} = N$ and the nonnegativity of f, we have for each $t \geq 0$

(4.1.22)

$$\begin{aligned}
\int_0^t (f\overline{G})(t-s)d\overline{K}_s^{(N)} &= \int_0^t (f\overline{G})(t-s)d\overline{E}_s^{(N)} - f(0)(\overline{X}_t^{(N)}-1)^+ \\
&\quad - \int_0^t (\overline{X}_s^{(N)}-1)^+(f\overline{G})'(t-s)ds \\
&\leq \int_0^t (f\overline{G})(t-s)d\overline{E}_s^{(N)} + \int_0^t (\overline{X}_s^{(N)}-1)^+|(f\overline{G})'(t-s)|ds \qquad (4.1.23)
\end{aligned}$$

When $V_0^{(N)}$ is distributed as $\pi_*^{(N)}$, $E^{(N)}$ is a stationary renewal process and satisfies $\mathbb{E}[\overline{E}^{(N)}(t)] = \overline{\lambda}^{(N)}t$. Therefore, using integration by parts twice and invoking Fubini's theorem, and using the facts that $\overline{\lambda}^{(N)} \leq 1$ and f is non-negative, we have

$$\begin{aligned}
\mathbb{E}\left[\int_0^t (f\overline{G})(t-s)d\overline{E}_s^{(N)}\right] &= f(0)\mathbb{E}\left[\overline{E}_t^{(N)}\right] + \int_0^t \mathbb{E}\left[\overline{E}_s^{(N)}\right](f\overline{G})'(t-s)ds \\
&= \overline{\lambda}^{(N)}f(0)t + \overline{\lambda}^{(N)}\int_0^t s(f\overline{G})'(t-s)ds \\
&\leq \int_0^t (f\overline{G})(s)ds. \qquad (4.1.24)
\end{aligned}$$

Finally, by Corollary 4.7, for every $s \geq 0$, $\mathbb{E}[(\overline{X}_s^{(N)} - 1)^+] = \mathbb{E}[\widehat{X}_s^{(N)+}]/\sqrt{N} \leq C_Q$, and hence,

$$\mathbb{E}\left[\int_0^t (\overline{X}_s^{(N)} - 1)^+ |(f\overline{G})'(t-s)| ds\right] \leq \int_0^t \mathbb{E}\left[(\overline{X}_s^{(N)} - 1)^+\right] |(f\overline{G})'(t-s)| ds$$
$$\leq C_Q \int_0^t |(f\overline{G})'(s)| ds. \tag{4.1.25}$$

Equation (4.1.17) then follows on substituting the relations (4.1.20)-(4.1.25) in (4.1.19).

For the second claim, note that by Assumptions II-III and recalling the constants H and H_2 defined in Remark 2.2, for every $s, r \geq 0$, the function $f_* \doteq \overline{G}^2(s + r + \cdot)g(\cdot)/\overline{G}^3(\cdot)$ is non-negative, absolutely continuous and bounded, and satisfies $|f_*\overline{G}(x)| \leq H\overline{G}(x + s + r)$. Also,

$$(f_*\overline{G})'(x) = \frac{-2\overline{G}(x+s+r)g(x+s+r)g(x) + \overline{G}^2(x+s+r)g'(x)}{\overline{G}(x)^2} + \frac{\overline{G}^2(x+s+r)g^2(x)}{\overline{G}^3(x)},$$

which implies $|(f_*\overline{G})'(x)| \leq (3H^2 + H_2)\overline{G}(x + s + r)$. Therefore, replacing f with f_* and t by $t - s$ in (4.1.17), we have

$$\mathbb{E}\left[\overline{\nu}_{t-s}^{(N)}\left(\frac{\overline{G}^2(\cdot + s + r)}{\overline{G}^2(\cdot)}\frac{g(\cdot)}{\overline{G}(\cdot)}\right)\right] \leq H\int_0^\infty \overline{G}(s+r+x)dx$$
$$+ C_Q(3H^2 + H_2)\int_0^{t-s} \overline{G}(s+r+x)dx,$$

and hence, by Tonelli's theorem, (4.1.18) holds with

$$R_M(r) \doteq (H + C_Q(3H^2 + H_2))\int_r^\infty \int_s^\infty \overline{G}(x)dx\, ds.$$

The function R_M is bounded and integrable by Assumption IV.a (see Remark 2.4).

□

The next lemma shows absolute continuity of the the martingale measure stochastic integral term. A similar result is obtained in [**62**, Lemma E.1].

LEMMA 4.12. Suppose Assumptions II-III hold. Then, for every $N \in \mathbb{N}$ and every $t, r \geq 0$,

$$\mathcal{M}_t^{(N)}(\Psi_{t+r}\mathbf{1}) = \mathcal{H}_t^{(N)}(\mathbf{1}) - \int_0^r \mathcal{M}_t^{(N)}(\Psi_{t+u}h)du, \tag{4.1.26}$$

and

$$\mathcal{M}_t^{(N)}(\Psi_{t+r}h) = \mathcal{H}_t^{(N)}(h) + \int_0^r \mathcal{M}_t^{(N)}(\Psi_{t+u}h_2)du, \tag{4.1.27}$$

where $h_2 = g/\overline{G}$ is as defined in Remark 2.2.

PROOF. Fix $t \geq 0$. By Assumption III, the function

$$(\Psi_{t+u}h)(x,s) = \frac{g(x+t+u-s)}{\overline{G}(x)},$$

is bounded, continuous, and satisfies

$$\int_0^r (\Psi_{t+u}h)(x,s)du = \frac{\overline{G}(x+t-s)}{\overline{G}(x)} - \frac{\overline{G}(x+t+r-s)}{\overline{G}(x)} = \Psi_t\mathbf{1}(x,s) - \Psi_{t+r}\mathbf{1}(x,s).$$

Therefore, by the stochastic Fubini theorem for martingale measures (e.g. see [**98**, Theorem 2.6], and note that $\mathcal{M}^{(N)}$ is an orthogonal, and therefore a worthy martingale measure, see definitions therein) we have

$$\int_0^r \mathcal{M}_t^{(N)}(\Psi_{t+u}h)du = \mathcal{M}_t^{(N)}\left(\int_0^r \Psi_{t+u}h\, du\right) = \mathcal{M}_t^{(N)}(\Psi_t\mathbf{1}) - \mathcal{M}_t^{(N)}(\Psi_{t+r}\mathbf{1}),$$

and (4.1.26) follows from definition (2.3.9) of $\mathcal{H}^{(N)}$. Similarly, by Assumption III,

$$(\Psi_{t+u}h_2)(x,s) = \frac{g'(x+t+u-s)}{\overline{G}(x)},$$

is bounded continuous and satisfies

$$\int_0^r (\Psi_{t+u}h_2)(x,s)du = \frac{g(x+t+r-s)}{\overline{G}(x)} - \frac{g(x+t-s)}{\overline{G}(x)} = \Psi_{t+r}h(x,s) - \Psi_t h(x,s).$$

The equality (4.1.27) then follows from another application of the stochastic Fubini theorem for martingale measures. $\square$

Now we can prove tightness of the martingale measure integral term.

LEMMA 4.13. Suppose Assumptions I-III and IV.a hold, and $V_0^{(N)}$ is distributed as $\pi_*^{(N)}$. Then, the family $\{\widehat{\mathcal{M}}_t^{(N)}(\Psi_{t+r}\mathbf{1}); t \geq 0, N \in \mathbb{N}\}$ is tight in $\mathbb{H}^1(0,\infty)$.

PROOF. By definition (2.3.7) of Ψ_s and quadratic variation (2.3.6) of integrals with respect to martingale measure $\mathcal{M}^{(N)}$, and using the bound (4.1.18), for every $t, r \geq 0$ and $N \in \mathbb{N}$ we have

$$\mathbb{E}\left[\widehat{\mathcal{M}}_t^{(N)}(\Psi_{t+r}\mathbf{1})^2\right] = \mathbb{E}\left[\int_0^t \overline{\nu}_{t-s}^{(N)}\left(\frac{\overline{G}^2(s+r+\cdot)}{\overline{G}^2(\cdot)}\frac{g(\cdot)}{\overline{G}(\cdot)}\right)ds\right] \leq R_M(r). \tag{4.1.28}$$

Furthermore, by (4.1.26), $\widehat{\mathcal{M}}_t^{(N)}(\Psi_{t+\cdot}\mathbf{1})$ has a density $\widehat{\mathcal{M}}_t^{(N)}(\Psi_{t+\cdot}h)$, which satisfies an analogous relationship

$$\begin{aligned}\mathbb{E}\left[\widehat{\mathcal{M}}_t^{(N)}(\Psi_{t+r}h)^2\right] &= \mathbb{E}\left[\int_0^t \overline{\nu}_{t-s}^{(N)}\left(\frac{g^2(s+r+\cdot)}{\overline{G}^2(\cdot)}\frac{g(\cdot)}{\overline{G}(\cdot)}\right)ds\right] \\ &\leq H^2\mathbb{E}\left[\int_0^t \overline{\nu}_{t-s}^{(N)}\left(\frac{\overline{G}^2(s+r+\cdot)}{\overline{G}^2(\cdot)}\frac{g(\cdot)}{\overline{G}(\cdot)}\right)ds\right]\end{aligned}$$

where the inequality holds by Assumption III. Therefore, the bound (4.1.18) implies

$$\mathbb{E}\left[\widehat{\mathcal{M}}_t^{(N)}(\Psi_{t+r}\, h)^2\right] \leq H^2 R_M(r). \tag{4.1.29}$$

Likewise, by (4.1.27), $\widehat{\mathcal{M}}_t^{(N)}(\Psi_{t+\cdot}h)$ has a density $\widehat{\mathcal{M}}_t^{(N)}(\Psi_{t+\cdot}h_2)$, which, by Assumption III, satisfies

(4.1.30)

$$\mathbb{E}\left[\widehat{\mathcal{M}}_t^{(N)}(\Psi_{t+r}h_2)^2\right] = \mathbb{E}\left[\int_0^t \overline{\nu}_{t-s}^{(N)}\left(\frac{|g'(s+r+\cdot)|^2}{\overline{G}^2(\cdot)}\frac{g(\cdot)}{\overline{G}(\cdot)}\right)ds\right] \leq H_2^2 R_M(r).$$

The lemma follows from (4.1.28)-(4.1.30), the integrability of the function R_M guaranteed by Lemma 4.11, and the tightness criteria of Lemma 4.10, with $R_1 = R_M$, $R_2 = H^2 R_M$, and $R_3 = H_2^2 R_M$. □

4.1.3.3. *Tightness of Other Terms.* In the next four lemmas, we prove tightness of other sequences associated with each of the other components of $\widehat{Z}^{(N)}$ on the right-hand side of (2.3.20).

LEMMA 4.14. Suppose Assumptions II and III hold, and $V_0^{(N)}$ is distributed as $\pi_*^{(N)}$. Then the family $\{\widehat{Z}_0^{(N)}(t+\cdot); t \geq 0, N \in \mathbb{N}\}$ is tight in $\mathbb{H}^1(0,\infty)$.

PROOF. By the representation for $Z^{(N)}$ given in (2.2.12) and definition (2.2.7) of $\overline{Z}$, we have

$$\widehat{Z}_0^{(N)}(t+r) = \frac{1}{\sqrt{N}}\left(Z_0^{(N)}(t+r) - N\overline{Z}(t+r)\right) = \frac{1}{\sqrt{N}}\left(\nu_0^{(N)}(\Phi_{t+r}\mathbf{1}) - N\overline{Z}(t+r)\right).$$

Since $\nu_0^{(N)}$ has the same distribution as $\nu_*^{(N)}$ defined in (4.1.1), we have

$$\widehat{Z}_0^{(N)}(t+\cdot) \stackrel{(d)}{=} \frac{1}{\sqrt{N}}\sum_{j=1}^N \left(\frac{\overline{G}(a_j^*+t+\cdot)}{\overline{G}(a_j^*)} - \int_0^\infty \overline{G}(x+t+\cdot)dx\right),$$

where the $\{a_j^*; j \geq 1\}$ is i.i.d with p.d.f. $\overline{G}$. Hence, the sequence $\{\overline{G}(a_j^*+t+r)/\overline{G}(a_j^*)\}_{j\geq 1}$ is also i.i.d. with mean $\int_0^\infty \overline{G}(x+t+r)dx$, and,

$$\begin{aligned}\mathbb{E}\left[\widehat{Z}_0^{(N)}(t+r)^2\right] &= Var\left(\frac{\overline{G}(a_1^*+t+r)}{\overline{G}(a_1^*)}\right) \leq \mathbb{E}\left[\left|\frac{\overline{G}(a_1^*+t+r)}{\overline{G}(a_1^*)}\right|^2\right] \\ &= \int_0^\infty \frac{\overline{G}^2(x+t+r)}{\overline{G}(x)}\,dx \leq \int_{t+r}^\infty \overline{G}(x)dx.\end{aligned}$$

Therefore, for the function $R_1(r) \doteq \int_r^\infty \overline{G}(x)dx$ which is non-increasing and integrable by Assumption II.b, we have

$$\mathbb{E}\left[\widehat{Z}_0^{(N)}(t+r)^2\right] \leq R_1(t+r) \leq R_1(r). \tag{4.1.31}$$

Furthermore, since $\overline{G}$ has density g, $\widehat{Z}_0^{(N)}(t+\cdot)$ has density $(\widehat{Z}_0^{(N)})'(t+\cdot)$, where

$$(\widehat{Z}_0^{(N)})'(t+\cdot) \stackrel{(d)}{=} -\frac{1}{\sqrt{N}}\sum_{j=1}^N \left(\frac{g(a_j^*+t+\cdot)}{\overline{G}(a_j^*)} - \int_0^\infty g(x+t+\cdot)dx\right).$$

Similar to (4.1.31), using Assumption III, it can be shown that $\zeta'_{N,t}$ satisfies

(4.1.32)

$$\mathbb{E}\left[(\widehat{Z}_0^{(N)})'(t+r)^2\right] \leq \mathbb{E}\left[\left|\frac{g(a_1^*+t+r)}{\overline{G}(a_1^*)}\right|^2\right] \leq H^2\mathbb{E}\left[\left|\frac{\overline{G}(a_1^*+t+r)}{\overline{G}(a_1^*)}\right|^2\right] \leq H^2R_1(r).$$

Finally, by Assumption III, g has density g', and hence $(\widehat{Z}_0^{(N)})'(t+\cdot)$ has density $(\widehat{Z}_0^{(N)})''(t+\cdot)$, where

$$(\widehat{Z}_0^{(N)})''(t+\cdot) \stackrel{(d)}{=} -\frac{1}{\sqrt{N}}\sum_{j=1}^N \left(\frac{g'(a_j^*+t+\cdot)}{\overline{G}(a_j^*)} - \int_0^\infty g'(x+t+\cdot)dx\right),$$

which similarly satisfies

(4.1.33)

$$\mathbb{E}\left[(\widehat{Z}_0^{(N)})''(t+r)^2\right] \leq \mathbb{E}\left[\left|\frac{g'(a_1^*+t+r)}{\overline{G}(a_1^*)}\right|^2\right] \leq H_2^2\mathbb{E}\left[\left|\frac{\overline{G}(a_1^*+t+r)}{\overline{G}(a_1^*)}\right|^2\right] \leq H_2^2 R_1(r).$$

The result follows from (4.1.31)-(4.1.33) and the tightness criteria of Lemma 4.10 with $R_2 = H^2 R_1$, and $R_3 = H_2^2 R_1$. □

LEMMA 4.15. Suppose Assumptions I, II and IV.b hold and $V_0^{(N)}$ is distributed as $\pi_*^{(N)}$. Then, there exists $C_E < \infty$ such that for every (deterministic) absolutely continuous function f on $(0,\infty)$,

(4.1.34)

$$\mathbb{E}\left[\left|\int_0^t f(t-s)d\widehat{E}_s^{(N)}\right|^2\right] \leq C_E t|f(t)|^2 + C_E\left(\int_0^t s|f'(s)|ds\right)\left(\int_0^t |f'(s)|ds\right)$$

$$+ C_E\left(\int_0^t f(s)ds\right)^2. \tag{4.1.35}$$

Moreover, the family

$$\left\{\int_0^t \overline{G}(\cdot+t-s)d\widehat{E}_s^{(N)}; t\geq 0, N\in\mathbb{N}\right\}$$

is tight in $\mathbb{H}^1(0,\infty)$.

PROOF. To prove (4.1.34), fix $t \geq 0$ and f as in the statement of the lemma. Since $E^{(N)}$ is a stationary renewal process when $V_0^{(N)}$ is distributed as $\pi_*^{(N)}$, $\{E_s^{(N)}, 0 \leq s \leq t\}$ has the same distribution (as a process) as $\{E_t^{(N)} - E_{t-s}^{(N)}, 0 \leq s \leq t\}$. Hence, since $\overline{E} = \mathbf{Id}$ by (2.3.2), the process $\check{E}^{(N)}$ defined as

$$\check{E}_s^{(N)} \doteq \widehat{E}_t^{(N)} - \widehat{E}_{t-s}^{(N)} = \frac{1}{\sqrt{N}}\left(E_t^{(N)} - E_{t-s}^{(N)} - s\right)$$

has the same distribution as $\widehat{E}^{(N)}$ on $[0,t]$, and therefore, for every fixed $t \geq 0$,

$$\int_0^t f(t-s)d\widehat{E}_s^{(N)} = \int_0^t f(s)d\check{E}_s^{(N)} \stackrel{(d)}{=} \int_0^t f(s)d\widehat{E}_s^{(N)}.$$

Now, using integration by parts and the fact that $N = \lambda^{(N)} + \beta\sqrt{N}$ by (2.1.1), we have

$$\begin{aligned}\int_0^t f(s)d\widehat{E}_s^{(N)} &= \frac{1}{\sqrt{N}}\int_0^t f(s)dE_s^{(N)} - \frac{\lambda^{(N)}}{\sqrt{N}}\int_0^t f(s)ds - \beta\int_0^t f(s)ds \\ &= \frac{f(t)}{\sqrt{N}}\left(E_t^{(N)} - \lambda^{(N)}t\right) - \frac{1}{\sqrt{N}}\int_0^t f'(s)\left(E_s^{(N)} - \lambda^{(N)}s\right)ds - \beta\int_0^t f(s)ds.\end{aligned}$$

Then, combined with the inequality $(a+b+c)^2 \leq 8(a^2+b^2+c^2)$ and the Cauchy-Schwartz inequality, this implies

$$\begin{aligned}\mathbb{E}\left[\left|\int_0^t f(t-s)d\widehat{E}_s^{(N)}\right|^2\right] &\leq \frac{8|f(t)|^2}{N}\,\mathbb{E}\left[\left|E_t^{(N)} - \lambda^{(N)}t\right|^2\right] \\ &\quad + \frac{8}{N}\left(\int_0^t |f'(s)|\mathbb{E}\left[\left|E_s^{(N)} - \lambda^{(N)}s\right|^2\right]ds\right)\left(\int_0^t |f'(s)|ds\right) \\ &\quad + 8\beta^2\left(\int_0^t f(s)ds\right)^2. \end{aligned} \tag{4.1.36}$$

Moreover, recall by Assumption I that $E^{(N)}(\cdot) = \tilde{E}(\lambda^{(N)}\cdot)$ is a stationary renewal process, satisfying $\mathbb{E}[E_t^{(N)}] = \lambda^{(N)}t$, and let $\tilde{U}$ be the renewal function associated to $\tilde{E}$. By the equation (2.12) in [**100**, Theorem 7.2.4] for the variance of a stationary renewal process and Lorden's inequality $\tilde{U}(t) \leq t + \mathbb{E}[\tilde{u}_1^2]$ (e.g. see [**3**, Proposition V.6.2]), we have for all $s \in [0,t]$,

$$\begin{aligned}\mathbb{E}\left[\left|E_s^{(N)} - \lambda^{(N)}s\right|^2\right] &= \mathrm{Var}(E_s^{(N)}) = \mathrm{Var}(\tilde{E}(\lambda^{(N)}s)) \\ &= 2\int_0^{\lambda^{(N)}s}(\tilde{U}(v) - v - \frac{1}{2})dv \leq 2\mathbb{E}[\tilde{u}_1^2]\lambda^{(N)}s,\end{aligned} \tag{4.1.37}$$

where $\mathbb{E}[\tilde{u}_1^2]$ is finite by Assumption I. Substituting the above inequality and the inequality $\overline{\lambda}^{(N)} \leq 1$ in (4.1.36), we obtain (4.1.34) with $C_E = 8\max\{\mathbb{E}[\tilde{u}_1^2], \beta^2, 1\} < \infty$.

Now define $\zeta_{N,t}(r) \doteq \int_0^t \overline{G}(t-s+r)d\widehat{E}_s^{(N)}$ for $r \geq 0$. Substituting $f(\cdot) = \overline{G}(\cdot + r)$ in (4.1.34) and using the facts that $\overline{G}$ is decreasing, $g \leq H\overline{G}$ by Assumption

III, and $\int_r^\infty \overline{G}(s)ds \leq \int_0^\infty \overline{G}(s)ds = 1$ by Assumption II.a, we have

$$\begin{aligned}
\mathbb{E}\left[|\zeta_{N,t}(r)|^2\right] &\leq C_E \left\{ t\overline{G}^2(t+r) + \int_0^t sg(s+r)ds \int_0^t g(s+r)ds + \left(\int_0^t \overline{G}(s+r)ds\right)^2 \right\} \\
&\leq C_E \left\{ (t\overline{G}(t))\overline{G}(r) + H^2 \int_0^\infty s\overline{G}(s)ds \int_r^\infty \overline{G}(s)ds + \int_r^\infty \overline{G}(s)ds \right\}.
\end{aligned}$$

By Assumption II.b, $\overline{G}(t) = \mathcal{O}(t^{-(2+\epsilon)})$ for some $\epsilon > 0$ as $t \to \infty$, and hence, $t\overline{G}(t)$ is uniformly bounded by a finite constant c_1, the constant $c_2 \doteq \int_0^\infty s\overline{G}(s)ds$ is finite, and also the functions $\overline{G}$ and $\int_\cdot^\infty \overline{G}(x)dx$ are integrable (see Remark 2.3). Therefore, the function

$$R(r) \doteq c_1\overline{G}(r) + (1+c_2)\int_r^\infty \overline{G}(s)ds$$

is integrable and we have

$$\mathbb{E}\left[|\zeta_{N,t}(r)|^2\right] \leq C_E(1+H^2)R(r). \tag{4.1.38}$$

Moreover, since $\overline{G}$ has density $-g$, almost surely, $\zeta_{N,t}$ has density $\zeta'_{N,t}$ on $(0,\infty)$ with $\zeta'_{N,t}(r) = -\int_0^t g(t-s+r)d\widehat{E}_s^{(N)}$. Substituting $f = g(\cdot + r)$ in (4.1.34) and using the facts that $g \leq H\overline{G}$ and $|g'| \leq H_2\overline{G}$ by Assumption III, we obtain the analogous bound:

$$\begin{aligned}
\mathbb{E}\left[|\zeta'_{N,t}(r)|^2\right] &\leq C_E \left\{ tg^2(t+r) + \int_0^t s|g'(s+r)|ds \int_0^t |g'(s+r)|ds + \left(\int_0^t g(s+r)ds\right)^2 \right\} \\
&\leq C_E \left\{ H^2(t\overline{G}(t))\overline{G}(r) + H_2^2 \int_r^\infty \overline{G}(s)ds \int_0^\infty s\overline{G}(s)ds + H^2 \int_r^\infty \overline{G}(s)ds \right\} \\
&\leq C_E(H^2 + H_2^2)R(r).
\end{aligned} \tag{4.1.39}$$

Finally, since g has density g' by Assumption III, almost surely, $\zeta'_{N,t}$ has density $\zeta''_{N,t}$ on $(0,\infty)$ with $\zeta''_{N,t}(r) = -\int_0^t g'(t-s+r)d\widehat{E}_s^{(N)}$. Substituting $f = g'(\cdot + r)$ in

(4.1.34) and using the fact that $|g'| \leq H_2\overline{G}$ by Assumption III, we have

$$\begin{aligned}\mathbb{E}\left[|\zeta''_{N,t}(r)|^2\right] &\leq C_E\left\{t(g'(t+r))^2 + \left(\int_0^t (1+s)|g''(s+r)|ds\right)^2 + \left(\int_0^t |g'(s+r)|ds\right)^2\right\} \\ &\leq C_E\left\{H_2^2(t\overline{G}(t)) + \left(\int_0^\infty (1+s)|g''(s)|ds\right)^2 + H_2^2\left(\int_0^\infty \overline{G}(s)ds\right)^2\right\}.\end{aligned}$$

The first term on the right-hand side above is bounded, as discussed above, by $C_E H_2^2 c_1 < \infty$ using Assumption II.b. Also, since g'' is bounded and satisfies $g''(x) = \mathcal{O}(x^{-(2+\epsilon)})$ as $x \to \infty$ by Assumption IV.b, the integral in the second term is bounded by a constant $c_3 < \infty$. The last term is also bounded by $C_E H_2^2$, using Assumption II.a. Hence,

$$\mathbb{E}\left[|\zeta''_{N,t}(r)|^2\right] \leq C_E(H_2^2 c_1 + c_3 + H_2^2) < \infty. \tag{4.1.40}$$

The result follows from (4.1.38)-(4.1.40) and the tightness criteria in Lemma 4.10 with $R_1 = C_E(1+H^2)R$, $R_2 = C_E(H^2+H_2^2)R$ and $R_3 = C_E H_2^2(1+c_1)+c_2$. ☐

LEMMA 4.16. Suppose Assumptions I-III hold, and $V_0^{(N)}$ is distributed as $\pi_*^{(N)}$. Then, the family

$$\left\{-\widehat{X}_t^{(N)+}\,\overline{G}(\cdot); t \geq 0, N \in \mathbb{N}\right\}$$

is tight in $\mathbb{H}^1(0,\infty)$.

PROOF. Define $\zeta_{N,t}(r) \doteq -\widehat{X}_t^{(N)+}\overline{G}(r)$. Then, by Corollary 4.7, for every interval $I \subset (0,\infty)$, $N \in \mathbb{N}$ and $t \geq 0$,

(4.1.41)

$$\mathbb{P}\left\{\|\zeta_{N,t}\|_{\mathbb{L}^2(I)} > \lambda\right\} = \mathbb{P}\left\{\widehat{X}_t^{(N)+}\|\overline{G}\|_{\mathbb{L}^2(I)} > \lambda\right\} \leq \frac{\mathbb{E}\left[\widehat{X}_t^{(N)+}\right]\|\overline{G}\|_{\mathbb{L}^2(I)}}{\lambda} \leq \frac{C_Q\|\overline{G}\|_{\mathbb{L}^2(I)}}{\lambda}.$$

Since $\overline{G} \in \mathbb{L}^2(0,\infty)$ by Assumption II.a, for every $L > 0$, substituting $I = (0,L)$ in (4.1.41), we have

$$\lim_{\lambda\to\infty}\sup_{t\geq 0, N\in\mathbb{N}} \mathbb{P}\left\{\|\zeta_{N,t}\|_{\mathbb{L}^2(0,L)} > \lambda\right\} \leq \lim_{\lambda\to\infty}\frac{C_Q\|\overline{G}\|_{\mathbb{L}^2(0,\infty)}}{\lambda} = 0, \tag{4.1.42}$$

and for every $\lambda > 0$ replacing I with (L, ∞) in (4.1.41), we have

$$\text{(4.1.43)} \qquad \lim_{L\to\infty} \sup_{t\geq 0, N\in\mathbb{N}} \mathbb{P}\left\{\|\zeta_{N,t}\|_{\mathbb{L}^2(L,\infty)} > \lambda\right\} \leq \lim_{L\to\infty} \frac{C_Q\|\overline{G}\|_{\mathbb{L}^2(L,\infty)}}{\lambda} = 0.$$

Moreover, since $\overline{G}$ has density $-g$, $\zeta_{N,t}$ has density $\zeta'_{N,t}(r) = \widehat{X}_t^{(N)+}g(r)$. By Assumption III and definition of the constant H in Remark 2.2, $|\zeta'_{N,t}| \leq H|\zeta_{N,t}|$, and therefore, by (4.1.42) and (4.1.43), for every $L > 0$ and $\delta > 0$, we have

(4.1.44)

$$\lim_{\lambda\to\infty} \sup_{t\geq 0, N\in\mathbb{N}} \mathbb{P}\left\{\|\zeta'_{N,t}\|_{\mathbb{L}^2(0,L)} > \lambda\right\} = 0, \quad \lim_{L\to\infty} \sup_{t\geq 0, N\in\mathbb{N}} \mathbb{P}\left\{\|\zeta'_{N,t}\|_{\mathbb{L}^2(L,\infty)} > \delta\right\} = 0$$

Finally, by Assumption III and definition of the constant H_2 in Remark 2.2, g has density g' with $|g'| \leq H_2\overline{G}$, and hence, $\zeta'_{N,t}$ has density $\zeta''_{N,t}(r) = \widehat{X}_t^{(N)+}g'(r)$ which satisfies $|\zeta''_{N,t}| \leq H_2|\zeta_{N,t}|$. Therefore, by (4.1.42), for every $L > 0$ we have

$$\text{(4.1.45)} \qquad \lim_{\lambda\to\infty} \sup_{t\geq 0, N\in\mathbb{N}} \mathbb{P}\left\{\|\zeta''_{N,t}\|_{\mathbb{L}^2(0,L)} > \lambda\right\} = 0.$$

The result holds by (4.1.42)-(4.1.45), and the tightness criteria of Proposition 4.9. □

LEMMA 4.17. Suppose Assumptions I-III and IV.b hold, and $V_0^{(N)}$ is distributed as $\pi_*^{(N)}$. Then, the family

$$\left\{\int_0^t \widehat{X}_s^{(N)+}g(\cdot + t - s)ds; t \geq 0, N \in \mathbb{N}\right\}$$

is tight in $\mathbb{H}^1(0,\infty)$.

PROOF. Define $\zeta_{N,t}(r) \doteq \int_0^t \widehat{X}_s^{(N)+}g(t + r - s)ds$. Since $g \leq H\overline{G}$ by Assumption III and definition of H in Remark 2.2, and $\overline{G}$ is decreasing,

$$|\zeta_{N,t}(r)| \leq H\int_0^t \widehat{X}_s^{(N)+}\overline{G}(t + r - s)ds \leq H\overline{G}(r)^{1/2}\int_0^t \widehat{X}_{t-s}^{(N)+}\ \overline{G}(s)^{1/2}\, ds.$$

The bound (4.1.10) of Corollary 4.7 then implies that for every interval $I \subset (0,\infty)$,

$$\begin{aligned}\mathbb{P}\left\{\left\|\int_0^t \widehat{X}_s^{(N)+}\overline{G}(\cdot + t - s)ds\right\|_{\mathbb{L}^2(I)} > \lambda\right\} &\leq \frac{1}{\lambda}\|\overline{G}^{\ 1/2}\|_{\mathbb{L}^2(I)}\int_0^t \mathbb{E}\left[\widehat{X}_{t-s}^{(N)+}\right]\overline{G}(s)^{1/2}\, ds \\ \text{(4.1.46)} \qquad &\leq \frac{1}{\lambda}C_Q\|\overline{G}^{\ 1/2}\|_{\mathbb{L}^2(I)}\|\overline{G}^{\ 1/2}\|_{\mathbb{L}^1(0,\infty)}.\end{aligned}$$

By Assumption II.b, $\overline{G}(r) = \mathcal{O}(r^{-(2+\epsilon)})$ as $r \to \infty$, and $\overline{G}$ is also bounded by one, and thus, $\overline{G}^{\ 1/2} \in \mathbb{L}^1(0,\infty) \cap \mathbb{L}^2(0,\infty)$. Therefore, for every $L > 0$, substituting $I = (0, L)$ in (4.1.46), we have

(4.1.47)
$$\lim_{\lambda\to\infty} \sup_{t\geq 0, N\in\mathbb{N}} \mathbb{P}\left\{\|\zeta_{N,t}\|_{\mathbb{L}^2(0,L)} > \lambda\right\} \leq \lim_{\lambda\to\infty} \frac{1}{\lambda} HC_Q \|\overline{G}^{\ 1/2}\|_{\mathbb{L}^2(0,\infty)} \|\overline{G}^{\ 1/2}\|_{\mathbb{L}^1(0,\infty)} = 0,$$

and for every $\lambda > 0$, replacing I with (L,∞) in (4.1.46), we have

(4.1.48)
$$\lim_{L\to\infty} \sup_{t\geq 0, N\in\mathbb{N}} \mathbb{P}\left\{\|\zeta_{N,t}\|_{\mathbb{L}^2(L,\infty)} > \lambda\right\} \leq \lim_{L\to\infty} \frac{1}{\lambda} HC_Q \|\overline{G}^{\ 1/2}\|_{\mathbb{L}^2(L,\infty)} \|\overline{G}^{\ 1/2}\|_{\mathbb{L}^1(0,\infty)} = 0.$$

Furthermore, by Assumption III and Remark 2.2, g has a density g' with $|g'| \leq H_2\overline{G}$, and hence, $\zeta_{N,t}$ has density $\zeta'_{N,t}$, where

$$|\zeta'_{N,t}(r)| = \left|\int_0^t \widehat{X}_s^{(N)+}\ g'(t+r-s)ds\right| \leq H_2 \int_0^t \widehat{X}_s^{(N)+}\ \overline{G}(t+r-s)ds.$$

Using the above inequality and (4.1.46), exactly analogous to (4.1.47) and (4.1.48), for every $L < \infty$ and $\delta > 0$ we have

(4.1.49)
$$\lim_{\lambda\to\infty} \sup_{t\geq 0, N\in\mathbb{N}} \mathbb{P}\left\{\|\zeta'_{N,t}\|_{\mathbb{L}^2(0,L)} > \lambda\right\} = 0, \quad \lim_{L\to\infty} \sup_{t\geq 0, N\in\mathbb{N}} \mathbb{P}\left\{\|\zeta'_{N,t}\|_{\mathbb{L}^2(L,\infty)} > \delta\right\} = 0.$$

Finally, by Assumption IV.b, g' has bounded density g'', and hence, $\zeta'_{N,t}$ has density

$$\zeta''_{N,t}(r) = \int_0^t \widehat{X}_s^{(N)+} g''(t-s+r)ds.$$

By the same assumption, g'' is uniformly bounded by a constant (say) $C_{g''}$ and satisfies $g''(x) = \mathcal{O}(x^{-(2+\epsilon)})$ as $x \to \infty$, that is, there exist $T \in (0,\infty)$ and $C_T < \infty$ such that $|g''(x)| \leq C_T x^{-(2+\epsilon)}$ for all $x \geq T$. Therefore, for $r \geq 0$,

$$|\zeta''_{N,t}(r)| \leq \int_0^t \widehat{X}_s^{(N)+}|g''(t-s+r)|ds \leq C_{g''}\int_0^{T\wedge t} \widehat{X}_s^{(N)+}ds + C_T \int_{T\wedge t}^t \widehat{X}_{t-s}^{(N)+}\ s^{-(2+\epsilon)}ds.$$

Another use of the bound (4.1.10) shows that for every $L < \infty$,

$$\begin{aligned}\mathbb{P}\left\{\|\zeta''_{N,t}\|_{\mathbb{L}^2(0,L)} > \lambda\right\} &\leq \frac{1}{\lambda}\mathbb{E}\left[\|\zeta''_{N,t}\|_{\mathbb{L}^2(0,L)}\right] \\ &\leq \frac{\sqrt{L}}{\lambda}\mathbb{E}\left[\|\zeta''_{N,t}\|_{L}\right] \\ &\leq \frac{C_Q\sqrt{L}}{\lambda}\left(C_{g''}T + C_T\int_0^\infty s^{-(2+\epsilon)}ds\right),\end{aligned}$$

and hence,

$$\lim_{\lambda\to\infty}\sup_{t\geq 0, N\in\mathbb{N}} \mathbb{P}\left\{\|\zeta''_{N,t}\|_{\mathbb{L}^2(0,L)} > \lambda\right\} = 0. \tag{4.1.50}$$

The lemma now follows from (4.1.47)-(4.1.50), and the criteria of Lemma 4.10 for tightness in $\mathbb{H}^1(0,\infty)$. $\square$

PROOF OF PROPOSITION 4.8. The result follows from the representation (2.3.20) of $\widehat{Z}_t^{(N)}$ and Lemmas 4.13-4.17. $\square$

4.1.4. Proof of Theorem 4.2.

PROOF OF THEOREM 4.2. First, we claim that the family $\{\widehat{X}_t^{(N)}; t \geq 0, N \in \mathbb{N}\}$ is tight in $\mathbb{R}$. It clearly suffices to establish tightness of $\{\widehat{X}_t^{(N)+}; t \geq 0, N \in \mathbb{N}\}$ and $\{\widehat{X}_t^{(N)-}; t \geq 0, N \in \mathbb{N}\}$. Tightness of the first family follows from Corollary 4.7, which shows that $\{\widehat{X}^{(N)+}\}$ is (uniformly) bounded in $\mathbb{L}^1$. To see the tightness of $\{\widehat{X}^{(N)-}\}$, note that since $Y_t^{(N)}$ takes values in $\mathcal{Y}$ (see Theorem 2.7), it follows that for every $t \geq 0$ and $N \in \mathbb{N}$, $\widehat{X}_t^{(N)-} = -\widehat{Z}_t^{(N)}(0) = -\Pi_0(\widehat{Z}_t^{(N)})$, where $\Pi_0 : f \in \mathbb{H}^1(0,\infty) \mapsto f(0)$. By Lemma 2.5, the embedding $I : \mathbb{H}^1(0,\infty) \mapsto \mathbb{C}_{\mathbb{R}}[0,\infty)$ is continuous, and hence, the mapping Π_0 is continuous from $\mathbb{H}^1(0,\infty)$ to $\mathbb{R}$. Therefore, the tightness of $\{\widehat{X}_t^{(N)-}\}$ follows from the tightness of $\{\widehat{Z}_t^{(N)}\}$ in $\mathbb{H}^1(0,\infty)$ proved in Proposition 4.8, and the continuous mapping theorem. This proves the claim.

By another application of Proposition 4.8 and the claim above, the family $\{\widehat{Y}_t^{(N)} = (\widehat{X}_t^{(N)}, \widehat{Z}_t^{(N)}); t \geq 0, N \in \mathbb{N}\}$ is tight in $\mathcal{Y}$, when $V_0^{(N)}$ is initialized at $\pi_*^{(N)}$. Finally, by Theorem 2.7, $\mathscr{L}aw(\widehat{Y}_t^{(N)})$ converges in distribution to $\widehat{\pi}^{(N)}$, as $t \to \infty$. The result then follows from Remark 4.3. $\square$

4.2. Convergence of Subsequential limits

We now prove the convergence of the sequence $\{\widehat{\pi}^{(N)}\}_{N\in\mathbb{N}}$ of stationary distributions. Since the sequence is tight by Theorem 4.2, it remains to uniquely characterize subsequential limits of $\widehat{\pi}^{(N)}$. We prove that every subsequential limit of $\widehat{\pi}^{(N)}$ is an invariant distribution of the limit process Y introduced in Section 3.1.

We first show that under certain assumptions on the sequence of initial conditions, finite-dimensional projections of $\widehat{Y}_t^{(N)}$ converges to those of Y (Proposition 4.19). This is analogous to the path-wise C.L.T. result obtained in **[62]**, but does not directly follow from it since the state representation used there is different. However, we make use of a convergence result from **[62]** which is recalled in Section 4.2.1. The argument in the proof of Theorem 4.1 involves invoking Proposition 4.19 in the case when each Markovian state process $V^{(N)}$ is initialized according to its stationary distributions $V_\infty^{(N)}$. We show in Section 4.2.2 that the sequence $\{V_\infty^{(N)}\}$, as initial condition, satisfies the assumptions of Proposition 4.19. The proof of Theorem 4.1 is concluded at the end of Section 4.2.3.

4.2.1. A Central Limit Theorem. Recall from Section 3.1 that $\{B_t; t \geq 0\}$ is a Brownian motion and $\mathcal{M}$ is a martingale measure, with covariance functional defined in (3.1.1), that is independent of B. Also, define $E_t \doteq B_t - \beta t$ for $t \geq 0$. The following result from **[62]** shows that $E^{(N)}$ and $\mathcal{M}^{(N)}$ are asymptotic independent and converge in distribution to E and $\mathcal{M}$, respectively.

THEOREM 4.18. Suppose Assumptions I-III hold, and the sequence of initial conditions $\{V_0^{(N)} = (R_0^{(N)}, X_0^{(N)}, \nu_0^{(N)})\}$ satisfies the condition (2.3.4). Moreover, for $N \in \mathbb{N}$, define $\widehat{Y}_0^{(N)} \doteq \hat{\Psi}(V_0^{(N)})$ to be the corresponding initial condition for the centered and scaled process $\widehat{Y}^{(N)}$, where $\widehat{\Psi}$ is defined in (2.2.14), and assume that there exits a $\mathcal{Y}$-valued random variable $Y_0 = (X_0, Z_0)$ such that, as $N \to \infty$,

$$\widehat{Y}_0^{(N)} \Rightarrow Y_0, \qquad \text{in } \mathcal{Y}. \tag{4.2.1}$$

Then, for every $t \geq 0$, $k \in \mathbb{N}$, and bounded continuous functions $\varphi_1, ..., \varphi_k$ on $[0,\infty) \times [0,\infty)$, we have

$$\begin{aligned}
&\left(\widehat{E}^{(N)}, \widehat{X}_0^{(N)}, \widehat{Z}_0^{(N)}, \widehat{\mathcal{H}}^{(N)}(\mathbf{1}), \widehat{\mathcal{M}}_t^{(N)}(\varphi_1), ..., \widehat{\mathcal{M}}_t^{(N)}(\varphi_k)\right) \\
&\qquad\qquad \Rightarrow (E, X_0, Z_0, \mathcal{H}(\mathbf{1}), \mathcal{M}_t(\varphi_1), ..., \mathcal{M}_t(\varphi_k)), \qquad (4.2.2)
\end{aligned}$$

weakly in $\mathbb{D}_{\mathbb{R}}[0,\infty) \times \mathbb{R} \times \mathbb{H}^1 \times \mathbb{D}_{\mathbb{R}}[0,\infty) \times \mathbb{R}^k$, where $\mathbb{D}_{\mathbb{R}}[0,\infty)$ is equipped with the topology of uniform convergence on compact sets.

PROOF. We first explain why assumptions 1-4 in **[62]** are satisfied under our assumptions. As discussed before, Assumption 1 of **[62]** holds by condition (2.3.4) and Proposition 2.10.a, and (condition a. of) Assumption 2 of **[62]** holds because h is bounded by Assumption III. Moreover, by Assumption I, $E^{(N)}$ is a renewal process with inter-arrival times $\{u_n^{(N)}; n \geq 1\}$ with $u_n^{(N)} = \tilde{u}_n/\lambda^{(N)}$, where $\{\tilde{u}_n; n \in \mathbb{N}\}$ is an i.i.d. sequence with mean 1 and variance $\tilde{\sigma}_A^2$. Hence, $u_1^{(N)}$ satisfies

$$\mathbb{E}\left[u_1^{(N)}\right] = \frac{\mathbb{E}\left[\tilde{u}_1\right]}{\lambda^{(N)}} = \frac{1}{\lambda^{(N)}}, \quad \mathrm{Var}(u_1^{(N)}) = \frac{\mathrm{Var}(\tilde{u}_1)}{(\lambda^{(N)})^2} = \frac{\tilde{\sigma}_A^2}{(\lambda^{(N)})^2},$$

and since $\mathbb{E}[\tilde{u}_1^2] < \infty$ and $\lambda^{(N)}/\sqrt{N} \to \infty$ as $N \to \infty$, the Lindeberg condition

$$\lim_{N\to\infty} N^2 \mathbb{E}\left[(u_1^{(N)})^2 \mathbb{1}_{\{u_1^{(N)}\sqrt{N} > \epsilon\}}\right] = \lim_{N\to\infty} \frac{N^2}{(\lambda^{(N)})^2} \mathbb{E}\left[(\tilde{u}_1)^2 \mathbb{1}_{\{\tilde{u}_1 > \frac{\lambda^{(N)}\epsilon}{\sqrt{N}}\}}\right] = 0,$$

holds. Therefore, Assumption 3 of **[62]** is satisfied. Finally, by **[62**, Remark 5.2**]**, Assumption 4 of **[62]** also follows from boundedness of h, which holds by Assumption III.

The convergence result (4.2.2) then follows from (4.2.1), Corollary 8.7 in **[62]** and the second display in the proof of that corollary with $\tilde{f}_j$ replaced by φ_j and $f_1 \doteq \mathbf{1}$. Note that Assumption 5 in **[62]** is not required for other assertions to hold, as we are not claiming the convergence of the distribution-valued process $\widehat{\nu}^{(N)}$. $\square$

Next, we use the result of Theorem 4.18 to establish the convergence of sequence of finite-dimensional projections of $\widehat{Y}^{(N)}$. Recall that $\mathbb{H}^1(0,\infty)$ is a separable Hilbert space with the inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}^1}$ defined in (1.4.1). Let $\{e_k; k \in \mathbb{N}\}$

be a sequence of functions in $\mathbb{H}^1(0,\infty)$ with compact support, that form a countable orthonormal basis for $\mathbb{H}^1(0,\infty)$ (e.g., Daubechies Wavelets, see [**97**, Section 9.2]), and let e_j' denote the density of $e_j, j \in \mathbb{N}$.

PROPOSITION 4.19. Suppose Assumptions I-III hold, and the sequence of initial conditions $\{V_0^{(N)} = (R_0^{(N)}, X_0^{(N)}, \nu_0^{(N)})\}$ satisfies the condition (2.3.4). Moreover, for $N \in \mathbb{N}$, define $\widehat{Y}_0^{(N)} \doteq \hat{\Psi}(V_0^{(N)})$ to be the corresponding initial condition for the centered and scaled process $\widehat{Y}^{(N)}$, where $\widehat{\Psi}$ is defined in (2.2.14), and assume that there exits a $\mathcal{Y}$-valued random variable $Y_0 = (X_0, Z_0)$ such that, as $N \to \infty$,

$$\widehat{Y}_0^{(N)} \Rightarrow Y_0, \qquad \text{in } \mathcal{Y}. \tag{4.2.3}$$

Also, let $Y = (X, Z)$ be the limit process associated with initial condition Y_0. Then, for every $t \geq 0$ and $k \in \mathbb{N}$,

$$\left(\widehat{X}_t^{(N)}, \langle \widehat{Z}_t^{(N)}, e_1\rangle_{\mathbb{H}^1}, ..., \langle \widehat{Z}_t^{(N)}, e_k\rangle_{\mathbb{H}^1}\right) \Rightarrow \left(X_t, \langle Z_t, e_1\rangle_{\mathbb{H}^1}, ..., \langle Z_t, e_k\rangle_{\mathbb{H}^1}\right), \tag{4.2.4}$$

as $N \to \infty$.

PROOF. For this proof, we use the convention that all continuity results on $\mathbb{D}_{\mathbb{R}}[0,\infty)$ are with respect to the topology of uniform convergence on compact sets.

Fix $t \geq 0$ and $k \in \mathbb{N}$. First, recall that by equations (2.3.16) and (3.1.11) for $(K^{(N)}, X^{(N)})$ and (K, X), respectively, we have

$$(\widehat{K}^{(N)}, \widehat{X}^{(N)}) = \Lambda(\widehat{E}^{(N)}, \widehat{X}_0^{(N)}, \widehat{Z}_0^{(N)} - \widehat{\mathcal{H}}^{(N)}(\mathbf{1})), \qquad (K, X) = \Lambda(E, X_0, Z_0 - \mathcal{H}(\mathbf{1})), \tag{4.2.5}$$

where, by Lemma 2.13, the centered many-server mapping $\Lambda : \mathbb{D}_{\mathbb{R}}^0[0,\infty) \times \mathbb{R} \times \mathbb{D}_{\mathbb{R}}[0,\infty) \mapsto \mathbb{D}_{\mathbb{R}}^0[0,\infty) \times \mathbb{D}_{\mathbb{R}}[0,\infty)$ is continuous.

Next, we compute the projections of $\widehat{Z}_t^{(N)}$ and Z_t onto the directions $e_1, ..., e_k$. Recall the representations (2.3.18) of $\widehat{Z}_t^{(N)}$ and (2.3.19) of $(\widehat{Z}^{(N)})'$. By definition (2.3.7) of $\Psi_s, s \geq 0$, Z_t defined in (3.1.12) admits the representation

$$Z_t(r) \doteq Z_0(t+r) - \mathcal{M}_t(\Psi_{t+r}\mathbf{1}) + K_t\overline{G}(r) - \int_0^t K_s g(t-s+r)ds, \tag{4.2.6}$$

and by equation (3.3.1) of Proposition 3.17, Z_t has density Z_t' where

$$Z_t'(r) \doteq Z_0'(t+r) + \mathcal{M}_t(\Psi_{t+r}h) - K_t g(r) - \int_0^t K_s g'(t-s+r)ds. \tag{4.2.7}$$

First, since the translation mapping $f \mapsto f(t+\cdot)$ is continuous from $\mathbb{L}^2(0,\infty)$ to itself, and for any $u \in \mathbb{L}^2$, the projection mapping $f \mapsto \langle f, u\rangle_{\mathbb{L}^2}$ is continuous from $\mathbb{L}^2(0,\infty)$ to $\mathbb{R}$, there exist continuous mappings $F_j^1, F_j^2 : \mathbb{L}^2(0,\infty) \mapsto \mathbb{R}, j = 1, ..., k$, such that

$$\langle \widehat{Z}_0^{(N)}(t+\cdot), e_j\rangle_{\mathbb{L}^2} = F_j^1(\widehat{Z}_0^{(N)}), \quad \langle Z_0(t+\cdot), e_j\rangle_{\mathbb{L}^2} = F_j^1(Z_0), \tag{4.2.8}$$

and

$$\langle (\widehat{Z}_0^{(N)})'(t+\cdot), e_j'\rangle_{\mathbb{L}^2} = F_j^2((\widehat{Z}_0^{(N)})'), \quad \langle Z_0'(t+\cdot), e_j'\rangle_{\mathbb{L}^2} = F_j^2(Z_0'). \tag{4.2.9}$$

Next, for every $j = 1, \dots, k$ define the function $\varphi_j : [0,\infty) \times [0,\infty) \mapsto \mathbb{R}$ as

$$\varphi_j(x,s) \doteq \langle \Psi_{t+\cdot}\mathbf{1}(x,s), e_j\rangle_{\mathbb{L}^2} = \int_0^\infty \Psi_{t+r}\mathbf{1}(x,s)e_j(r)\, dr = \int_0^\infty \frac{\overline{G}(x+t-s+r)}{\overline{G}(x)} e_j(r)dr, \tag{4.2.10}$$

which is bounded by $\|e_j\|_{\mathbb{L}^1} < \infty$ (recall that e_j has compact support by choice). Also, by the continuity of $\overline{G}$ and since $1/\overline{G}$ is bounded on finite intervals, the dominated convergence theorem shows that φ_j is continuous. Since $\varphi_j \in \mathbb{C}_b([0,\infty) \times [0,\infty))$, by the stochastic Fubini theorem for orthogonal martingale measures $\widehat{\mathcal{M}}^{(N)}$ (see e.g. [**98**, Theorem 2.6]), we have

$$\langle \widehat{\mathcal{M}}_t^{(N)}(\Psi_{t+\cdot}\mathbf{1}), e_j\rangle_{\mathbb{L}^2} = \widehat{\mathcal{M}}_t^{(N)}\left(\int_0^\infty \Psi_{t+r}\mathbf{1}e_j(r)dr\right) = \widehat{\mathcal{M}}_t^{(N)}(\varphi_j). \tag{4.2.11}$$

Similarly, the function $\tilde{\varphi}_j : [0,\infty) \times [0,\infty) \mapsto \mathbb{R}$ defined as

$$\tilde{\varphi}_j(x,s) \doteq \langle \Psi_{t+\cdot}h(x,s), e_j'\rangle_{\mathbb{L}^2} = \int_0^\infty \Psi_{t+r}h(x,s)e_j'(r)\, dr = \int_0^\infty \frac{g(x+t-s+r)}{\overline{G}(x)} e_j'(r)dr, \tag{4.2.12}$$

is also continuous and bounded (by $H\|e_j'\|_{\mathbb{L}^1} < \infty$) by Assumption III. By another use of the stochastic Fubini theorem for orthogonal martingales,

$$\langle \widehat{\mathcal{M}}_t^{(N)}(\Psi_{t+\cdot}h), e_j'\rangle_{\mathbb{L}^2} = \widehat{\mathcal{M}}_t^{(N)}\left(\int_0^\infty \Psi_{t+r}he_j'(r)dr\right) = \widehat{\mathcal{M}}_t^{(N)}(\tilde{\varphi}_j). \tag{4.2.13}$$

Similarly, replacing $\widehat{\mathcal{M}}^{(N)}$ with $\mathcal{M}_t$, we have

$$\langle \mathcal{M}_t(\Psi_{t+\cdot}\mathbf{1}), e_j\rangle_{\mathbb{L}^2} = \mathcal{M}_t(\varphi_j), \qquad \langle \mathcal{M}_t(\Psi_{t+\cdot}h), e_j'\rangle_{\mathbb{L}^2} = \mathcal{M}_t(\tilde{\varphi}_j). \tag{4.2.14}$$

Finally, since $\widehat{K}^{(N)}$ is almost surely bounded on finite intervals, for every $j = 1, \dots, k$, by Fubini's theorem,

$$\langle \widehat{K}_t^{(N)}\overline{G}(\cdot) - \int_0^t \widehat{K}_s^{(N)} g(t-s+\cdot)ds, e_j\rangle_{\mathbb{L}^2} = \widehat{K}_t^{(N)}\langle \overline{G}, e_j\rangle_{\mathbb{L}^2} - \int_0^t \widehat{K}_s^{(N)}\langle g(t-s+\cdot), e_j\rangle_{\mathbb{L}^2} ds.$$

By assumptions II.a and III, both $\overline{G}$ and g are in $\mathbb{L}^2(0,\infty)$ (see Remark 2.3), and hence the constant $\langle \overline{G}, e_j\rangle_{\mathbb{L}^2}$ is finite, and the function $s \mapsto \langle g(t-s+\cdot), e_j\rangle$ is bounded and continuous on $[0,t]$. Moreover, for every $t \geq 0$ and continuous and bounded function u, the mapping $f \mapsto \int_0^t f(s)u(s)ds$ from $\mathbb{D}_{\mathbb{R}}[0,\infty)$ to $\mathbb{R}$ is continuous. Therefore, by equation (4.2.5), there exist continuous functions $\tilde{F}_j^3 : \mathbb{D}_{\mathbb{R}}[0,\infty) \mapsto \mathbb{R}$ and $F_j^3 : \mathbb{D}_{\mathbb{R}}[0,\infty) \times \mathbb{R} \times \mathbb{D}_{\mathbb{R}}[0,\infty) \mapsto \mathbb{R}$, such that

$$\begin{aligned}\langle \widehat{K}_t^{(N)}\overline{G}(\cdot) - \int_0^t \widehat{K}_s^{(N)} g(t-s+\cdot)ds, e_j\rangle_{\mathbb{L}^2} &= \tilde{F}_j^3(\widehat{K}^{(N)}) \\ &= F_j^3(\widehat{E}^{(N)}, \widehat{X}_0^{(N)}, \widehat{Z}_0^{(N)} - \widehat{\mathcal{H}}^{(N)}(\mathbf{1})).\end{aligned} \tag{4.2.15}$$

By the same argument,

$$\langle K_t\overline{G}(\cdot) - \int_0^t K_s g(t-s+\cdot)ds, e_j\rangle_{\mathbb{L}^2} = F_j^3(E, X_0, Z_0 - \mathcal{H}(\mathbf{1})). \tag{4.2.16}$$

Similarly, by Assumption III, g' is also in $\mathbb{L}^2$, and hence the constant $\langle g, e_j'\rangle$ is finite and the mapping $s \mapsto \langle g'(t-s+\cdot)e_j'\rangle$ is bounded and continuous on $[0,t]$. Therefore, there exists a continuous function $F_j^4 : \mathbb{D}_{\mathbb{R}}[0,\infty) \times \mathbb{R} \times \mathbb{D}_{\mathbb{R}}[0,\infty) \mapsto \mathbb{R}$ such that

$$\langle \widehat{K}_t^{(N)} g(\cdot) + \int_0^t \widehat{K}_s^{(N)} g'(t-s+\cdot)ds, e_j'\rangle_{\mathbb{L}^2} = F_j^4(\widehat{E}^{(N)}, \widehat{X}_0^{(N)}, \widehat{Z}_0^{(N)} - \widehat{\mathcal{H}}^{(N)}(\mathbf{1})), \tag{4.2.17}$$

and

$$\langle K_t g(\cdot) + \int_0^t K_s g'(t-s+\cdot)ds, e_j'\rangle_{\mathbb{L}^2} = F_j^4(E, X_0, Z_0 - \mathcal{H}(\mathbf{1})). \tag{4.2.18}$$

Therefore, combining equation (4.2.5), representations (2.3.18) of $\widehat{Z}_t^{(N)}$, (2.3.19) of $(\widehat{Z}^{(N)})'$, (4.2.6) of Z_t and (4.2.7) of Z_t', equations (4.2.8)-(4.2.18), and the fact that addition is continuous on $\mathbb{D}_{\mathbb{R}}[0,\infty)$, we conclude that there exists a continuous function

$$F : \mathcal{I} \doteq \mathbb{R} \times \mathbb{H}^1(0,\infty) \times \mathbb{D}_{\mathbb{R}}[0,\infty) \times \mathbb{D}_{\mathbb{R}}[0,\infty) \times \mathbb{R}^{2k} \mapsto \mathbb{R}^{k+1},$$

such that

$$\left(X_t^{(N)}, \langle \widehat{Z}_t^{(N)}, e_1\rangle_{\mathbb{H}^1}, ..., \langle \widehat{Z}_t^{(N)}, e_k\rangle_{\mathbb{H}^1}\right) = F\left(\widehat{I}^{(N)}\right) \tag{4.2.19}$$

and

$$(X_t, \langle Z_t, e_1\rangle_{\mathbb{H}^1}, ..., \langle Z_t, e_k\rangle_{\mathbb{H}^1}) = F(I), \tag{4.2.20}$$

where

$$\widehat{I}^{(N)} = \left(\widehat{X}_0^{(N)}, \widehat{Z}_0^{(N)}, \widehat{E}^{(N)}, \widehat{\mathcal{H}}^{(N)}(\mathbf{1}), \widehat{\mathcal{M}}_t^{(N)}(\varphi_j), \widehat{\mathcal{M}}^{(N)}(\tilde{\varphi}_k); j = 1, \ldots, k\right),$$

and

$$I = \left(X_0, Z_0, E, \mathcal{H}(\mathbf{1}), \mathcal{M}_t(\varphi_j), \mathcal{M}_t(\tilde{\varphi}_j); j = 1, \ldots k\right).$$

Finally, by Theorem 4.18 with φ_j and $\tilde{\varphi}_j; j = 1, ..., k$, defined in (4.2.10) and (4.2.12), we have

$$\widehat{I}^{(N)} \Rightarrow I, \tag{4.2.21}$$

in $\mathcal{I}$. The result in (4.2.4) follows from the (4.2.19), (4.2.20), (4.2.21), and an application of the continuous mapping theorem. $\square$

4.2.2. Fluid-Scaling Limit of $V_\infty^{(N)}$. In order to use Proposition 4.19 with the sequence of initial conditions $V_0^{(N)} = V_\infty^{(N)}, N \in \mathbb{N}$, we need to verify the conditions of that proposition. Recall that for every $N \in \mathbb{N}$,

$$V_\infty^{(N)} = (R_\infty^{(N)}, X_\infty^{(N)}, \nu_\infty^{(N)}), \quad \text{and} \quad \overline{V}_\infty^{(N)} = (\overline{R}_\infty^{(N)}, \overline{X}_\infty^{(N)}, \overline{\nu}_\infty^{(N)}) = \frac{1}{N}(R_\infty^{(N)}, X_\infty^{(N)}, \nu_\infty^{(N)}).$$

We start with the following lemma.

LEMMA 4.20. Suppose Assumptions I-III hold. Then, the sequence $\{\overline{\nu}_\infty^{(N)}\}_{N\in\mathbb{N}}$ is tight in $\mathbb{M}_F[0,\infty)$.

PROOF. For every $N \in \mathbb{N}$, consider the process $\{V_t = (R_t^{(N)}, X_t^{(N)}, \nu_t^{(N)}), t \geq 0\}$ initialized at $V_*^{(N)}$. For every $t \geq 0$, $\overline{\nu}_t^{(N)}$ is a sub-probability measure and hence

$$\sup_{N \in \mathbb{N},\, t \geq 0} \mathbb{E}\left[\overline{\nu}_t^{(N)}(\mathbf{1})\right] \leq 1. \tag{4.2.22}$$

Now for every $c \geq 0$, consider the piece-wise linear function f_c defined as

$$f_c(x) \doteq \begin{cases} 0 & \text{if } 0 \leq x < c, \\ \frac{1}{c}(x-c) & \text{if } c \leq x < 2c, \\ 1 & \text{if } x \geq 2c, \end{cases}$$

which is non-negative, bounded and absolutely continuous with density $f_c' = \frac{1}{c}\mathbb{1}_{(c,2c)}$ and satisfies $f_c(0) = 0$ and $\mathbb{1}_{[2c,\infty)} \leq f_c \leq \mathbb{1}_{[c,\infty)}$. By (4.1.17) in Lemma 4.11 with f replaced by f_c, we have

$$\begin{aligned} \mathbb{E}\left[\overline{\nu}_t^{(N)}(f_c)\right] &\leq \int_0^\infty f_c(x)\overline{G}(x)dx + C_Q \int_0^t (f_c'(x)\overline{G}(x) + f_c(x)g(x))dx \\ &\leq (1 + \frac{C_Q}{c}) \int_c^\infty \overline{G}(x)dx + C_Q \int_c^\infty g(x)dx. \end{aligned}$$

The right-hand side of the equation above does not depend on N or t, and converges to zero as $c \to \infty$ because g and $\overline{G}$ are integrable by Assumption II.a. This implies that

$$\lim_{c\to\infty} \sup_{N \in \mathbb{N},\, t \geq 0} \mathbb{E}\left[\overline{\nu}_t^{(N)}(\mathbb{1}_{[2c,\infty)})\right] \leq \lim_{c\to\infty} \sup_{N \in \mathbb{N},\, t \geq 0} \mathbb{E}\left[\overline{\nu}_t^{(N)}(\mathbb{1}_{f_c})\right] = 0. \tag{4.2.23}$$

Therefore, by (4.2.22)-(4.2.23), the family $\{\overline{\nu}_t^{(N)}; N \in \mathbb{N}, t \geq 0\}$ satisfies the criteria in Proposition 4.24, and hence, is tight in $\mathbb{M}_F[0,\infty)$. Finally, by Proposition 2.9, $\overline{\nu}_t^{(N)}$ converges in distribution to $\overline{\nu}_\infty^{(N)}$ in $\mathbb{M}_F[0,\infty)$ as $t \to \infty$, and so tightness of the sequence $\{\overline{\nu}_\infty^{(N)}\}_{N \in \mathbb{N}}$ follows from Remark 4.3. $\square$

PROPOSITION 4.21. Suppose Assumptions I-III hold. Then the condition (2.3.4) is satisfied with $(\overline{X}_0^{(N)}, \overline{\nu}_0^{(N)})$ replaced by $(\overline{X}_\infty^{(N)}, \overline{\nu}_\infty^{(N)})$, $N \in \mathbb{N}$.

PROOF. First, we claim that when $V_0^{(N)}$ is distributed as $\pi_*^{(N)}$, for every $f \in \mathbb{C}_b^1[0,\infty)$, there exists a constant $C_f < \infty$ such that

$$\sup_{t\geq 0, N\in\mathbb{N}} \frac{1}{\sqrt{N}} \mathbb{E}\left[|\nu_t^{(N)}(f) - N\overline{\nu}(f)| \right] \leq C_f. \tag{4.2.24}$$

Fix $f \in \mathbb{C}_b^1[0,\infty)$, $t \geq 0$ and $N \in \mathbb{N}$. By the definition (2.3.3) of $\overline{\nu}$,

$$\begin{aligned}
\overline{\nu}(f) &= \int_0^\infty f(s)\overline{G}(s)ds \\
&= \int_0^\infty (f\overline{G})(s+t)ds + \int_0^t (f\overline{G})(s)ds \\
&= \overline{\nu}(\Phi_t f) + \int_0^t (f\overline{G})(t-s)ds.
\end{aligned}$$

Therefore, substituting $K^{(N)}$ from (2.3.11) into the representation (2.3.10) of $\overline{\nu}$, using integration by parts, the fact that $X_0^{(N)} = N$, and recalling the diffusion scaling form of (2.2.6), we have

$$\begin{aligned}
\frac{1}{\sqrt{N}}\left(\nu_t^{(N)}(f) - N\overline{\nu}(f)\right) = \frac{1}{\sqrt{N}}\left(\nu_0^{(N)}(\Phi_t f) - N\overline{\nu}(\Phi_t f)\right) - \widehat{\mathcal{H}}_t^{(N)}(f) - f(0)\widehat{X}_t^{(N)+} \\
+ \int_0^t (f\overline{G})(t-s)d\widehat{E}_s^{(N)} + \int_0^t (f\overline{G})'(t-s)\widehat{X}_s^{(N)+}ds.
\end{aligned} \tag{4.2.25}$$

Since $\nu_0^{(N)}$ has the same distribution as ν_* defined in (4.1.1), $\nu_0^{(N)}(\Phi_t f) - N\overline{\nu}(\Phi_t f) \stackrel{(d)}{=} \sum_{j=1}^N \zeta_j^{N,t}$, where

$$\zeta_j^{N,t} \doteq \delta_{a_j}(\Phi_t f) - \overline{\nu}(\Phi_t f) = f(a_j+t)\frac{\overline{G}(a_j+t)}{\overline{G}(a_j)} - \int_0^\infty f(x+t)\overline{G}(x+t)dx,$$

and $\{a_j; j\geq 1\}$ are i.i.d. with p.d.f $\overline{G}$. Hence, $\zeta_j^{N,t}; j = 1, ..., N$, are also i.i.d. with $\mathbb{E}[\zeta_1^{N,t}] = 0$ and

$$\mathrm{Var}(\zeta_1^{N,t}) \leq \mathbb{E}\left[\delta_{a_j}(\Phi_t f)^2\right] = \int_0^\infty \frac{f^2(x+t)\overline{G}^2(x+t)}{\overline{G}^2(x)}\overline{G}(x)dx \leq \|f\|_\infty^2.$$

Therefore, $\sum_{j=1}^{N} \zeta_j^{N,t}$ has zero mean and finite variance equal to $N\mathrm{Var}(\zeta_1^{N,t})$, and hence,

(4.2.26)

$$\frac{1}{\sqrt{N}}\mathbb{E}\left[\left|\nu_0^{(N)}(\Phi_t f) - N\overline{\nu}(\Phi_t f)\right|\right] \leq \frac{1}{\sqrt{N}}\left(\mathbb{E}\left[(\sum_{j=1}^{N}\zeta_j^{N,t})^2\right]\right)^{\frac{1}{2}} = \mathrm{Var}(\zeta_1^{N,t})^{\frac{1}{2}} \leq \|f\|_\infty.$$

Next, by definition (2.3.9) of $\mathcal{H}_t^{(N)}$, the covariance function (2.3.6) of $\mathcal{M}^{(N)}$, and the bound (4.1.18) of Lemma 4.11, we have

$$\mathbb{E}\left[\widehat{\mathcal{H}}_t^{(N)}(f)^2\right] = \mathbb{E}\left[\int_0^t \overline{\nu}_{t-s}^{(N)}\left(\frac{f^2(\cdot+s)\overline{G}^2(\cdot+s)}{\overline{G}^2(\cdot)}\frac{g(\cdot)}{\overline{G}(\cdot)}\right)\, ds\right] \leq \|f\|_\infty^2 R_M(0),$$

and $R_M(0) < \infty$ by Lemma 4.11. Therefore,

$$\mathbb{E}\left[\left|\widehat{\mathcal{H}}_t^{(N)}(f)\right|\right] \leq \left(\mathbb{E}\left[\widehat{\mathcal{H}}_t^{(N)}(f)^2\right]\right)^{\frac{1}{2}} \leq \|f\|_\infty\sqrt{R_M(0)}. \tag{4.2.27}$$

Moreover, note that under Assumption III, for $f \in \mathbb{C}_b^1[0,\infty)$, $f\overline{G}$ is absolutely continuous with

$$\|(f\overline{G})'\|_\infty \leq H\|f\|_\infty + \|f'\|_\infty \leq (1+H)(\|f\|_\infty + \|f'\|_\infty). \tag{4.2.28}$$

Therefore, replacing f with $f\overline{G}$ in (4.1.34) of Lemma 4.15, we have

$$\begin{aligned}\mathbb{E}\left[\left|\int_0^t (f\overline{G})(t-s)d\widehat{E}_s^{(N)}\right|^2\right] &\leq C_E\|f\|_\infty^2 t\overline{G}(t) \\ &\quad + C_E(1+H)^2(\|f\|_\infty + \|f'\|_\infty)^2\left(\int_0^t s\overline{G}(s)ds\right)\left(\int_0^t \overline{G}(s)ds\right) \\ &\quad + C_E\|f\|_\infty^2\left(\int_0^t \overline{G}(s)ds\right)^2.\end{aligned}$$

Since $\overline{G}(t) = (O)(t^{-(2+\epsilon)})$ by Assumption II.b, $t\overline{G}(t)$, $\int_0^t \overline{G}(s)ds$ and $\int_0^t s\overline{G}(s)ds$ are all uniformly bounded in t. Therefore, there exists a constant $C < \infty$ such that

(4.2.29)

$$\mathbb{E}\left[\left|\int_0^t (f\overline{G})(t-s)d\widehat{E}_s^{(N)}\right|\right] \leq \left(\mathbb{E}\left[\left|\int_0^t (f\overline{G})(t-s)d\widehat{E}_s^{(N)}\right|^2\right]\right)^{\frac{1}{2}} \leq C(\|f\|_\infty + \|f'\|_\infty).$$

Finally, using the bounds (4.1.10) and (4.2.28), we have

$$\mathbb{E}\left[\left|-f(0)\widehat{X}_t^{(N)+}+\int_0^t (f\overline{G})'(t-s)\widehat{X}_s^{(N)+}ds\right|\right] \tag{4.2.30}$$

$$\begin{aligned}&\le |f(0)|\mathbb{E}\left[\widehat{X}_t^{(N)+}\right]+\int_0^t |(f\overline{G})'|(t-s)\mathbb{E}\left[\widehat{X}_s^{(N)+}\right]ds\\ &\le C_Q|f(0)|+C_Q(1+H)(\|f\|_\infty+\|f'\|_\infty)\int_0^t \overline{G}(x)dx\end{aligned}$$

$$\le 2C_Q(1+H)(\|f\|_\infty+\|f'\|_\infty). \tag{4.2.31}$$

The claim (4.2.24) follows from the equation (4.2.25) and the bounds (4.2.26)-(4.2.30). In addition, since by Proposition 2.9 for every $N\to\infty$ $\nu_t^{(N)}\Rightarrow\nu_\infty^{(N)}$ in $\mathbb{M}_F[0,\infty)$ as $t\to\infty$, and for every $f\in\mathbb{C}_b^1[0,\infty)$, $\nu_t^{(N)}(f)\Rightarrow\nu_\infty^{(N)}(f)$, the portmanteau theorem for weak convergence shows that

$$\sup_{N\in\mathbb{N}}\frac{1}{\sqrt{N}}\mathbb{E}\left[|\nu_\infty^{(N)}(f)-N\overline{\nu}(f)|\right]\le \sup_{N\in\mathbb{N}}\liminf_{t\to\infty}\frac{1}{\sqrt{N}}\mathbb{E}\left[|\nu_t^{(N)}(f)-N\overline{\nu}(f)|\right]\le C_f. \tag{4.2.32}$$

Now we return to the proof of the proposition. For every $N\in\mathbb{N}$,

$$\mathbb{E}\left[\left|\overline{X}_\infty^{(N)}-1\right|\right]=\frac{1}{\sqrt{N}}\mathbb{E}\left[\left|\widehat{X}_\infty^{(N)}\right|\right]\le\frac{1}{\sqrt{N}}\left(\mathbb{E}\left[\widehat{X}_\infty^{(N)+}\right]+\mathbb{E}\left[\widehat{X}_\infty^{(N)-}\right]\right). \tag{4.2.33}$$

By Proposition 2.9, $V_t^{(N)}\Rightarrow V_\infty^{(N)}$ in $\mathcal{V}^{(N)}$ as $t\to\infty$, and in particular, $X_t^{(N)}\Rightarrow X_\infty^{(N)}$. Hence, by Corollary 4.7 and the portmanteau theorem for weak convergence,

$$\mathbb{E}\left[\widehat{X}_\infty^{(N)+}\right]\le\liminf_{t\to\infty}\mathbb{E}\left[\widehat{X}_t^{(N)+}\right]\le C_Q. \tag{4.2.34}$$

Moreover, recall that by the non-idling condition (2.1.8), $\nu_t^{(N)}(\mathbf{1})=N\wedge X_t^{(N)}$, and by another use of Proposition 2.9, as $t\to\infty$,

$$\nu_t^{(N)}(\mathbf{1})-N\wedge X_t^{(N)}\Rightarrow\nu_\infty^{(N)}(\mathbf{1})-N\wedge X_\infty^{(N)}.$$

Therefore, since $\overline{\nu}(\mathbf{1})=1$,

$$\widehat{X}_\infty^{(N)-}=-\frac{(X_\infty^{(N)}-N)}{\sqrt{N}}\wedge 0=-\frac{X_\infty^{(N)}\wedge N-N}{\sqrt{N}}=-\frac{\nu_\infty^{(N)}(\mathbf{1})-N\overline{\nu}(\mathbf{1})}{\sqrt{N}}.$$

Hence, substituting $f=\mathbf{1}$ in (4.2.32),

$$\mathbb{E}\left[\widehat{X}_\infty^{(N)-}\right]=\frac{1}{\sqrt{N}}\mathbb{E}\left[|\nu_\infty^{(N)}(f)-N\overline{\nu}(f)|\right]\le C_{\mathbf{1}}. \tag{4.2.35}$$

We conclude from (4.2.33) and the bounds (4.2.34) and (4.2.35) that

$$\lim_{N\to\infty} \mathbb{E}\left[\left|\overline{X}_\infty^{(N)} - 1\right|\right] = \lim_{N\to\infty} \frac{1}{\sqrt{N}} \mathbb{E}\left[\left|\widehat{X}_\infty^{(N)}\right|\right] \le \lim_{N\to\infty} \frac{1}{\sqrt{N}}(C_Q + C_{\mathbf{1}}) = 0.$$

In other words, $N \to \infty$, $\overline{X}_\infty^{(N)}$ converges to $\overline{X} \equiv 1$ in $\mathbb{L}^1$, which implies both $\overline{X}^{(N)} \Rightarrow \overline{X}$ and $\mathbb{E}[\overline{X}^{(N)}] \to \overline{X}$.

Finally, by another use of (4.2.32), for every $f \in \mathbb{C}_b^1[0,\infty)$,

$$\lim_{N\to\infty} \mathbb{E}\left[\overline{\nu}_\infty^{(N)}(f) - \overline{\nu}(f)\right] \le \lim_{N\to\infty} \frac{1}{\sqrt{N}} C_f = 0,$$

which implies $\overline{\nu}_\infty^{(N)}(f) \Rightarrow \overline{\nu}(f)$. The result then follows from the tightness of the sequence $\{\overline{\nu}_\infty^{(N)}\}_{N\in\mathbb{N}}$ from Lemma 4.20 and the characterization of tightness in Lemma 4.25. □

4.2.3. Convergence of Stationary Distributions. Finally, we conclude the proof of Theorem 4.1 by uniquely characterizing subsequential limits of the sequence $\{\widehat{Y}_\infty^{(N)}\}_{N\in\mathbb{N}}$. We start with the following Lemma.

LEMMA 4.22. Suppose Assumptions I-IV hold. If along some subsequence $\{N_j\}$, as $j \to \infty$, $\widehat{\pi}^{(N_j)} \Rightarrow \pi_0$ in $\mathcal{Y}$, then π_0 is an invariant distribution of limit process Y defined in Section 3.1.1.

PROOF. First, let $Y = (X, Z)$ be the limit process with initial condition Y_0 distributed as π_0. Since $\widehat{Y}_0^{(N)}$ has distribution $\widehat{\pi}^{(N)}$ for every $N \in \mathbb{N}$, Y_0 has distribution π_0, and the projection map on a Hilbert space is continuous, by the continuous mapping theorem we have, as $N \to \infty$,

$$\left(\widehat{X}_0^{(N)}, \langle \widehat{Z}_0^{(N)}, e_1\rangle_{\mathbb{H}^1}, ..., \langle \widehat{Z}_0^{(N)}, e_k\rangle_{\mathbb{H}^1}\right) \Rightarrow (X_0, \langle Z_0, e_1\rangle_{\mathbb{H}^1}, ..., \langle Z_0, e_k\rangle_{\mathbb{H}^1}).$$

Next, let $V_0^{(N)}$ have the same distribution as the stationary distribution $V_\infty^{(N)}$. Then, for every $t \ge 0$, $V_t^{(N)} \overset{(d)}{=} V_\infty^{(N)}$, and by continuity (and thus, measurability) of the mapping $\widehat{\Psi}$ and the projection mappings on Hilbert spaces, we have

$$\left(\widehat{X}_0^{(N)}, \langle \widehat{Z}_0^{(N)}, e_1\rangle_{\mathbb{H}^1}, ..., \langle \widehat{Z}_0^{(N)}, e_k\rangle_{\mathbb{H}^1}\right) \overset{(d)}{=} \left(\widehat{X}_t^{(N)}, \langle \widehat{Z}_t^{(N)}, e_1\rangle_{\mathbb{H}^1}, ..., \langle \widehat{Z}_t^{(N)}, e_k\rangle_{\mathbb{H}^1}\right).$$

Moreover, since $(X_0^{(N)}, \nu_0^{(N)})$ has the same distribution as $(X_\infty^{(N)}, \nu_\infty^{(N)})$, by Proposition 4.21, the sequence $\{(\overline{X}_0^{(N)}, \overline{\nu}_0^{(N)})\}_N$ satisfies the condition (2.3.4), and $\{\widehat{Y}_0^{(N)}\}_N$ satisfies (4.2.3) along $\{N_j\}$ by assumption. Hence, by Proposition 4.19, for any $t > 0$ we have as $j \to \infty$,

$$\left(\widehat{X}^{(N_j)}, \langle \widehat{Z}_t^{(N_j)}, e_1\rangle_{\mathbb{H}^1}, ..., \langle \widehat{Z}_t^{(N_j)}, e_k\rangle_{\mathbb{H}^1}\right) \Rightarrow \left(X_t, \langle Z_t, e_1\rangle_{\mathbb{H}^1}, ..., \langle Z_t, e_k\rangle_{\mathbb{H}^1}\right),$$

The last three displays imply that for every $t \geq 0$,

$$\left(X_0, \langle Z_0, e_1\rangle_{\mathbb{H}^1}, ..., \langle Z_0, e_k\rangle_{\mathbb{H}^1}\right) \stackrel{(d)}{=} \left(X_t, \langle Z_t, e_1\rangle_{\mathbb{H}^1}, ..., \langle Z_t, e_k\rangle_{\mathbb{H}^1}\right),$$

and hence, Y_0 and Y_t have the same finite-dimensional distributions on the chain of increasing finite-dimensional subspaces $\mathcal{L}_k = \text{span}\{e_1, ...e_k\}$. Therefore, $Y_0 \stackrel{(d)}{=} Y_t$ (e.g., see the discussion in [**93**, Chapter I.1]), and $\pi_0 = \mathscr{L}aw(Y_0)$ is an invariant distribution of the limit process. □

As a corollary of the previous lemma and Theorem 4.2, we obtain the existence of an invariant distribution for the limit process Y.

COROLLARY 4.23. Under Assumptions I-IV, the limit process Y has a unique invariant distribution π.

PROOF. By Theorem 4.2, the sequence $\{\widehat{\pi}^{(N)}\}$ is tight in $\mathcal{Y}$. Therefore, there exists a subsequence $\{N_j\}$ such that $\{\widehat{\pi}^{(N_j)}\}$ converges in distribution to a probability distribution π on $\mathcal{Y}$, as $j \to \infty$. By Lemma 4.22, this subsequential limit is an invariant distribution of the limit process, which proves the existence. The uniqueness of the invariant distribution is proved in Theorem 3.26. □

Finally, we conclude the proof of Theorem 4.1.

PROOF OF THEOREM 4.1. In light of the tightness of $\{\widehat{\pi}^{(N)}\}$ established in Theorem 4.2, a standard argument by contradiction then shows that $\{\widehat{\pi}^{(N)}\}$ converges in distribution to the unique invariant distribution π of the limit process Y. Suppose $\{\widehat{\pi}^{(N)}\}$ has a subsequence $\{N_j\}$ such that $\lim_{j\to\infty} d_p(\widehat{\pi}^{(N_j)}, \pi) > 0$, where d_p represents the Prohorov metric on the space of probability measures on $\mathcal{Y}$. Then, since $\{\widehat{\pi}^{(N_j)}\}$ is tight by Theorem 4.2, we can assume without loss of generality that $\widehat{\pi}^{(N_j)} \Rightarrow \pi_0$ in $\mathcal{Y}$, for some probability distribution π_0 on $\mathcal{Y}$. By the claim above, π_0 is an invariant distribution of the limit process Y. But, since the limit process has a unique invariant distribution π by Corollary 4.23, it follows that $\pi_0 = \pi$, and hence, that $\lim_{j\to\infty} d_p(\widehat{\pi}^{(N_j)}, \pi) > 0$, which is a contradiction. $\square$

4.A. Various Tightness Criteria

4.A.1. Tightness and Convergence of Random Measures. A criteria for tightness of a family of $\mathbb{M}_F[0,\infty)$-valued random variables is given in Proposition 4.1 of **[61]**.

PROPOSITION 4.24. A family $\{\mu_\alpha; \alpha \in \mathcal{A}\}$ of $\mathbb{M}_F[0,\infty)$-valued random variables is tight if and only if the following two conditions hold:

a. $\sup_{\alpha\in\mathcal{A}} \mathbb{E}\left[|\langle \mathbf{1}, \mu_\alpha\rangle|\right] < \infty$;
b. $\lim_{c\to\infty} \sup_{\alpha\in\mathcal{A}} \mathbb{E}\left[|\mu_\alpha[c,\infty)|\right] = 0$.

The following Lemma gives a characterization for the limit of a sequence of random measures in $\mathbb{M}_F[0,\infty)$.

LEMMA 4.25. Suppose the sequence $\{\mu_n\}_{n\in\mathbb{N}}$ of random measures is tight in $\mathbb{M}_F[0,\infty)$. Assume there exists a deterministic measure $\mu \in \mathbb{M}_F[0,\infty)$ such that for every $f \in \mathbb{C}_b^1[0,\infty)$, $\mu_n(f) \Rightarrow \mu(f)$ in $\mathbb{R}$ as $n \to \infty$. Then, $\mu_n \Rightarrow \mu$ in $\mathbb{M}_F[0,\infty)$.

PROOF. Since $\{\mu_n\}_{n\in\mathbb{N}}$ is tight, every subsequence of $\{\mu_n\}$ has a further converging subsequence $\{\mu_{n_k}\}$. Let $\tilde{\mu}$ be any such sub-sequential limit. For every $f \in \mathbb{C}_b^1[0,\infty)$, the mapping $\mu \mapsto \mu(f)$ from $\mathbb{M}_F[0,\infty)$ to $\mathbb{R}$ is continuous, and hence, by the continuous mapping theorem, $\mu_{n_k}(f) \Rightarrow \tilde{\mu}(f)$. On the other hand, $\mu_{n_k}(f) \Rightarrow \mu(f)$ by the assumption of the lemma. Hence, $\tilde{\mu}(f) \stackrel{(d)}{=} \mu(f)$, and since μ is deterministic, $\tilde{\mu}(f) = \mu(f)$, almost surely. As a consequence, for a countable dense subspace $\mathcal{C}$ of $\mathbb{C}_b^1[0,\infty)$, almost surely, the equality

$$\tilde{\mu}(f) = \mu(f), \qquad \forall f \in \mathcal{C},$$

holds simultaneously. This implies $\tilde{\mu} = \mu$, almost surely, which uniquely characterizes the sub-sequential limits of $\{\mu_n\}$, and completes the proof. $\square$

4.A.2. Tightness in $\mathbb{H}^1(0,\infty)$. To obtain a criterion for the tightness of a set of probability measures or random elements in $\mathbb{H}^1(0,\infty)$, we first obtain a characterization of compact sets in $\mathbb{H}^1(0,\infty)$. We start with a version of the Sobolev Embedding Theorem.

PROPOSITION 4.26. [**1**, Theorem 6.2 on page 144, Part II, equation (6)] Let Ω_0 be a bounded subinterval of $(0,\infty)$. Then the embedding

$$\mathbb{H}^1(\Omega_0) \mapsto \mathbb{L}^2(\Omega_0)$$

is compact.

The following result helps to extend the last result to $\mathbb{L}^2(\Omega)$ when $\Omega = (0,\infty)$ is unbounded.

PROPOSITION 4.27. [**1**, Theorem 2.22 on page 35] A subset $K \subset \mathbb{L}^2(\Omega)$ is pre-compact if there exist sub-domains Ω_j, $j \in \mathbb{N}$, of Ω such that

(a) $\Omega_j \subset \Omega_{j+1}$ for $j \in \mathbb{N}$;
(b) $K_j = \{f|_{\Omega_j}; f \in K\}$ is pre-compact in $\mathbb{L}^2(\Omega_j)$ for $j \in \mathbb{N}$;
(c) for every $\epsilon > 0$, there exists an index $j \in \mathbb{N}$ such that

$$\int_{\Omega \setminus \Omega_j} |f(x)|^2 dx < \epsilon, \quad \forall f \in K.$$

Propositions 4.26 and 4.27 provide criteria for the tightness of a sequence of random elements in $\mathbb{L}^2(0,\infty)$.

LEMMA 4.28. A subset $K \subset \mathbb{L}^2(0,\infty)$ is compact if

a. $\sup_{f\in K} \|f\|_{\mathbb{H}^1(0,L)} < \infty, \quad \forall L \in \mathbb{N}$,
b. $\lim_{L\to\infty} \sup_{f\in K} \|f\|_{\mathbb{L}^2(L,\infty)} = 0$.

Moreover, a family $\{\zeta_\alpha, \alpha \in \mathcal{A}\}$ of $\mathbb{H}^1(0,\infty)$-valued random elements defined on probability spaces $(\Omega_\alpha, \mathbb{P}_\alpha)$ is tight in $\mathbb{L}^2(0,\infty)$ if

a. $\lim_{\lambda \uparrow \infty} \sup_{\alpha \in \mathcal{A}} \mathbb{P}_\alpha \left\{ \|\zeta_\alpha\|_{\mathbb{H}^1(0,L)} > \lambda \right\} = 0, \quad \forall L \in \mathbb{N},$

b. $\lim_{L \to \infty} \sup_{\alpha \in \mathcal{A}} \mathbb{P}_\alpha \left\{ \|\zeta_\alpha\|_{\mathbb{L}^2(L,\infty)} > \epsilon \right\} = 0, \quad \forall \epsilon > 0.$

PROOF. The first claim is a direct sequence of Propositions 4.26 and 4.27, with $\Omega_N = (0, N)$. For every $N \in \mathbb{N}$, Ω_N is bounded, and Proposition 4.26 asserts that K is compact in $\mathbb{L}^2(\Omega_N)$ because it is uniformly bounded in the $\mathbb{H}^1(\Omega_N)$ norm. The second claim follows from the first claim by exactly the same argument as in [**59**, Theorem 4.10]. □

PROOF OF PROPOSITION 4.9. The result is a consequence of the second claim in Lemma 4.28, the fact that ζ_α is tight in $\mathbb{H}^1(0,\infty)$ if ζ_α and ζ_α' are tight in $\mathbb{L}^2(0,\infty)$, the definition of the $\mathbb{H}^1(0,\infty)$ norm and the elementary inequality $\mathbb{P}(A + B > 2\lambda) \leq \mathbb{P}(A > \lambda) + \mathbb{P}(B + \lambda)$. □

Part II

The Hydrodynamic Limit of a Randomized Load-Balancing Network

CHAPTER 5

N-Server Load-Balancing Network: Description and Dynamics

In this chapter, we study in detail an N-server network with general service distribution under the "supermarket" randomized algorithm. As mentioned in the introduction, we introduce a sequence of interacting measure-valued processes that describes the state of the network at each time. Moreover, we describe the dynamics of the network, by deriving equations that govern the evolution of these measure-valued state process.

Section 5.1 contains the definition of the measure-valued Markovian state process $(\nu_\ell^{(N)}(t); \ell \geq, t \geq 0)$, as well as the assumptions we impose on the arrival process and service time distribution. (An extra condition, Assumption IV, is required for some of our results, which is stated in Chapter 6.) In Section 5.2, we study the dynamics of the measure-valued state process. The results of Section 5.2 are fairly general, and hold for a large class of networks in which each queue is served using a head-of-the-line policy, and does not depend on the specific choice of the load-balancing algorithm. Finally, in Section 5.3, we establish martingale decompositions for certain stochastic processes that describe arrivals and departures to and from the network. These decompositions are essential for establishing hydrodynamic limits later in Chapter 7.

5.1. Model Description and State Representation

In this section, we provide more details on the description and dynamics of N-server network. First in Section 5.1.1, we describe the N-server load balancing network and $SQ(d)$ load-balancing algorithm, as well as the assumptions on the arrival process and service time distribution. Then in Section 5.1.2, we precisely define the sequence of interacting measure-valued processes $(\nu_\ell^{(N)}; \ell \geq 1)$ which

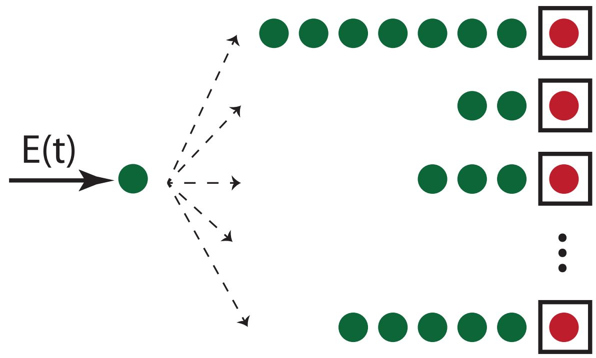


FIGURE 1. The load-balancing network.

serve as the state descriptor, as well as its state space $\mathbb{S}$. In Section 5.1.3, some auxiliary processes are defined which help us describe the dynamics of the state variables in a more succinct and intuitive fashion. Finally in Section 5.2, we introduce and prove a set of equations governing the dynamics of the state variables.

REMARK 5.1. To make it easy to follow the notation, throughout the paper, we use the superscript i to denote queue indices and subscript j to denote jobs.

5.1.1. Basic Definitions and Assumptions. Consider a network of N servers, each equipped with an infinite capacity queue, serving a common stream of arriving jobs (see Figure 1). For each $i = 1, ..., N$, the length $X^{(N),i}(t)$ of a queue i at a time $t \geq 0$ is the number of jobs in that queue including the job in service, and $X^{(N)}(t)$ denotes the total number of jobs in the network. Each server is assumed to be non-idling, that is, as long as the length of a queue is non-zero, there is a job receiving service at that queue.

Let $E^{(N)}$ denote the cumulative arrival process, that is, for every $t \geq 0$, $E^{(N)}(t)$ describes the number of jobs that arrive to the N-server network during the interval $[0, t]$. We impose the following assumption on the arrival process $E^{(N)}$, although our results holds for more general arrival processes. Define $\widetilde{E}$ to be a delayed renewal process with delay $\widetilde{u}_0$ and inter-arrival times $\widetilde{u}_n, n \geq 1$. Inter-arrival times have a common distribution $G_{\widetilde{E}}$ which has mean λ^{-1} density $g_{\widetilde{E}}$, where $\lambda > 0$

is a constant. Also, the delay $\widetilde{u}_0$ has distribution

$$\mathbb{P}\left\{\widetilde{u}_0 > r\right\} = \frac{1 - G_{\widetilde{E}}(\tilde{R} + r)}{1 - G_{\widetilde{E}}(\tilde{R})},$$

for some $\tilde{R} \geq 0$.

ASSUMPTION I. The arrival process satisfies $E^{(N)}(t) = \widetilde{E}(Nt), t \geq 0$, and there, $E^{(N)}$ is a delayed renewal process that has delay $u_0^{(N)} = \widetilde{u}_0/N$, and inter-arrival times $u_n^{(N)} \doteq \widetilde{u}_n/N, n \geq 1$, with common distribution $G_E^{(N)}(x) \doteq G_{\widetilde{E}}(Nx)$, $x \geq 0$, and probability density function $g_E^{(N)}$.

Note that the $u_0^{(N)}$ has distribution

$$\mathbb{P}\left\{u_0^{(N)} > r\right\} = \frac{\overline{G}_E^{(N)}(R^{(N)} + r)}{\overline{G}_E^{(N)}(R^{(N)})}, \tag{5.1.1}$$

where $R^{(N)} \doteq \widetilde{R}/N$. Let $R^{(N)}$ be the backward recurrence time corresponding to the arrival process, defined as

$$R^{(N)}(t) \doteq \begin{cases} R^{(N)} + t & \text{if } 0 \leq t < u_0 \\ t - \sup\{s \geq 0, E^{(N)}(s) < E^{(N)}(t).\} & \text{if } t \geq u_0, \end{cases} \tag{5.1.2}$$

where $R^{(N)}$ is as in Assumption I. In particular, $R^{(N)}(0) = R^{(N)}$.

REMARK 5.2. The results of this paper hold can easily be extended to a larger class of arrival processes. In particular, we can show that the our results hold for time-inhomogeneous Poisson processes (although we have this result, the proof is not included in this dissertation).

The sequence of service times, denoted by $\{v_j; j \in \mathbb{Z}\}$, are assumed to be i.i.d, and independent of the arrival process. We let G and $\overline{G}$ be, respectively, the cumulative distribution function and the complementary cumulative distribution function of the service times. When G has density g, the hazard rate h associated with the service distribution is defined as

$$h(x) \doteq \frac{g(x)}{1 - G(x)}, \qquad x \in [0, L), \tag{5.1.3}$$

where L is the right-end of the support of the service time distribution:

$$L \doteq \sup\{x \in [0,\infty) : G(x) < 1\}. \tag{5.1.4}$$

The following very mild conditions are imposed on G.

ASSUMPTION II. The service time distribution G has the following properties.

(a) G has density g and finite mean which is set to 1.

(b) There exists $\ell_0 < L$ such that h is either bounded or lower semi-continuous on (ℓ_0, L).

For each job j, let $\alpha_j^{(N)}$ represent the time at which job j arrives into the system, let $\beta_j^{(N)}$ be the time at which it enters service, and let $\gamma_j^{(N)}$ be the time at which it departs the queue after completing service. Note that $\alpha_j^{(N)} = \beta_j^{(N)}$ if the job is routed to an empty queue, and also that $\gamma_j^{(N)} = \beta_j^{(N)} + v_j$, where $\{v_j\}$ is the i.i.d sequence of service times. We use the convention that jobs initially in network are indexed by non-positive numbers $j = j_0^{(N)}, ..., 0$ with $j_0^{(N)} \doteq -X^{(N)}(0) + 1$ being the smallest job index, and job which arrive after time 0 are indexed by $j \geq 1$, and according to their arrival time ($0 < \alpha_j^{(N)} < \alpha_{j+1}^{(N)}$ for $j \geq 1$, almost surely). Also, we define the age $a_j^{(N)}(t)$ of job j at time t as follows:

$$a_j^{(N)}(t) \doteq \begin{cases} 0 & \text{if } t < \beta_j^{(N)}, \\ t - \beta_j^{(N)} & \text{if } \beta_j^{(N)} \leq t < \gamma_j^{(N)}, \\ v_j & \text{if } t \geq \gamma_j^{(N)}. \end{cases} \tag{5.1.5}$$

Thus, the age of a job is zero before service entry time, then increases linearly until the departure time, and remain constant afterwards.

Throughout this paper, we assume that the following randomized algorithm is being used, unless stated otherwise.

***SQ(d)* Load-Balancing Algorithm:** Upon arrival, each job chooses d queue indices independently and uniformly at random, and independent of arrival process and service times. The job is then routed to the queue with the shortest queue length, and ties are broken uniformly at random.

Let $\iota_j = (\iota_j(1), ..., \iota_j(d))$ be the vector of d queue indices chosen by job $j \geq 1$, and $\kappa_j^{(N)}$ be the queue with the shortest queue length to which job j is routed. Hence,

$$\kappa_j^{(N)} \doteq \text{argmin} \left(X^{(N),\iota_j(1)}(\alpha_j^{(N)}-), ..., X^{(N),\iota_j(d)}(\alpha_j^{(N)}-) \right), \tag{5.1.6}$$

that is, $\kappa_j^{(N)} = \iota_j(i)$ for some $i = 1, ..., d$ if $X^{(N),\iota_j(d')}(\alpha_j^{(N)}-) > X^{(N),\iota_j(i)}(\alpha_j^{(N)}-)$ for all $1 \leq d' < i$ and $X^{(N),\iota_j(i)}(\alpha_j^{(N)}-) \leq X^{(N),\iota_j(d')}(\alpha_j^{(N)}-)$ for all $i < d' \leq d$. According to (5.1.6), given that subset of queue indices are chosen and have equal queue lengths which is the shortest among all d chosen queues, the job is equally likely to be set to each of those queues. With a slight abuse of notation, we use $\{\kappa_j^{(N)}(t); t \geq 0\}$ to also denote the queue index process, defined as

$$\kappa_j^{(N)}(t) \doteq \mathbb{1}(t \geq \alpha_j^{(N)})\kappa_j^{(N)}.$$

All these random variables are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

5.1.2. A Markovian Measure-Valued Representation. In our representation, the state of the N-server network at each time t is described by a sequence $(\nu_\ell^{(N)}(t);$ $\ell \geq 1)$ if measure-valued random variables.

State Space. Recall that $\mathbb{M}_{\leq 1}[0, L)$ is the space of sub-probability measures on $[0, L)$ equipped with the topology of weak convergence, which is compatible with the Prohorov's metric d_P. Consider the space $\mathbb{S}$ of infinite vectors $\mu = (\mu_\ell; \ell \geq 1)$ of ordered sub-probability measures, in the sense that for every non-negative, bounded continuous function f, $\langle f, \mu_\ell \rangle \geq \langle f, \mu_{\ell+1} \rangle$. Therefore,

$$\mathbb{S} = \{(\mu_\ell; \ell \geq 1) \in \mathbb{M}_{\leq 1}[0, L)^{\mathbb{N}} \; ; \; \langle f, \mu_\ell \rangle \geq \langle f, \mu_{\ell+1} \rangle, \forall \ell \geq 1, f \in \mathbb{C}_b[0, \infty), f \geq 0\}. \tag{5.1.7}$$

The space $\mathbb{S}$ is utilized with a metric $d_{\mathbb{S}}$ defined as

$$d_{\mathbb{S}}(\mu, \tilde{\mu}) = \sup_{\ell \geq 1} \frac{d_P(\mu_\ell, \tilde{\mu}_\ell)}{\ell}, \tag{5.1.8}$$

It is easy to see a sequence μ^n converges to μ in $\mathbb{M}_{\leq 1}[0, L)^{\mathbb{N}}$ if and only if for every $\ell \geq 1$, μ_ℓ^n converges weakly to μ_ℓ, and that $(\mathbb{M}_{\leq 1}[0, L)^{\mathbb{N}}, d_{\mathbb{S}})$ is a Polish space. It is also straightforward to show that $\mathbb{S}$ is a closed subset of $\mathbb{M}_{\leq 1}[0, L)^{\mathbb{N}}$ under the metric $d_{\mathbb{S}}$, and hence $(\mathbb{S}, d_{\mathbb{S}})$ is itself a Polish space. Recall that $\mathbb{C}_{\mathbb{S}}[0,\infty)$ and $\mathbb{D}_{\mathbb{S}}[0,\infty)$ are the spaces of $\mathbb{S}$-valued functions on $[0,\infty)$ which are continuous, and right-continuous with finite left limits at every point in $(0,\infty)$, respectively.

State Variables. Recall that $\nu_\ell^{(N)}(t)$ is a (random) finite measure on $[0,\infty)$ that has a unit delta mass at the age (i.e., amount of time spent in service) of each job that, at time t, is in service at a queue of length no less than ℓ. Define

$$\nu^{(N)}(t) \doteq (\nu_\ell^{(N)}(t); \ell \geq 1).$$

For each $\ell \geq 1$, $\nu_\ell^{(N)}(t)$ takes values on the space $\mathbb{M}_{\leq 1}[0, L)$, and $\nu^{(N)}(t)$ takes values in $\mathbb{S}$ for every $t \geq 0$. The $\mathbb{R} \times \mathbb{S}$-valued random element $(R^{(N)}, \nu^{(N)})$ with $\nu^{(N)} = (\nu_\ell^{(N)}; \ell \geq 1)$ serves as the state descriptor of the N-server network. Recalling the notation $\mathbf{1}$ for the constant function that is identically one everywhere on $[0,\infty)$, $S_\ell^{(N)}(t) \doteq \langle \mathbf{1}, \nu_\ell^{(N)}(t) \rangle$ is the number of queues with length at least ℓ at time t.

The state descriptor $(\nu_\ell^{(N)}(t)\ ; \ell \geq 1)$ can be expressed in terms of primitives of the network. Note that a job j receives service during the interval $[\beta_j^{(N)}, \gamma_j^{(N)})$, and hence,

$$\mathcal{V}^{(N)}(t) \doteq \left\{ j \geq j_0 \ : \ \beta_j^{(N)} \leq t < \gamma_j^{(N)} \right\}, \tag{5.1.9}$$

is the set of jobs receiving service at time t. Also, for a job j, let $\chi_j^{(N)}(t)$ defined on $[\alpha_j^{(N)}, \infty)$ denote the length at time t of the queue to which job j was routed, which can be represented as

$$\chi_j^{(N)}(t) = X^{(N),\kappa_j^{(N)}}(t), \quad t \geq \alpha_j^{(N)}. \tag{5.1.10}$$

The value of $\chi_j^{(N)}(t)$ for $t < \alpha_j^{(N)}$ is irrelevant. Therefore, for $\ell \geq 1$,

$$\mathcal{U}_\ell^{(N)}(t) \doteq \left\{ j \geq j_0 : \mathbb{1}\left(\alpha_j^{(N)} \leq t\right) \chi_j^{(N)}(t) \geq \ell \right\} \tag{5.1.11}$$

is the set of jobs in queues with length at least ℓ at time t. Using this notation, $\nu_\ell(t)$ can be represented as

$$\nu_\ell^{(N)}(t) \doteq \sum_{j=j_0}^{\infty} \mathbb{1}\left(\beta_j^{(N)} \leq t\right) \mathbb{1}\left(\gamma_j^{(N)} > t\right) \mathbb{1}\left(\chi_j^{(N)}(t) \geq \ell\right) \delta_{a_j^{(N)}(t)} \tag{5.1.12}$$

$$= \sum_{j=j_0}^{\infty} \mathbb{1}\left(j \in \mathcal{V}^{(N)}(t)\right) \mathbb{1}\left(j \in \mathcal{U}_\ell^{(N)}(t)\right) \delta_{a_j^{(N)}(t)}. \tag{5.1.13}$$

Initial Conditions. The initial state of the network can essentially be described only using $R^{(N)}(0)$ and $\nu^{(N)}(0)$. In particular, each delta mass in $\nu_1^{(N)}(0)$ corresponds to a non-empty queue, with the initial age of the job receiving service in the corresponding queue is the location of that delta mass, and the length of the corresponding queue is the number of $\nu_\ell^{(N)}(0)$, $\ell \geq 1$, in which that delta mass is present. Hence, the number of initially non-empty queues is the number of delta mass in $\nu_1^{(N)}(0)$, that is, $\langle \mathbf{1}, \nu_1^{(N)}(0)\rangle$, and the rest of N queues, if any, are empty. Also, the rest of $X^{(N)}(0) = \sum_{\ell=1}^{\infty} \langle \mathbf{1}, \nu_\ell^{(N)}(0)\rangle$ jobs initially in network are initially waiting in queues, and have initial age 0. However, we need to give each queue an index $i \in \{1, ..., N\}$ an each job initially in network an index $j \in \{-X^{(N)}(0) + 1, ..., N\}$, base on a convention, such that $\kappa_j^{(N)}$ is known for every $j = -X^{(N)}(0) + 1, ..., 0$. We choose the convention such that the the jobs initially in service have the smaller indices $j \in \{-X^{(N)}(0) + 1, ..., -X^{(N)}(0) + \langle \mathbf{1}, \nu_1^{(N)}(0)\rangle\}$. Such indexing convention is arbitrary as the dynamics of the network does not depend on the particular convention one chooses. Let

$$I_0 = \left(R^{(N)}(0), a_j^{(N)}(0), \kappa_j^{(N)}; j = -X^{(N)}(0) + 1, ..., 0\right), \tag{5.1.14}$$

and note that for a fixed indexing convention, I_0 and $(R^{(N)}(0), \nu^{(N)}(0))$ can be obtained from each other. We impose the following condition on initial conditions.

ASSUMPTION III. The sequence of initial conditions satisfies the followings.

(a) For every $N \in \mathbb{N}$, $X^{(N)}(0) = \sum_{\ell \geq 1} \langle \mathbf{1}, \nu_\ell^{(N)}(0)\rangle < \infty$ almost surely, the arrival process and random queue choices in the load balancing algorithm are independent of $\nu^{(N)}(0)$, and every service time v_j is conditionally independent of $\nu^{(N)}(0)$ given the initial age $a_j(0)$, that is, for every finite subset $\mathcal{K}$ of jobs,

$$\mathbb{P}\left\{v_j > b_1; j \in \mathcal{K} | I_0\right\} = \prod_{j\in\mathcal{K}} \frac{\overline{G}(a_j^{(N)}(0)+b_j)}{\overline{G}(a_j^{(N)}(0))}. \tag{5.1.15}$$

In particular, for every job j which is not initially in service, v_j is independent of I_0.

5.1.3. Auxiliary Processes and Filtration. To describe the dynamics of the processes $\nu_\ell^{(N)}$, it will be convenient to introduce a number of auxiliary processes. For every queue $i \in \{1, ..., N\}$, let $E^{(N),i}$ denote the cumulative arrival process to queue i, defined as

$$E^{(N),i}(t) \doteq \sum_{j=1}^{\infty} \mathbb{1}\left(\alpha_j^{(N)} \le t\right) \mathbb{1}\left(\kappa_j^{(N)} = i\right), \qquad t \ge 0. \tag{5.1.16}$$

For $\ell \ge 1$, let $\nu_\ell^{(N),i}$ denote the contribution of the i^{th} queue to the measure $\nu_\ell^{(N)}$, that is, for $t \ge 0$,

$$\nu_\ell^{(N),i}(t) \doteq \sum_{j=j_0}^{\infty} \mathbb{1}\left(\beta_j^{(N)} \le t\right) \mathbb{1}\left(\gamma_j^{(N)} > t\right) \mathbb{1}\left(\chi_j^{(N)}(t) \ge \ell\right) \mathbb{1}\left(\kappa_j^{(N)} = i\right) \delta_{a_j^{(N)}(t)}. \tag{5.1.17}$$

Clearly, we have

$$\nu_\ell^{(N)} = \sum_{i=1}^{N} \nu_\ell^{(N),i}. \tag{5.1.18}$$

Also, for $\ell \ge 1$ and $\varphi \in \mathbb{C}_b([0, L) \times \mathbb{R}_+)$, let $\mathcal{R}_{\varphi,\ell}^{(N)}$ be the cumulative φ-weighted routing measure process to the queues with length exactly $\ell - 1$, defined as follows for all $t \ge 0$:

$$\mathcal{R}_{\varphi,\ell}^{(N)}(t) \doteq \begin{cases} \sum_{i=1}^N \int_{(0,t]} \varphi(0,s)\mathbf{1}(X^{(N),i}(s-)=0)dE^{(N),i}(s) & \text{if } \ell = 1, \\ \sum_{i=1}^N \int_{(0,t]} \langle \varphi(\cdot,s), \nu_{\ell-1}^{(N),i}(s-) - \nu_\ell^{(N),i}(s-)\rangle dE^{(N),i}(s), & \text{if } \ell \ge 2. \end{cases} \tag{5.1.19}$$

Moreover, let $D_\ell^{(N)} = \{D_\ell^{(N)}(t); t \ge 0\}$ be the counting process of departures from queues with length at least ℓ right before departure, that is, for every $t \ge 0$

and for $\ell \in \mathbb{N}$, $D_\ell^{(N)}(t)$ denotes the total number of jobs that completed service in the interval $[0,t]$ at a queue that had length ℓ just prior to service completion. To express $D_\ell^{(N)}$ and other process in a more succinct way, we use the following notation for values of the queue length process $\chi_j^{(N)}$ of a job j (defined in (5.1.10)) right after the arrival time, right after service entry time and right before the departure time of that job:

$$\chi_j^{E,(N)} \doteq \chi_j^{(N)}(\alpha_j^{(N)}), \quad \chi_j^{K,(N)} \doteq \chi_j^{(N)}(\beta_j^{(N)}), \quad \chi_j^{D,(N)} \doteq \chi_j^{(N)}(\gamma_j^{(N)}-). \tag{5.1.20}$$

Using this notation, $D_\ell^{(N)}$ can be expressed as:

$$D_\ell^{(N)}(t) \doteq \sum_{j=j_0}^{\infty} \mathbb{1}\left(\gamma_j^{(N)} \le t\right) \mathbb{1}\left(\chi_j^{D,(N)} \ge \ell\right), \qquad t \ge 0. \tag{5.1.21}$$

REMARK 5.3. Note that for every job j, $\chi_j^{D,(N)} \ge 1$ and $\chi_j^{K,(N)} \ge 1$ because a queue is never empty just prior to a departure or right after a service entry. Also, The process $D^{(N)} \doteq D_1^{(N)}$ is the total cumulative departure process. As the total number of departures cannot exceed the total number of arrivals plus the number of jobs initially in network, we have the bound

$$D^{(N)}(t) \le X^{(N)}(0) + E^{(N)}(t). \tag{5.1.22}$$

For every $\ell \ge 1$ and $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$, let $\mathcal{D}_{\varphi,\ell}^{(N)}$ be the cumulative φ-weighted departure process from queues of length at least ℓ, defined by

$$\mathcal{D}_{\varphi,\ell}^{(N)}(t) \doteq \sum_{j=j_0}^{\infty} \varphi\left(v_j, \gamma_j^{(N)}\right) \mathbb{1}\left(\gamma_j^{(N)} \le t\right) \mathbb{1}\left(\chi_j^{D,(N)} \ge \ell\right), \qquad t \ge 0. \tag{5.1.23}$$

Note that for $\varphi = \mathbf{1}$, $\mathcal{D}_{\mathbf{1},\ell}^{(N)} = D_\ell^{(N)}$, and hence using (5.1.22), for all $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$, $\ell \ge 1$ and $t \ge 0$, the bound

$$|\mathcal{D}_{\varphi,\ell}^{(N)}(t)| \le \|\varphi\|_\infty \left(X^{(N)}(0) + E^{(N)}(t)\right), \tag{5.1.24}$$

holds. Moreover, for every queue $i \in \{1, ..., N\}$, let $D^{(N),i}$ denote the cumulative departure process from queue i, defined as

$$D^{(N),i}(t) \doteq \sum_{j=j_0}^{\infty} \mathbb{1}\left(\gamma_j^{(N)} \le t\right) \mathbb{1}\left(\kappa_j^{(N)} = i\right), \qquad t \ge 0. \tag{5.1.25}$$

The queue length $X^{(N),i}$ satisfies the following mass balance equation

$$X^{(N),i}(t) = X^{(N),i}(0) + E^{(N),i}(t) - D^{(N),i}(t). \tag{5.1.26}$$

Finally, we define the filtration $\{\mathcal{F}_t^{(N)}; t \geq 0\}$, which is generated by the initial conditions of the network, plus arrival to and departure processes to all queues. Define

$$\mathcal{F}_t^{(N)} \doteq \bigvee_{i=1}^{N} \left(\mathcal{F}_t^{E^{(N),i}} \vee \mathcal{F}_t^{D^{(N),i}} \right) \vee \sigma(I_0), \tag{5.1.27}$$

where I_0 is defined in (5.1.14) and $\mathcal{F}_t^{E^{(N),i}}$ and $\mathcal{F}_t^{D^{(N),i}}$ are filtrations generated by $E^{(N),i}$ and $D^{(N),i}$, respectively. In Proposition 5.21, we show that all state variables and auxiliary processes are $\mathcal{F}_t^{(N)}$-adapted.

REMARK 5.4. We could alternatively define $\{\mathcal{F}_t^{(N)}\}$ to be the filtration generated by the age and queue index processes $a_j^{(N)}(\cdot), \kappa_j^{(N)}(\cdot); j \geq 1$, which resembles the filtration defined in **[61]**. While one can easily show that the two definitions are equivalent, our definition allows us to directly apply the compensator result from **[14]** in Section 5.3, and hence is preferable in this paper.

REMARK 5.5. For every $N \in \mathbb{N}$, one can show that $\{(R^{(N)}(t), \nu^{(N)}(t)); t \geq 0\}$ is a Markov process with respect to the filtration $\{\mathcal{F}_t^{(N)}; t \geq 0\}$. We do not use this property and hence, do not include the proof in this paper.

5.2. Equations Governing the Dynamics of the N-Server Network

Using the notation defined in the last section, the dynamics of the state variable $\nu^{(N)}$ for a fixed N is studied in the next proposition. This result does not depend on our assumptions on arrival process and service time distribution, and hold for a very general class of networks and load balancing algorithms, as long as all arrival and departure times are distinct, almost surely.

To make this notion precise, denote by $\Omega_{s,\delta}$ the set of realizations for which at most one arrival or one departure occurs during $(s, s+\delta]$. Also, define Ω_t to be the set of realizations for which there exists a partition $\{(\frac{k}{n}, \frac{k+1}{n}]; k = 0, ..., \lfloor nt \rfloor\}$ of

$(0,t]$ such that at most one arrival or one departure occurs in each subinterval, that is,

$$\Omega_t \doteq \bigcup_{n=1}^{\infty} \bigcap_{k=0}^{\lfloor nt \rfloor} \Omega_{\frac{k}{n},\frac{1}{n}}. \tag{5.2.1}$$

The result in Proposition 5.6 below holds as long as for every $t \geq 0$, Ω_t has full measure. In particular, it holds under our Assumptions I and II.a, as shown in Corollary 5.16.

PROPOSITION 5.6. Consider an N-server network with any arrival process, service times and load balancing algorithm such that $\mathbb{P}\{\Omega_t\} = 1$ for every $t \geq 0$. Then, for every $N \in \mathbb{N}$, $\varphi \in \mathbb{C}_c^{1,1}([0,L) \times \mathbb{R}_+)$, $\ell \geq 1$ and $t \geq 0$,

$$\begin{aligned}\langle \varphi(\cdot,t), \nu_\ell^{(N)}(t)\rangle =& \langle \varphi(\cdot,0), \nu_\ell^{(N)}(0)\rangle + \int_0^t \langle \varphi_s(\cdot,s) + \varphi_x(\cdot,s), \nu_\ell^{(N)}(s)\rangle ds - \mathcal{D}_{\varphi,\ell}^{(N)}(t) \\ &+ \int_{[0,t]} \varphi(0,s) dD_{\ell+1}^{(N)}(s) + \mathcal{R}_{\varphi,\ell}^{(N)}(t),\end{aligned} \tag{5.2.2}$$

almost surely. Moreover, equation (5.2.2) is also valid for $\varphi = \mathbf{1}$, that is,

$$\langle \mathbf{1}, \nu_\ell^{(N)}(t)\rangle = \langle \mathbf{1}, \nu_\ell^{(N)}(0)\rangle - D_\ell^{(N)}(t) + D_{\ell+1}^{(N)}(t) + \mathcal{R}_{\mathbf{1},\ell}^{(N)}(t), \tag{5.2.3}$$

almost surely.

The rest of this section is devoted to the proof of this proposition. Through this section, for ease of notation, by $\sum_j$ we denotes the sum over all job indices $j \in \{j_0, ..., -1, 0, 1, ...\}$.

Fix $\ell \geq 1$, $t \geq 0$. For every $\varphi \in \mathbb{C}_b([0,L) \times [0,\infty))$, since $t \mapsto \langle \varphi(\cdot,t), \nu_\ell^{(N)}(t)\rangle$ is right-continuous, we can write

$$\begin{aligned}&\langle \varphi(\cdot,t), \nu_\ell^{(N)}(t)\rangle - \langle \varphi(\cdot,0), \nu_\ell^{(N)}(0)\rangle \\ &\qquad = \lim_{n\to\infty} \sum_{k=0}^{\lfloor nt \rfloor} \langle \varphi(\cdot, \frac{k+1}{n}), \nu_\ell^{(N)}(\frac{k+1}{n})\rangle - \langle \varphi(\cdot, \frac{k}{n}), \nu_\ell^{(N)}(\frac{k}{n})\rangle \\ &\qquad = \lim_{n\to\infty} \left(\mathcal{I}_1^{(n)} + \mathcal{I}_2^{(n)}\right),\end{aligned} \tag{5.2.4}$$

where

$$\mathcal{I}_1^{(n)} \doteq \lim_{n\to\infty} \sum_{k=0}^{\lfloor nt \rfloor} \langle \varphi(\cdot, \frac{k+1}{n}) - \varphi(\cdot, \frac{k}{n}), \nu_\ell^{(N)}(\frac{k+1}{n}) \rangle \tag{5.2.5}$$

captures the changes in φ and

$$\mathcal{I}_2^{(n)} \doteq \lim_{n\to\infty} \sum_{k=0}^{\lfloor nt \rfloor} \langle \varphi(\cdot, \frac{k}{n}), \nu_\ell^{(N)}(\frac{k+1}{n}) \rangle - \langle \varphi(\cdot, \frac{k}{n}), \nu_\ell^{(N)}(\frac{k}{n}) \rangle \tag{5.2.6}$$

captures the changes in $\nu_\ell^{(N)}$.

We first compute $\mathcal{I}_1^{(n)}$ for $\varphi \in \mathbb{C}_c^{1,1}([0,L) \times \mathbb{R}_+)$. For every $x \in [0,L)$, by the mean value theorem for $\varphi(x,\cdot)$,

$$\varphi(x, \frac{k+1}{n}) - \varphi(x, \frac{k}{n}) = \frac{1}{n}\varphi_s(x, \frac{k+1}{n} - \delta(x,n,k))$$

for some $\delta(x,n,k) \in [0, \frac{1}{n}]$. Therefore, we have

$$\begin{aligned}
&\lim_{n\to\infty} \left| \mathcal{I}_1^{(n)} - \int_0^t \langle \varphi_s(\cdot, s), \nu_\ell^{(N)}(s) \rangle ds \right| \\
&\qquad = \lim_{n\to\infty} \left| \frac{1}{n} \sum_{k=0}^{\lfloor nt \rfloor} \langle \varphi_s(\cdot, \frac{k+1}{n}), \nu_\ell^{(N)}(\frac{k+1}{n}) \rangle - \int_0^t \langle \varphi_s(\cdot, s), \nu_\ell^{(N)}(s) \rangle ds \right| \\
&\qquad + \lim_{n\to\infty} \left| \frac{1}{n} \sum_{k=0}^{\lfloor nt \rfloor} \langle \varphi_s(\cdot, \frac{k+1}{n} - \delta(\cdot,n,k)) - \varphi_s(\cdot, \frac{k+1}{n}), \nu_\ell^{(N)}(\frac{k+1}{n}) \rangle \right|
\end{aligned} \tag{5.2.7}$$

The first term on the right-hand side above is equal to zero because $\nu_\ell^{(N)}$ is right-continuous and φ_s is continuous, and hence $s \mapsto \langle \varphi_s(\cdot, s), \nu_\ell^{(N)}(s) \rangle$ is Riemann integrable in $[0,t]$. Also, since φ is a $\mathbb{C}^1$ function with support in $[0,m] \times [0,T]$ for some $x < L$ and $T < \infty$, φ_s is uniformly continuous, and hence given any $\epsilon > 0$, for all sufficiently large n

$$\left| \varphi_s(x, \frac{k+1}{n} - \delta(x,n,k)) - \varphi_s(x, \frac{k+1}{n}) \right| \le \epsilon, \qquad \forall x \in [0,m], k \ge 1.$$

Therefore, using the bound $\langle 1, \nu_\ell^{(N)}(t) \rangle \le N$,

$$\left| \frac{1}{n} \sum_{k=0}^{\lfloor nt \rfloor} \langle \varphi_s(\cdot, \frac{k+1}{n} - \delta(\cdot,n,k)) - \varphi_s(\cdot, \frac{k+1}{n}), \nu_\ell^{(N)}(\frac{k+1}{n}) \rangle \right| \le (t+1)\epsilon N.$$

Since ϵ is arbitrary, the second term on the right-hand side of (5.2.7) is also equal to zero. Therefore, we have shown that

$$\lim_{n\to\infty} \mathcal{I}_1^{(n)} = \int_0^t \langle \varphi_s(\cdot, s), \nu_\ell^{(N)}(s)\rangle ds. \tag{5.2.8}$$

To compute the limit of $\mathcal{I}_2^{(n)}$, first for every $s, \delta \geq 0$ we express $\nu_\ell^{(N)}(s+\delta)$ in a convenient form using the expression for $\nu_\ell^{(N)}$ given in (5.1.13), and the definition (5.1.5) of $a_j^{(N)}$, which in particular shows that $a_j^{(N)}(s+\delta) = a_j^{(N)}(s) + \delta$ if $\beta_j^{(N)} \leq s \leq s+\delta \leq \gamma_j^{(N)}$, or equivalently, $j \in \mathcal{V}^{(N)}(s) \cap \mathcal{V}^{(N)}(s+\delta)$. Thus, for every $s, \delta \geq 0$,

For $s, \delta \geq 0$,

$$\begin{aligned}
\nu_\ell^{(N)}(s+\delta) &= \sum_j \mathbb{1}\left(j \in \mathcal{V}^{(N)}(s+\delta) \cap \mathcal{U}_\ell^{(N)}(s+\delta)\right) \delta_{a_j^{(N)}(s+\delta)} \\
&= \sum_j \mathbb{1}\left(j \in \mathcal{V}^{(N)}(s+\delta) \cap \mathcal{U}_\ell^{(N)}(s+\delta) \backslash \mathcal{V}^{(N)}(s)\right) \delta_{a_j^{(N)}(s+\delta)} \\
&\quad + \sum_j \mathbb{1}\left(j \in \mathcal{V}^{(N)}(s+\delta) \cap \mathcal{V}^{(N)}(s) \cap \mathcal{U}_\ell^{(N)}(s+\delta) \backslash \mathcal{U}_\ell^{(N)}(s)\right) \delta_{a_j^{(N)}(s)+\delta} \\
&\quad + \sum_j \mathbb{1}\left(j \in \mathcal{V}^{(N)}(s+\delta) \cap \mathcal{V}^{(N)}(s) \cap \mathcal{U}_\ell^{(N)}(s+\delta) \cap \mathcal{U}_\ell^{(N)}(s)\right) \delta_{a_j^{(N)}(s)+\delta},
\end{aligned} \tag{5.2.9}$$

where recall $\sum_j$ denotes the sum over all job indices $j \in \{j_0, ..., -1, 0, 1, ...\}$. In turn, applying the elementary set theoretic identity

$$C \cap \tilde{C} \cap F \cap \tilde{F} = \tilde{C} \cap \tilde{F} \backslash [\tilde{C} \cap \tilde{F} \backslash C] \cup \left([\tilde{C} \cap C \cap \tilde{F}] \backslash F\right)$$

with $C = \mathcal{V}^{(N)}(s+\delta)$, $\tilde{C} = \mathcal{V}^{(N)}(s)$, $F = \mathcal{U}_\ell^{(N)}(s+\delta)$ and $\tilde{F} = \mathcal{U}_\ell^{(N)}(s)$, we have

(5.2.10)

$$\begin{aligned}
&\sum_j \mathbb{1}\left(j \in \mathcal{V}^{(N)}(s+\delta) \cap \mathcal{V}^{(N)}(s) \cap \mathcal{U}_\ell^{(N)}(s+\delta) \cap \mathcal{U}_\ell^{(N)}(s)\right) \delta_{a_j^{(N)}(s)+\delta} \\
&\qquad = \sum_j \mathbb{1}\left(\mathcal{V}^{(N)}(s) \cap \mathcal{U}_\ell^{(N)}(s)\right) \delta_{a_j(s)+\delta} \\
&\qquad\quad - \sum_j \mathbb{1}\left(\mathcal{V}^{(N)}(s) \cap \mathcal{U}_\ell^{(N)}(s) \backslash \mathcal{V}^{(N)}(s+\delta)\right) \delta_{a_j(s)+\delta} \\
&\qquad\quad - \sum_j \mathbb{1}\left(\mathcal{V}^{(N)}(s) \cap \mathcal{U}_\ell^{(N)}(s) \cap \mathcal{V}^{(N)}(s+\delta) \backslash \mathcal{U}_\ell^{(N)}(s+\delta)\right) \delta_{a_j(s)+\delta}
\end{aligned}$$

Note that we have not used any property of $\mathcal{U}_\ell^{(N)}$ in this calculation. This expression can be further simplified when there is only a single arrival or departure occurs during $(s, s+\delta]$, that is, on the set of realizations $\Omega_{s,\delta}$.

LEMMA 5.7. For $\ell \geq 1$ and $\varphi \in \mathbb{C}_b([0,L) \times [0,\infty))$, $s \geq 0$ and $\delta > 0$, on $\Omega_{s,\delta}$,

(5.2.11)
$$\begin{aligned}
\langle \varphi(\cdot, s), \nu_\ell^{(N)}(s+\delta)\rangle &= \langle \varphi(\cdot + \delta, s), \nu_\ell^{(N)}(s)\rangle \\
&\quad - \sum_j \varphi\left(a_j^{(N)}(s) + \delta, s\right) \mathbb{1}\left(\gamma_j^{(N)} \in (s, s+\delta]\right) \mathbb{1}\left(\chi_j^{D,(N)} \geq \ell\right) \\
&\quad + \sum_j \varphi\left(a_j^{(N)}(s+\delta), s\right) \mathbb{1}\left(\beta_j^{(N)} \in (s, s+\delta]\right) \mathbb{1}\left(\chi_j^{K,(N)} \geq \ell\right) \\
&\quad + \mathbb{1}(\ell \geq 2) \sum_j \varphi\left(a_j^{(N)}(s+\delta), s\right) \mathbb{1}\left(\beta_j^{(N)} \leq s\right) \mathbb{1}\left(\gamma_j^{(N)} > s\right) \\
&\qquad \times \mathbb{1}\left(\chi_j^{(N)}(s) = \ell - 1\right) \sum_{j' \geq 1} \mathbb{1}\left(\alpha_{j'}^{(N)} \in (s, s+\delta]\right) \mathbb{1}\left(\kappa_{j'}^{(N)} = \kappa_j^{(N)}\right).
\end{aligned}$$

REMARK 5.8. Since $\chi_j^{D,(N)} \geq 1$ and $\chi_j^{K,(N)} \geq 1$, (see Remark 5.3,) for $\ell = 1$ (5.2.11) simplifies to

$$\begin{aligned}
\langle \varphi(\cdot, s), \nu_1^{(N)}(s+\delta)\rangle =& \langle \varphi(\cdot + \delta, s), \nu_1^{(N)}(s)\rangle \\
& - \sum_j \varphi\left(a_j^{(N)}(s) + \delta, s\right) \mathbb{1}\left(\gamma_j^{(N)} \in (s, s+\delta]\right) \\
& + \sum_j \varphi\left(a_j^{(N)}(s+\delta), s\right) \mathbb{1}\left(\beta_j^{(N)} \in (s, s+\delta]\right).
\end{aligned} \tag{5.2.12}$$

PROOF OF LEMMA 5.7. We first simplify the terms on the right-hand side of (5.2.9). First note that a job j receives service at time $s+\delta$, but not at s if and only if it entered service during $(s, s+\delta]$. Moreover, on $\Omega_{s,\delta}$, there could have been no arrivals in this interval $(s, s+\delta]$ and so, the length of the job's queue is constant in this interval. In particular, $j \in \mathcal{U}_\ell^{(N)}(s+\delta)$ if and only of $\chi_j^{K,(N)} \geq \ell$. Thus,

$$\begin{aligned}
&\sum_j \mathbb{1}\left(j \in \mathcal{V}^{(N)}(s+\delta) \backslash \mathcal{V}^{(N)}(s) \cap \mathcal{U}_\ell^{(N)}(s+\delta)\right) \delta_{a_j^{(N)}(s+\delta)} \\
&\qquad = \sum_j \mathbb{1}\left(\beta_j^{(N)} \in (s, s+\delta]\right) \mathbb{1}\left(\chi_j^{K,(N)} \geq \ell\right) \delta_{a_j^{(N)}(s+\delta)}.
\end{aligned} \tag{5.2.13}$$

The analysis of the second term on the right-hand side of (5.2.9) is slightly different for $\ell = 1$ and $\ell \geq 2$. If a job j received service throughout the period $(s, s+\delta]$, that is $j \in \mathcal{V}^{(N)}(s) \cap \mathcal{V}^{(N)}(s+\delta)$, then the corresponding queue could not have been empty at time s, and on $\Omega_{s,\delta}$, the difference between the queue length at time $s+\delta$ and time s is either zero (if there were no arrivals to that queue) or one (if there were precisely one arrival to that queue.) Therefore, when $\ell = 1$,

$$\mathcal{V}^{(N)}(s) \cap \mathcal{V}^{(N)}(s+\delta) \cap \mathcal{U}_1^{(N)}(s+\delta) \backslash \mathcal{U}_1^{(N)}(s) = \emptyset, \tag{5.2.14}$$

and for $\ell = 2$, using the representation (5.1.16) of $E^{(N),i}$,

(5.2.15)

$$\begin{aligned}
&\sum_j \mathbb{1}\left(j \in \mathcal{V}^{(N)}(s) \cap \mathcal{V}^{(N)}(s+\delta) \cap \mathcal{U}_\ell^{(N)}(s+\delta) \backslash \mathcal{U}_\ell^{(N)}(s)\right) \delta_{a_j^{(N)}(s)+\delta} \\
&\quad = \sum_j \mathbb{1}\left(\beta_j^{(N)} \leq s\right) \mathbb{1}\left(\gamma_j^{(N)} > s\right) \mathbb{1}\left(\chi_j^{(N)}(s) = \ell - 1\right) E^{(N),\kappa_j^{(N)}}(s, s+\delta] \delta_{a_j^{(N)}(s+\delta)} \\
&\quad = \sum_j \mathbb{1}\left(\beta_j^{(N)} \leq s\right) \mathbb{1}\left(\gamma_j^{(N)} > s\right) \mathbb{1}\left(\chi_j^{(N)}(s) = \ell - 1\right) \\
&\qquad\qquad \times \delta_{a_j^{(N)}(s+\delta)} \sum_{j' \geq 1} \mathbb{1}\left(\alpha_{j'}^{(N)} \in (s, s+\delta], \kappa_{j'}^{(N)} = \kappa_j^{(N)}\right).
\end{aligned}$$

For the third term on the right-hand side of (5.2.9), we consider the right-hand side of (5.2.10). First note that for every $\varphi \in \mathbb{C}_c^{1,1}([0, L) \times \mathbb{R}_+)$, by definition (5.1.13) of $\nu_\ell^{(N)}$,

$$\langle \varphi(\cdot, s), \sum_j \mathbb{1}\left(j \in \mathcal{V}^{(N)}(s) \cap \mathcal{U}_\ell^{(N)}(s)\right) \delta_{a_j^{(N)}(s)+\delta} \rangle = \langle \varphi(\cdot + \delta), s), \nu_\ell^{(N)}(s) \rangle. \tag{5.2.16}$$

Next, note that a job j has been departed from a queue during $(s, s+\delta]$ if and only if $j \in \mathcal{V}^{(N)}(s+\delta) \backslash \mathcal{V}^{(N)}(s)$. Moreover, on $\Omega_{s,\delta}$ there were no other arrivals during $(s, s+\delta]$, and hence, the queue length was constant on $(s, \gamma_j^{(N)}-)$. Therefore, $j \in \mathcal{U}_\ell^{(N)}(s)$ if and only if $\chi_j^{D,(N)} \geq \ell$. Hence, we have

$$\begin{aligned}
&\sum_j \mathbb{1}\left(j \in \mathcal{V}^{(N)}(s) \backslash \mathcal{V}^{(N)}(s+\delta) \cap \mathcal{U}_\ell^{(N)}(s)\right) \delta_{a_j^{(N)}(s)+\delta} \\
&\qquad = \sum_j \mathbb{1}\left(\gamma_j^{(N)} \in (s, s+\delta]\right) \mathbb{1}\left(\chi_j^{D,(N)} \geq \ell\right) \delta_{a_j^{(N)}(s)+\delta}.
\end{aligned} \tag{5.2.17}$$

Finally for the last term on the right-hand side of (5.2.10), note that if a job j receives service at a queue during $(s, s+\delta]$, then that queue length is non-decreasing on that interval. therefore,

$$\mathcal{V}^{(N)}(s) \cap \mathcal{V}^{(N)}(s+\delta) \cap \mathcal{U}_\ell^{(N)}(s) \backslash \mathcal{U}_\ell^{(N)}(s+\delta) = \emptyset. \tag{5.2.18}$$

The result follows from the representations (5.2.9)-(5.2.10) and equalities (5.2.13)-(5.2.18). $\square$

We continue with the identification of the limit of $\mathcal{I}_2^{(n)}$ using Lemma 5.7. Recall the definition (5.2.1) of Ω_t. Since $\mathbb{P}\{\Omega_t\} = 1$ by the assumptions of the theorem, there exists $n_0 \in \mathbb{N}$ such that almost surely, the identity (5.2.11) holds with $\delta = 1/n$ and $s = k/n$ simultaneously for every $n \geq n_0$ and $k = 0, 1, ..., \lfloor nt \rfloor$. Substitute the representation in (5.2.11), with $\delta = 1/n$ and $s = k/n$ into the expression (5.2.6) for $\mathcal{I}_2^{(n)}$, almost surely, for every $\varphi \in \mathbb{C}_b([0, L) \times [0, \infty))$ we have

(5.2.19)

$$\begin{aligned}
\lim_{n\to\infty} \mathcal{I}_2^{(n)} &= \lim_{n\to\infty} \sum_{k=0}^{\lfloor nt \rfloor} \left[\left\langle \varphi(\cdot + \frac{1}{n}, \frac{k}{n}) - \varphi(\cdot, \frac{k}{n}), \nu_\ell^{(N)}(\frac{k}{n}) \right\rangle \right] \\
&\quad - \lim_{n\to\infty} \sum_{k=0}^{\lfloor nt \rfloor} \sum_j \varphi\left(a_j^{(N)}(\frac{k}{n}) + \frac{1}{n}, \frac{k}{n} \right) \mathbb{1}\left(\gamma_j^{(N)} \in (\frac{k}{n}, \frac{k+1}{n}] \right) \mathbb{1}\left(\chi_j^{D,(N)} \geq \ell \right) \\
&\quad + \lim_{n\to\infty} \sum_{k=0}^{\lfloor nt \rfloor} \sum_j \varphi\left(a_j^{(N)}(\frac{k+1}{n}), \frac{k}{n} \right) \mathbb{1}\left(\beta_j^{(N)} \in (\frac{k}{n}, \frac{k+1}{n}] \right) \mathbb{1}\left(\chi_j^{K,(N)} \geq \ell \right) \\
&\quad + \mathbb{1}(\ell \geq 2) \lim_{n\to\infty} \sum_{k=0}^{\lfloor nt \rfloor} \sum_j \sum_{j' \geq 1} \varphi\left(a_j^{(N)}(\frac{k+1}{n}), \frac{k}{n} \right) \mathbb{1}\left(\beta_j^{(N)} \leq \frac{k}{n} \right) \\
&\qquad\qquad \times \mathbb{1}\left(\gamma_j^{(N)} > \frac{k}{n} \right) \mathbb{1}\left(\chi_j^{(N)}(\frac{k}{n}) = \ell - 1 \right) \\
&\qquad\qquad \times \mathbb{1}\left(\alpha_{j'}^{(N)} \in (\frac{k}{n}, \frac{k+1}{n}] \right) \mathbb{1}\left(\kappa_{j'}^{(N)} = \kappa_j^{(N)} \right).
\end{aligned}$$

For $\varphi \in \mathbb{C}_c^{1,1}([0, L) \times \mathbb{R}_+)$, the first term on the right-hand side of (5.2.19) can be characterized using computations exactly analogous to the derivation of limit for $\mathcal{I}_1^{(n)}$ in (5.2.5), (5.2.7) and (5.2.8), as follows:

$$\lim_{n\to\infty}\sum_{k=0}^{\lfloor nt\rfloor}\left[\left\langle \varphi(\cdot+\frac{1}{n},\frac{k}{n})-\varphi(\cdot,\frac{k}{n}),\nu_\ell^{(N)}(\frac{k}{n})\right\rangle\right]=\int_0^t\langle\varphi_x(\cdot,s),\nu_\ell^{(N)}(s)\rangle ds. \tag{5.2.20}$$

For the second term, changing the order of summation and setting $\gamma_{j,n}^{(N)} \doteq \frac{1}{n}\lfloor n\gamma_j^{(N)}\rfloor$, we have (as justified below),

(5.2.21)

$$\begin{aligned}
&\lim_{n\to\infty}\sum_{k=0}^{\lfloor nt\rfloor}\sum_j\varphi\left(a_j^{(N)}(\frac{k}{n})+\frac{1}{n},\frac{k}{n}\right)\mathbb{1}\left(\gamma_j^{(N)}\in(\frac{k}{n},\frac{k+1}{n}]\right)\mathbb{1}\left(\chi_j^{D,(N)}\geq\ell\right)\\
&=\lim_{n\to\infty}\sum_j\sum_{k=0}^{\lfloor nt\rfloor}\varphi\left(a_j^{(N)}(\frac{k}{n})+\frac{1}{n},\frac{k}{n}\right)\mathbb{1}\left(\gamma_j^{(N)}\in(\frac{k}{n},\frac{k+1}{n}]\right)\mathbb{1}\left(\chi_j^{D,(N)}\geq\ell\right)\\
&=\lim_{n\to\infty}\sum_j\varphi\left(a_j^{(N)}(\gamma_{j,n}^{(N)})+\frac{1}{n},\gamma_{j,n}^{(N)}\right)\mathbb{1}\left(\gamma_j^{(N)}\leq\frac{\lceil nt\rceil}{n}\right)\mathbb{1}\left(\chi_j^{D,(N)}\geq\ell\right)\\
&=\sum_j\varphi\left(v_j,\gamma_j^{(N)}\right)\mathbb{1}\left(\gamma_j^{(N)}\leq t\right)\mathbb{1}\left(\chi_j^{D,(N)}\geq\ell\right)\\
&=\mathcal{D}_{\varphi,\ell}^{(N)}(t).
\end{aligned}$$

Here, the third equality uses the facts that $\gamma_{j,n}^{(N)}\uparrow\gamma_j^{(N)}$ as $n\to\infty$ and $a_j^{(N)}(\gamma_j^{(N)})=v_j$ and the continuity of φ and $a_j^{(N)}$. The last equality follows from definition (5.1.23) of $\mathcal{D}_{\varphi,\ell}^{(N)}$.

For the third term on the right-hand side of (5.2.19), changing the order of summation and now setting $\beta_{j,n}^{(N)} \doteq \frac{1}{n}\lfloor n\beta_j^{(N)} \rfloor$, we obtain

(5.2.22)

$$\begin{aligned}
&\lim_{n\to\infty}\sum_{k=0}^{\lfloor nt\rfloor} \sum_j \varphi\left(a_j^{(N)}(\frac{k+1}{n}),\frac{k}{n}\right)\mathbb{1}\left(\beta_j^{(N)}\in(\frac{k}{n},\frac{k+1}{n}]\right)\mathbb{1}\left(\chi_j^{K,(N)}\geq\ell\right)\\
&\quad=\lim_{n\to\infty}\sum_j \sum_{k=0}^{\lfloor nt\rfloor}\varphi\left(a_j^{(N)}(\frac{k+1}{n}),\frac{k}{n}\right)\mathbb{1}\left(\beta_j^{(N)}\in(\frac{k}{n},\frac{k+1}{n}]\right)\mathbb{1}\left(\chi_j^{K,(N)}\geq\ell\right)\\
&\quad=\lim_{n\to\infty}\sum_j \varphi\left(a_j^{(N)}(\beta_{j,n}^{(N)}+\frac{1}{n}),\beta_{j,n}^{(N)}\right)\mathbb{1}\left(0\leq\beta_j^{(N)}\leq\frac{\lceil nt\rceil}{n}\right)\mathbb{1}\left(\chi_j^{K,(N)}\geq\ell\right)\\
&\quad=\sum_j\varphi\left(0,\beta_j^{(N)}\right)\mathbb{1}\left(0\leq\beta_j^{(N)}\leq t\right)\mathbb{1}\left(\chi_j^{K,(N)}\geq\ell\right),
\end{aligned}$$

where the last equality uses the facts that $a_j^{(N)}(\beta_j^{(N)}) = 0$ and $\beta_{j,n}^{(N)} \uparrow \beta_j^{(N)}$ as $n \to \infty$ and the continuity of φ and $a_j^{(N)}(\cdot)$. Further simplification of (5.2.22) is slightly different for $\ell = 1$ and $\ell \geq 2$. When $\ell \geq 2$, due to the non-idling assumption, the service entry time of any job to a queue of length at least ℓ just after service entry coincides with the departure time of another job from the same queue, which had the length of at least $\ell + 1$ just before the departure. Therefore, for $\ell \geq 2$, by definition (5.1.21) of $D_\ell^{(N)}$ we have

$$\begin{aligned}
&\sum_j\varphi\left(0,\beta_j^{(N)}\right)\mathbb{1}\left(0\leq\beta_j^{(N)}\leq t\right)\mathbb{1}\left(\chi_j^{K,(N)}\geq\ell\right)\\
&\qquad=\sum_j\varphi\left(0,\gamma_j^{(N)}\right)\mathbb{1}\left(\gamma_j^{(N)}\leq t\right)\mathbb{1}\left(\chi_j^{D,(N)}\geq\ell+1\right)\\
&\qquad=\int_{[0,t]}\varphi(0,s)dD_{\ell+1}^{(N)}(s). \qquad (5.2.23)
\end{aligned}$$

To simplify (5.2.22) for $\ell = 1$, we first define the cumulative service entry process $\{K^{(N)}(t); t \geq 0\}$ where $K^{(N)}(t)$ is the cumulative number of jobs that entered service in the interval $[0, t]$, that is

$$K^{(N)}(t) \doteq \sum_{j=j_0}^{\infty}\mathbb{1}\left(0\leq\beta_j^{(N)}\leq t\right). \qquad (5.2.24)$$

The following lemma gives a representation for $K^{(N)}$ in terms of the process $D_2^{(N)}$ defined in (5.1.21) and queue length $X^{(N),i}$ and arrival processes $E^{(N),i}$ to queues $i = 1, ..., N$, defined and (5.1.16).

LEMMA 5.9. For all $t \geq 0$, $K^{(N)}(t)$ satisfies

$$K^{(N)}(t) = D_2^{(N)}(t) + \sum_{i=1}^{N} \int_0^t \mathbb{1}\left(X^{(N),i}(u-) = 0\right) dE^{(N),i}(u) \tag{5.2.25}$$

PROOF. Service entries can be classified into two types, based on the status of the queue just prior to service entry. Thus we can expand (5.2.24), based on whether or not the queue was empty right before to the service entry, as

$$\begin{aligned} K^{(N)}(t) = & \sum_j \mathbb{1}\left(0 \leq \beta_j^{(N)} \leq t\right) \mathbb{1}\left(\chi_j^{(N)}(\beta_j^{(N)}-) \geq 1\right) \\ & + \sum_j \mathbb{1}\left(0 \leq \beta_j^{(N)} \leq t\right) \mathbb{1}\left(\chi_j^{(N)}(\beta_j^{(N)}-) = 0\right). \end{aligned}$$

Due to the non-idling assumption, the service entry time of of the first type coincides with the departure time of another job from the same queue, which had length of at least 2 just before departure. On the other hand, a service entry time of the second type coincides with the arrival time of the same job (to a queue that was empty right before arrival). Therefore, recalling that $\kappa_j^{(N)}$ is the queue index of job j, we can then write

$$\begin{aligned} K^{(N)}(t) &= \sum_j \mathbb{1}\left(\gamma_j^{(N)} \leq t\right) \mathbb{1}\left(\chi_j^{D,(N)} \geq 2\right) + \sum_{j \geq 1} \mathbb{1}\left(\alpha_j^{(N)} \leq t\right) \mathbb{1}\left(\chi_j^{E,(N)} = 1\right) \\ &= \sum_j \mathbb{1}\left(\gamma_j^{(N)} \leq t\right) \mathbb{1}\left(\chi_j^{D,(N)} \geq 2\right) \\ &\quad + \sum_{i=1}^{N} \sum_{j=1}^{\infty} \mathbb{1}\left(X^{(N),i}(\alpha_j^{(N)}) = 0\right) \mathbb{1}\left(\alpha_j^{(N)} \leq t\right) \mathbb{1}\left(\kappa_j^{(N)} = i\right). \end{aligned}$$

Equation (5.2.25) then follows from the definitions $D_2^{(N)}$ and $E^{(N),i}$ from (5.1.21) and (5.1.16), respectively. $\square$

By definition (5.2.24) of $K^{(N)}$ and the fact that $\mathbb{1}(\chi_j^{K,(N)} \geq 1) = 1$ (see Remark 5.3), representation (5.2.25) and definition (5.1.19) of $\mathcal{R}_{\varphi,1}^{(N)}$, we have

(5.2.26)

$$\begin{aligned}\sum_j \varphi\left(0, \beta_j^{(N)}\right) \mathbb{1}\left(0 \leq \beta_j^{(N)} \leq t\right) \mathbb{1}\left(\chi_j^{K,(N)} \geq 1\right) &= \int_{[0,t]} \varphi(0,u) dK^{(N)}(u) \\ &= \int_{[0,t]} \varphi(0,u) dD_2^{(N)}(u) + \sum_{j=1}^{N} \int_0^t \varphi(0,u) \mathbb{1}\left(X^{(N),i}(u-) = 0\right) dE^{(N),i}(u) \\ &= \int_{[0,t]} \varphi(0,u) dD_2^{(N)}(u) + \mathcal{R}_{\varphi,1}^{(N)}.\end{aligned}$$

Finally, the last term on the right-hand side of (5.2.19) is zero for $\ell = 1$. For $\ell \geq 2$, changing the order of summation, setting $\alpha_{j',n}^{(N)} = \frac{1}{n}\lfloor n\alpha_{j'}^{(N)} \rfloor$ and noting that on Ω_t, the arrival of j' should be the only event happening the last term on the right-hand side of (5.2.19) is equal to

$$\begin{aligned}\lim_{n\to\infty} \sum_{j'\geq 1} \sum_j \varphi\left(a_j^{(N)}(\alpha_{j',n}^{(N)} + \frac{1}{n}), \alpha_{j',n}^{(N)}\right) \mathbb{1}\left(\beta_j^{(N)} \leq \alpha_{j',n}^{(N)}\right) \mathbb{1}\left(\gamma_j^{(N)} > \alpha_{j',n}^{(N)}\right) \\ \times \mathbb{1}\left(\chi_j^{(N)}(\alpha_{j',n}^{(N)}) = \ell - 1\right) \mathbb{1}\left(\alpha_{j'}^{(N)} \leq \frac{\lceil nt \rceil}{n}\right) \mathbb{1}\left(\kappa_{j'}^{(N)} = \kappa_j^{(N)}\right).\end{aligned}$$

Since $\alpha_{j',n}^{(N)} \uparrow \alpha_{j'}^{(N)}$ as $n \to \infty$, the fact that on Ω_t, $\beta_j^{(N)}, \gamma_j^{(N)} \neq \alpha_{j'}^{(N)}$ for all $j \neq j'$, and by the continuity of φ and $a_j^{(N)}$, the last display is equal to

$$\begin{aligned}(5.2.27) \quad \sum_{j'\geq 1} \sum_j \varphi\left(a_j^{(N)}(\alpha_{j'}^{(N)}), \alpha_{j'}^{(N)}\right) \mathbb{1}\left(\beta_j^{(N)} < \alpha_{j'}^{(N)}\right) \mathbb{1}\left(\gamma_j^{(N)} \geq \alpha_{j'}^{(N)}\right) \\ \times \mathbb{1}\left(\chi_j^{(N)}(\alpha_{j'}^{(N)}-) = \ell - 1\right) \mathbb{1}\left(\alpha_{j'}^{(N)} \leq t\right) \mathbb{1}\left(\kappa_{j'}^{(N)} = \kappa_j^{(N)}\right).\end{aligned}$$

Summing over all the queues i, and all the arrivals j' to the queue, and using definition (5.1.17) of $\nu_\ell^{(N),i}$, we can re-write the last display as

(5.2.28)

$$\begin{aligned}
&\sum_{i=1}^{N} \sum_{j'\geq 1} \mathbb{1}\left(\kappa_{j'}^{(N)}=i\right)\mathbb{1}\left(\alpha_{j'}^{(N)}\leq t\right)\sum_j \varphi\left(a_j^{(N)}(\alpha_{j'}^{(N)}),\alpha_{j'}^{(N)}\right)\mathbb{1}\left(\beta_j^{(N)}<\alpha_{j'}^{(N)}\right)\\
&\qquad\qquad \times \mathbb{1}\left(\gamma_j^{(N)}\geq \alpha_{j'}^{(N)}\right)\mathbb{1}\left(\chi_j^{(N)}(\alpha_{j'}^{(N)}-)=\ell-1\right)\mathbb{1}\left(\kappa_j^{(N)}=i\right)\\
&=\sum_{i=1}^{N}\sum_{j'\geq 1}\mathbb{1}\left(\kappa_{j'}^{(N)}=i\right)\mathbb{1}\left(\alpha_{j'}^{(N)}\leq t\right)\langle\varphi(\cdot,\alpha_{j'}^{(N)}),\nu_{\ell-1}^{(N),i}(\alpha_{j'}^{(N)}-)-\nu_\ell^{(N),i}(\alpha_{j'}^{(N)}-)\rangle\\
&=\sum_{i=1}^{N}\int_{(0,t]}\langle\varphi(\cdot,s),\nu_{\ell-1}^{(N),i}(s-)-\nu_\ell^{(N),i}(s-)\rangle dE^{(N),i}(s)\\
&=\mathcal{R}_{\varphi,\ell}^{(N)}(t).
\end{aligned}$$

where $E^{(N),i}$ is the cumulative arrival process to queue i, defined in (5.1.16).

Now we can conclude the proof.

PROOF OF PROPOSITION 5.6. Equation (5.2.2) then follows from equations (5.2.4), (5.2.8), (5.2.19)-(5.2.23), (5.2.26) and (5.2.28). Furthermore, to establish the mass balance equation (5.2.3), note that for $\varphi = \mathbf{1}$, $\mathcal{I}_1^{(n)}(t)$ and the first term on the right-hand side of (5.2.19) are zero for all $t \geq 0$. Since Lemma 5.7 and the calculation of other terms on the right-hand side of (5.2.19) are valid for all $\varphi \in \mathbb{C}_b([0,L)\times[0,\infty))$, (5.2.3) follows on setting $\varphi = \mathbf{1}$ in (5.2.2). □

REMARK 5.10. Equation (5.2.2) remains valid for functions φ on $[0,L)\times\mathbb{R}_+$ of the form $\varphi(x,s) = f(x)$ for some $f \in \mathbb{C}_c^1[0,L)$. For such functions, $\mathcal{I}_1^{(n)}$ is identically zero, and the calculation of $\mathcal{I}_2^{(n)}$ remains unchanged.

5.3. Martingale Decomposition for Routing and Departure Processes

In this section, we establish a martingale decomposition result for the φ-weighted routing process $\mathcal{R}_{\varphi,\ell}^{(N)}$ and the departure process $\mathcal{D}_{\varphi,\ell}^{(N)}$ defined in (5.1.19) and (5.1.23), respectively. We first state the result s in Section 5.3.1. These results are specific to our assumptions on Arrival process and load balancing algorithm, although they can be easily adapted to different models. In Section 5.3.2, we introduce a marked point process which provides an alternative characterization of the the filtration $\{\mathcal{F}_t\}$, which facilitates the proofs of the results in this section. The decomposition result for routing processes is proved in Section 5.3.3, and the proof for the departure process, which is similar to those in [**61**, Lemma 5.4], is given in Section 5.3.4.

5.3.1. The form of Compensators. For any $N \geq 1$, $\ell \geq 1$ and $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$, define the process $\{B^{(N)}_{\varphi,\ell}(t); t \geq 0\}$ as the following: for $t \geq 0$,

$$B^{(N)}_{\varphi,1}(t) \doteq \int_0^t h_E^{(N)}\left(R^{(N)}(u)\right)\varphi(0,u)(1-(\overline{S}^{(N)}_1(u))^d)du \tag{5.3.1}$$

and for $\ell \geq 2$,

$$B^{(N)}_{\varphi,\ell}(t) \doteq \int_0^t h_E^{(N)}\left(R^{(N)}(u)\right)\mathfrak{P}_d\left(\overline{S}^{(N)}_{\ell-1}(u), \overline{S}^{(N)}_{\ell}(u)\right)\langle\varphi(\cdot,u)\, , \overline{\nu}^{(N)}_{\ell-1}(u) - \overline{\nu}^{(N)}_{\ell}(u)\rangle du. \tag{5.3.2}$$

Here recall that $h_E^{(N)}$ is the hazard rate function of the inter-arrival time distribution $G_E^{(N)}$.

PROPOSITION 5.11. Suppose Assumptions I, II.a, and III.a hold. Then, for every $N \geq 1$, $\ell \geq 1$ and $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$, the process $\mathcal{N}^{(N)}_{\varphi,\ell}$ defined as

$$\mathcal{N}^{(N)}_{\varphi,\ell} \doteq \mathcal{R}^{(N)}_{\varphi,\ell} - B^{(N)}_{\varphi,\ell}, \tag{5.3.3}$$

is a local $\{\mathcal{F}^{(N)}_t\}$-martingale, and its quadratic variation is given by

$$[\mathcal{N}^{(N)}_{\varphi,\ell}](t) = \mathcal{R}^{(N)}_{\varphi^2,\ell}(t), \qquad t \geq 0. \tag{5.3.4}$$

The proof of Proposition 5.11 is given in Section 5.3.3.

Moreover, for every $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$, define the process $\{A^{(N)}_{\varphi,\ell}(t); t \geq 0\}$ by

$$A^{(N)}_{\varphi,\ell}(t) \doteq \int_0^t \langle\varphi(\cdot,s)h(\cdot), \nu^{(N)}_\ell(s)\rangle ds, \qquad \forall t \geq 0. \tag{5.3.5}$$

Using an argument exactly analogous to that used to obtain equations (5.24) and (5.25) in **[61]**, we can show that for any $\varphi \in \mathbb{C}_c^{1,1}([0,L) \times \mathbb{R}_+)$ with $\mathrm{supp}(\varphi) \subset [0,m] \times \mathbb{R}_+$, $m < L$,

$$|A_{\varphi,\ell}^{(N)}(t)| \leq \|\varphi\|_\infty \left(X^{(N)}(0) + E^{(N)}(t)\right)\left(\int_0^m h(x)dx\right), \qquad t \geq 0. \tag{5.3.6}$$

PROPOSITION 5.12. Suppose Assumptions I, II.a, and III.a hold. Then, for every $N \in \mathbb{N}$, $\ell \geq 1$ and $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$, the process $\mathcal{M}_{\varphi,\ell}^{(N)}$ defined as

$$\mathcal{M}_{\varphi,\ell}^{(N)} \doteq \mathcal{D}_{\varphi,\ell}^{(N)} - A_{\varphi,\ell}^{(N)}, \tag{5.3.7}$$

is a local $\{\mathcal{F}_t^{(N)}\}$-martingale, and its quadratic variation is given by

$$[\mathcal{M}_{\varphi,\ell}^{(N)}](t) = \mathcal{D}_{\varphi^2,\ell}^{(N)}(t), \qquad t \geq 0. \tag{5.3.8}$$

The result in Proposition 5.12 is similar in flavor to a result established in the proof of Lemma 5.4 in **[61]**. However, the model, and more importantly, the filtration used in **[61]** is different from the one used here, so the result of **[61]** does not directly apply. Instead, we adopt a completely different approach to the proof, which we believe would be more useful in the analysis of various other randomized load balancing schemes. The proof is given in Section 5.3.4.

REMARK 5.13. Substituting the martingale decompositions (5.3.3) and (5.3.7), and the definition (5.3.5) of $A_{\varphi,\ell}^{(N)}$ in (5.2.2), we have

$$\begin{aligned}\langle \varphi(\cdot,t), \nu_\ell^{(N)}(t)\rangle =& \langle \varphi(\cdot,0), \nu_\ell^{(N)}(0)\rangle + \int_0^t \langle \varphi_s(\cdot,s) + \varphi_x(\cdot,s) - \varphi(\cdot,s)h(\cdot), \nu_\ell^{(N)}(s)\rangle ds \\ &+ \int_{[0,t]} \varphi(0,s)dD_{\ell+1}^{(N)}(s) + B_{\varphi,\ell}^{(N)}(t) - \mathcal{M}_{\varphi,\ell}^{(N)}(t) + \mathcal{N}_{\varphi,\ell}^{(N)}(t).\end{aligned} \tag{5.3.9}$$

Before proceeding to the proofs of Propositions 5.11 and 5.12, we recall the definition of the (stochastic) intensity of a point process, and state an elementary lemma which is later used in those proofs. Suppose $(\Omega, \mathcal{G}, \{\mathcal{G}_t\}, \mathbb{P})$ is a filtered probability space which satisfies the usual conditions, and let $\{\xi(t); t \geq 0\}$ be a point process adapted to the filtration $\{\mathcal{G}_t\}$. A non-negative $\{\mathcal{G}_t\}$-progressive process $\{u(t); t \geq 0\}$ is called a $\{\mathcal{G}_t\}$-intensity of ξ if for all $t \geq 0$,

$$\int_0^t u(s)ds < \infty, \qquad \text{a. s.,} \tag{5.3.10}$$

and for every non-negative $\{\mathcal{G}_t\}$-predictable processes H,

$$\mathbb{E}\left[\int_0^\infty H(t)d\xi(t)\right] = \mathbb{E}\left[\int_0^\infty H(s)u(s)ds\right]. \tag{5.3.11}$$

LEMMA 5.14. Suppose $(\Omega, \mathcal{G}, \{\mathcal{G}_t\}, \mathbb{P})$ and the point process $\{\xi(t); t \geq 0\}$ are as above, and let u be a $\{\mathcal{G}_t\}$-intensity of ξ. Given a locally bounded, $\{\mathcal{G}_t\}$-predictable process θ, define the stochastic process $\{\zeta(t); t \geq 0\}$ by $\zeta(t) \doteq \int_0^t \theta(s)d\xi(s)$. Then

$$\zeta(t) - \int_0^t \theta(s)u(s)ds, \qquad t \geq 0,$$

is a local $\{\mathcal{G}_t\}$-martingale, and the quadratic variation of ζ is given by

$$[\zeta](t) = \int_0^t \theta^2(s)d\xi(s), \qquad t \geq 0. \tag{5.3.12}$$

PROOF. First, note that by (5.3.10) and the local boundedness of θ, for all $t \geq 0$,

$$\int_0^t |\theta(s)|u(s)ds \leq \sup_{s\in[0,t]} |\theta(s)| \int_0^t u(s)ds < \infty, \qquad a.s.$$

The first claim then follows from [**14**, Lemma II.L3]. For the second claim, since ξ is a pure jump process with unit jumps, ζ is also a pure jump process with jumps

$$\zeta(s) - \zeta(s-) = \theta(s)(\xi(s) - \xi(s-)). \tag{5.3.13}$$

Therefore, the quadratic variation of ζ is given by (see (18.1) in Chapter IV of [**90**])

$$[\zeta](t) = \sum_{0<s\leq t} (\zeta(s) - \zeta(s-))^2 = \sum_{0<s\leq t} \theta^2(s)(\xi(s) - \xi(s-)) = \int_0^t \theta^2(s)d\xi(s). \tag{5.3.14}$$

$\square$

5.3.2. A Marked Point Process Representation. In this section, we construct a point process $\mathcal{T}^{(N)}$ consisting of all arrival and departure times, marked by their type and their corresponding queue index. This point process has the property that its natural filtration, together with the σ-algebra generated by initial conditions, is equivalent to the filtration $\{\mathcal{F}_t^{(N)}; t \geq 0\}$ defined (5.1.27), and that all auxiliary processes defined in Section 5.1.3 has an intuitive representation as an integral with respect to $\mathcal{T}^{(N)}$. This allows us to directly apply standard compensator results (e.g. from [**14**]).

Consider the set

$$\mathbb{T}^{(N)} \doteq \left\{(\alpha_j^{(N)}, (\mathfrak{E}, \kappa_j^{(N)})); j \geq 1\right\} \cup \left\{(\gamma_j^{(N)}, (\mathfrak{D}, \kappa_j^{(N)})); j \geq j_0\right\},$$

which is the union of all arrival times $\alpha_j^{(N)}$, marked by the tag $\mathfrak{E}$ (indicating that it is an arrival time) and the queue index to which job j is routed, and all departure times $\gamma_j^{(N)}$, marked by the tag $\mathfrak{D}$ (indicating that it is a departure time) and the queue index from which job j has departed. Let $\tau_0 \doteq 0$ and z_0 be a constant (whose value is irrelevant), and define the sequence of *events* $\{(\tau_k^{(N)}, z_k^{(N)}); k \geq 1\}$, each composed of an *event time* $\tau_k^{(N)}$ and an *event mark* $z_k^{(N)}$, to be the relabeling (i.e. a one to one correspondence) of $\mathbb{T}^N$ sorted by lexicographic order, assuming $\mathfrak{D} < \mathfrak{E}$. That is, events are ordered first by event times ($\tau_k^{(N)} \leq \tau_{k+1}^{(N)}$), then by event type (departure first, then arrival) and finally by queue index (smaller indices first.) Note that since the inter-arrival distribution $G_E^{(N)}$ and service distribution G are absolutely continuous with respect to Lebesgue measure by Assumption I and Assumption II.a, at most one arrival and at most one departure from each queue can occur at any given time, almost surely. Let $\mathcal{T}^{(N)} = \{\mathcal{T}^{(N)}(t); t \geq 0\}$ be the corresponding marked point process. Clearly, for every index $i \in \{1, ..., N\}$,

$$\mathcal{T}^{(N)}(\mathfrak{E}, i) \doteq \sum_{k \geq 1} \mathbb{1}\left(\tau_k^{(N)} \leq t\right) \mathbb{1}\left(z_k^{(N)} = (\mathfrak{E}, i)\right) = E^{(N),i}(t), \tag{5.3.15}$$

and

$$\mathcal{T}^{(N)}(\mathfrak{D}, i) \doteq \sum_{k \geq 1} \mathbb{1}\left(\tau_k^{(N)} \leq t\right) \mathbb{1}\left(z_k^{(N)} = (\mathfrak{D}, i)\right) = D^{(N),i}(t). \tag{5.3.16}$$

Therefore, the filtration $\{\mathcal{F}_t^{(N)}\}$ defined in (5.1.27) has the following equivalent representation:

$$\mathcal{F}_t^{(N)} = \sigma(I_0) \vee \mathcal{F}_t^{\mathcal{T},(N)}, \qquad t \geq 0, \tag{5.3.17}$$

where $\mathcal{F}_t^{\mathcal{T},(N)}$ is the filtration generated by the marked point process $\mathcal{T}^{(N)}$ (see e.g. [**14**, Equation (1.2) in page 57]).

At any time t, a server i is called *busy* if $X^{(N),i}(t) \geq 1$, and is called *idle* otherwise. Note that by non-idling assumption, there is a job receiving service at any busy server i at time t, and $a^{(N),i}(t)$ is defined to be the age of that job. Using this notation, we can re-write the definition (5.1.17) of $\nu_\ell^{(N),i}$ as

$$\nu_\ell^{(N),i}(t) = \mathbb{1}\left(X^{(N),i}(t) \geq \ell\right)\delta_{a^{(N),i}(t)}, \tag{5.3.18}$$

and when combined with (5.1.18), we have

$$\nu_\ell^{(N)}(t) = \sum_{i=1}^{N} \mathbb{1}\left(X^{(N),i}(t) \geq \ell\right)\delta_{a^{(N),i}(t)}. \tag{5.3.19}$$

For $k \geq 0$, define

$$\mathfrak{B}_k^{(N)} \doteq \left\{i : X^{(N),i}(\tau_k^{(N)}) \geq 1\right\} \tag{5.3.20}$$

to be the set of busy server at time $\tau_k^{(N)}$, and note that for $i \in \mathfrak{B}_k^{(N)}$, $a^{(N),i}(\tau_k^{(N)})$ is well-defined. Define $\xi_{k+1}^{(N)}$ to be the next arrival time after $\tau_k^{(N)}$, and for $i = 1, ..., N$, define $\sigma_{k+1}^{(N),i}$ to be the next departure time from queue i if $i \in \mathfrak{B}_k$, and $\sigma_{k+1}^{(N),i} = \infty$, otherwise. When the event time $\tau_k^{(N)}$ is distinct, that is $\tau_k^{(N)} \neq \tau_{k'}^{(N)}$ for all $k \neq k'$, the next event time will be the minimum among the first arrival time after τ_k and the next departure time from the queues which are busy at time τ_k. Therefore, defining $\tilde{\Omega}_k$ to be the set of realizations on which event time $\tau_k^{(N)}$ is distinct:

$$\tilde{\Omega}_k \doteq \{\omega \in \Omega; \tau_k \neq \tau_{k'}, \quad \forall k \neq k'\}, \tag{5.3.21}$$

we have

$$\tau_{k+1}^{(N)} = \min\left(\xi_{k+1}^{(N)}, \sigma_{k+1}^{(N),i}; i = 1, ..., N\right) \qquad \text{on } \tilde{\Omega}_k. \tag{5.3.22}$$

The following lemma provides the joint conditional distribution of next arrival and departure times given $\mathcal{F}^{(N)}_{\tau^{(N)}_k}$.

LEMMA 5.15. Suppose Assumptions I, II.a, and III.a hold. Then for every $k \geq 0$, $\mathbb{P}\{\tilde{\Omega}_k\} = 1$. Also, $\xi^{(N)}_{k+1}$ and $\sigma^{(N),i}_{k+1}$, $i = 1, ..., N$, are conditionally independent given $\mathcal{F}^{(N)}_{\tau^{(N)}_k}$, and have the following conditional distributions: for every $b > 0$,

$$\mathbb{P}\left\{\xi^{(N)}_{k+1} - \tau^{(N)}_k > b \middle| \mathcal{F}^{(N)}_{\tau^{(N)}_k}\right\} = \frac{\overline{G}^{(N)}_E(R^{(N)}(\tau^{(N)}_k) + b)}{\overline{G}^{(N)}_E(R^{(N)}(\tau^{(N)}_k))}, \tag{5.3.23}$$

and for $i = 1, ..., N$,

$$\mathbb{1}\left(i \in \mathfrak{B}^{(N)}_k\right) \mathbb{P}\left\{\sigma^{(N),i}_{k+1} - \tau^{(N)}_k > b \middle| \mathcal{F}^{(N)}_{\tau^{(N)}_k}\right\} = \mathbb{1}\left(i \in \mathfrak{B}^{(N)}_k\right) \frac{\overline{G}(a^{(N),i}(\tau^{(N)}_k) + b)}{\overline{G}(a^{(N),i}(\tau^{(N)}_k))}. \tag{5.3.24}$$

Lemma 5.15 basically follows from independence of inter-arrival and service times and is proved in Appendix 5.A.3. Note that using Lemma 5.15, we can rewrite equation (5.3.22) as

$$\tau^{(N)}_{k+1} = \min\left(\xi^{(N)}_{k+1}, \sigma^{(N),i}_{k+1}; i = 1, ..., N\right), \qquad \text{a.s.} \tag{5.3.25}$$

The following Corollary is a consequence of Lemma 5.15. Recall that a sequence $\{t_n; n \in \mathbb{N}\}$ is called non-explosive if for every $T < \infty$, there are finitely many n with $t_n \leq T$.

COROLLARY 5.16. Suppose Assumptions I, II.a, and III.a hold. Then, almost surely, the sequence of event times $\{\tau^{(N)}_k; k \geq 0\}$ is non-explosive, and for every $t \geq 0$, Ω_t defined in (5.2.1) satisfies $\mathbb{P}\{\Omega_t\} = 1$.

PROOF. Fixe $t \geq 0$, define $\hat{\Omega}_t = \{\omega \; : \; X^{(N)}(0) < \infty, E^{(N)}(t) < \infty\}$. Since $X^{(N)}(0) < \infty$ almost surely by Assumption III.a and $E^{(N)}$ is a renewal process with non-degenerate inter-arrival time distribution by Assumption I, $\mathbb{P}\{\hat{\Omega}_t\} = 1$. Moreover, $D^{(N)}(t)$ is also finite on $\hat{\Omega}_t$ because of the bound (5.1.22). Thus, the total number of events up to t is finite and $\{\tau^{(N)}_k\}$ is non-explosive on $\hat{\Omega}_t$, and hence, almost surely.

Moreover, the set $\tilde{\Omega} \doteq \cup_{k\geq 0}\tilde{\Omega}_k$ of realizations on which all events are distinct has full measure by Lemma 5.15. For every $\omega \in \hat{\Omega}_t \cap \tilde{\Omega}$, the infimum $\Delta(\omega)$ of the inter-event times before time t,

$$\Delta(\omega) \doteq \inf_{k\in\{0,1,\dots,I(t)\}} \left(\tau_{k+1}^{(N)} - \tau_k^{(N)}\right),$$

is positive, because it is an infimum over finitely many positive numbers. This means that for $n > 1/\Delta$, all events before time t are more than $1/n$ away from each other. Therefore, $\omega \in \Omega_{\frac{k}{n},\frac{1}{n}}$ for all $k = 0,\dots,\lfloor nt\rfloor$, and hence $\omega \in \Omega_t$. This implies $\Omega_t \subset \hat{\Omega}_t \cap \tilde{\Omega}$ and hence $\mathbb{P}\{\Omega_t\} = 1$. $\square$

5.3.3. Compensator for the Weighted Routing Measure. We start by presenting an intensity for the process $E^{(N),i}$, defined in (5.1.16).

LEMMA 5.17. Suppose Assumptions I, II.a, and III.a hold. Then, for $i = 1,\dots,N$, the process

(5.3.26)
$$\left\{\frac{1}{N}h_E^{(N)}\left(R^{(N)}(t-)\right)\sum_{\ell=1}^{\infty}\mathbb{1}\left(X^{(N),i}(t-)=\ell-1\right)\mathfrak{P}_d\left(\overline{S}_{\ell-1}^{(N)}(t-),\overline{S}_{\ell}^{(N)}(t-)\right);t\geq 0\right\}.$$

is an $\{\mathcal{F}_t^{(N)}\}$-intensity of $E^{(N),i}$.

PROOF. Throughout this proof, we fix N and omit the superscript (N) for the ease of notation. Suppose that for every $k \geq 0$, the conditional density $f_{k+1}^{\mathfrak{E},i}$ defined by

(5.3.27)
$$\mathbb{P}\left\{\tau_{k+1}-\tau_k\in A, z_{k+1}=(\mathfrak{E},i)\middle|\mathcal{F}_{\tau_k}\right\} = \int_A f_{k+1}^{\mathfrak{E},i}(\omega,r)dr, \qquad \omega\in\Omega,\ A\in\mathcal{B}[0,\infty),$$

exist. Then, by representation (5.3.15) of E^i in terms of the marked point process $\mathcal{T}^{(N)}$ and since the sequence $\{\tau_k\}$ is non-explosive by Corollary 5.16, it follows from Theorem III.T7 in **[14]**, comment (β) below the theorem, and equation (2.10) of the following remark that the process

$$\sum_{k=0}^{\infty}\frac{f_{k+1}^{\mathfrak{E}}(t-\tau_k;i)}{\mathbb{P}\left\{\tau_{k+1}>t|\mathcal{F}_{\tau_k}\right\}}\mathbb{1}\left(\tau_k<t\leq\tau_{k+1}\right). \tag{5.3.28}$$

is an $\{\mathcal{F}_t\}$-intensity of E^i.

Now we compute the conditional probability in (5.3.27) and show that its corresponding conditional density exists. First, note that using (5.3.25), the next event after τ_k is an arrival to queue i (i.e. $z_{k+1} = (\mathfrak{E}, i)$) if the next arrival occurs before the next departure from each queue (i.e. $\xi_{k+1} \leq \sigma^{i'}_{k+1}, i' = 1, ..., N$), and the arriving job, which has index $E(\tau_{k+1})$, is routed to queue i. Therefore, defining $\sigma_{k+1} \doteq \min(\sigma^{i'}_{k+1}; i' = 1, ..., N)$, by the tower property of conditional expectation we have

(5.3.29)

$$\begin{aligned}
&\mathbb{P}\left\{\tau_{k+1} - \tau_k > t, z_{k+1} = (\mathfrak{E}, i) \middle| \mathcal{F}_{\tau_k}\right\} \\
&\qquad = \mathbb{P}\left\{\xi_{k+1} > \tau_k + t, \sigma_{k+1} > \xi_{k+1}, \kappa_{E(\tau_{k+1})} = i \middle| \mathcal{F}_{\tau_k}\right\} \\
&\qquad = \mathbb{E}\left[\mathbb{1}\left(\xi_{k+1} > \tau_k + t, \sigma_{k+1} > \xi_{k+1}\right) \mathbb{P}\left\{\kappa_{E(\tau_{k+1})} = i | \mathcal{F}_{\tau_k}, \xi_{k+1}, \sigma_{k+1}\right\} \middle| \mathcal{F}_{\tau_k}\right].
\end{aligned}$$

The queue to which the job $j = E(\tau_{k+1})$ is routed by the $SQ(d)$ algorithm is given by (5.1.6), which is a (deterministic) function of random queue choices vector ι_j and queue lengths at time τ_k. According to (5.1.6), this job is routed to a server of queue length exactly $\ell - 1$, if and only if all selected queue indices ι_j have lengths at least $\ell - 1$, and at least one of them has length exactly $\ell - 1$. Therefore, since ι_j is independent of all other random variables, the conditional probability (given $\mathcal{F}_{\tau_k}, \sigma_{k+1}$ and ξ_{k+1}) that the job is routed to a queue of length $\ell - 1$ is $(\overline{S}_{\ell-1}(\tau_k))^d - (\overline{S}_\ell(\tau_k))^d$. Moreover, the number of queues with length $\ell - 1$ is $S_{\ell-1}(\tau_k) - S_\ell(\tau_k)$, and the job j is equally likely to be routed to any of such queues. Hence,

$$\begin{aligned}
&\mathbb{1}\left(X^i(\tau_k) = \ell - 1\right) \mathbb{P}\left\{\kappa_{E(\tau_{k+1})} = i | \mathcal{F}_{\tau_k}, \xi_{k+1}, \sigma_{k+1}\right\} \\
&\qquad = \frac{1}{S_{\ell-1}(\tau_k) - S_\ell(\tau_k)} \left(\overline{S}_{\ell-1}(\tau_k)\right)^d - \left(\overline{S}_\ell(\tau_k)\right)^d \\
&\qquad = \frac{1}{N} \mathfrak{P}_d \left(\overline{S}_{\ell-1}(\tau_k), \overline{S}_\ell(\tau_k)\right)
\end{aligned}$$

where the polynomial $\mathfrak{P}_d$ is defined in (6.1.1). Therefore,

$$\begin{aligned}(5.3.30)\qquad \mathbb{P}\left\{\kappa_{E(\tau_{k+1})}=i|\mathcal{F}_{\tau_k},\xi_{k+1},\sigma_{k+1}\right\}\\ &=\sum_{\ell=1}^{\infty}\mathbb{1}\left(X^i(\tau_k)=\ell-1\right)\mathbb{P}\left\{\kappa_{E(\tau_{k+1})}=i|\mathcal{F}_{\tau_k},\xi_{k+1},\sigma_{k+1}\right\}\\ &=\frac{1}{N}\sum_{\ell=1}^{\infty}\mathbb{1}\left(X^i(\tau_k)=\ell-1\right)\mathfrak{P}_d\left(\overline{S}_{\ell-1}(\tau_k),\overline{S}_{\ell}(\tau_k)\right).\end{aligned}$$

Moreover, using equation (5.3.23) and the conditional independence, given $\mathcal{F}_{\tau_k}$, of ξ_{k+1} and $\sigma^i_{k+1}, i=1,...,N$ established in Lemma 5.15, we have

$$\begin{aligned}(5.3.31)\qquad \mathbb{P}\left\{\xi_{k+1}>\tau_k+t,\sigma_{k+1}>\xi_{k+1}|\mathcal{F}_{\tau_k}\right\}\\ &=\frac{1}{\overline{G}_E(R_E(\tau_k))}\int_t^{\infty}\mathbb{P}\left\{\sigma_{k+1}-\tau_k>s|\mathcal{F}_{\tau_k}\right\}g_E(s+R_E(\tau_k))ds.\end{aligned}$$

Therefore, by equations (5.3.29)-(5.3.31) and the fact that the right-hand side of (5.3.30) is a function of $X^i(\tau_k)$ and $S_\ell(\tau_k), \ell\geq 1$, and hence is $\mathcal{F}_{\tau_k}$-measurable (see Proposition 5.21), for $k\in\mathbb{Z}_+$, $i\in\{1,...,N\}$, the density $f^{\mathfrak{E},i}_{k+1}$ exists, and for $t\geq\tau_k$,

$$\begin{aligned}(5.3.32)\qquad f^{\mathfrak{E},i}_{k+1}(t-\tau_k)&=\mathbb{P}\left\{\sigma_{k+1}>t|\mathcal{F}_{\tau_k}\right\}\frac{g_E(t-\tau_k+R_E(\tau_k))}{N\,\overline{G}_E(R_E(\tau_k))}\\ &\quad\times\sum_{\ell=1}^{\infty}\mathbb{1}\left(X^i(\tau_k)=\ell-1\right)\mathfrak{P}_d\left(\overline{S}_{\ell-1}(\tau_k),\overline{S}_{\ell}(\tau_k)\right)\end{aligned}$$

Similarly, by (5.3.25), the conditional independence established in Lemma 5.15, and (5.3.23), for every $k\geq 0$ and $t\geq\tau_k$ we have

$$\begin{aligned}\mathbb{P}\left\{\tau_{k+1}>t|\mathcal{F}_{\tau_k}\right\}&=\mathbb{P}\left\{\xi_{k+1}\wedge\sigma_{k+1}>t|\mathcal{F}_{\tau_k}\right\}\\ &=\mathbb{P}\left\{\xi_{k+1}>t|\mathcal{F}_{\tau_k}\right\}\mathbb{P}\left\{\sigma_{k+1}>t|\mathcal{F}_{\tau_k}\right\}\\ &=\mathbb{P}\left\{\sigma_{k+1}>t|\mathcal{F}_{\tau_k}\right\}\frac{\overline{G}_E(t-\tau_k+R_E(\tau_k))}{\overline{G}_E(R_E(\tau_k))}.\qquad(5.3.33)\end{aligned}$$

Therefore, by substituting (5.3.32) and (5.3.33) in the representation (5.3.28) of $\{\mathcal{F}_t\}$-intensity of E^i, recalling the definition $h_E=g_E/\overline{G}_E$ of the hazard rate function of the inter-arrival times, and using the facts that for $t\in(\tau_k,\tau_{k+1}]$,

$X^i(t-) = X^i(\tau_k)$, $S_\ell(t-) = S_\ell(\tau_k)$, and $R_E(t-) = t - \tau_k + R_E(\tau_k)$, the process

(5.3.34)
$$\frac{1}{N}\sum_{k=0}^{\infty} h_E(t-\tau_k+R_E(\tau_k)) \sum_{\ell=1}^{\infty} \mathbb{1}\left(X^i(\tau_k)=\ell-1\right)\mathfrak{P}_d\left(\overline{S}_{\ell-1}(\tau_k),\overline{S}_\ell(\tau_k)\right)\mathbb{1}\left(\tau_k<t\le\tau_{k+1}\right)$$
$$= \frac{1}{N}\sum_{k=0}^{\infty} h_E(R_E(t-)) \sum_{\ell=1}^{\infty} \mathbb{1}\left(X^i(t-)=\ell-1\right)\mathfrak{P}_d\left(\overline{S}_{\ell-1}(t-),\overline{S}_\ell(t-)\right)\mathbb{1}\left(\tau_k<t\le\tau_{k+1}\right)$$
$$= \frac{1}{N} h_E(R_E(t-)) \sum_{\ell=1}^{\infty} \mathbb{1}\left(X^i(t-)=\ell-1\right)\mathfrak{P}_d\left(\overline{S}_{\ell-1}(t-),\overline{S}_\ell(t-)\right), \qquad t\ge 0,$$

is an $\{\mathcal{F}_t\}$-intensity of $E^{(N),i}$. $\square$

We now use Lemma 5.17 to prove the martingale decomposition for $\mathcal{R}^{(N)}_{\varphi,\ell}$, stated in Proposition 5.11.

PROOF OF PROPOSITION 5.11. The process $\mathcal{R}^{(N)}_{\varphi,\ell}$ defined in (5.1.19), can be written as

$$\mathcal{R}^{(N)}_{\varphi,\ell} = \sum_{i=1}^{N} \mathcal{R}^{(N),i}_{\varphi,\ell}, \tag{5.3.35}$$

where

$$\mathcal{R}^{(N),i}_{\varphi,\ell}(t) \doteq \begin{cases} \int_{(0,t]} \varphi(0,s)\mathbb{1}\left(X^{(N),i}(s-)=0\right) dE^{(N),i}(s) & \text{if } \ell=1, \\ \int_{(0,t]} \langle \varphi(\cdot,s), \nu^{(N),i}_{\ell-1}(s-) - \nu^{(N),i}_{\ell}(s-)\rangle dE^{(N),i}(s) & \text{if } \ell\ge 2. \end{cases} \tag{5.3.36}$$

Consider the setup of Lemma 5.14 with ξ replaced by $E^{(N),i}$, the filtration $\{\mathcal{G}_t\}$ replaced by $\{\mathcal{F}^{(N)}_t\}$, $\theta(s)$ replaced by $\varphi(0,s)\mathbb{1}(X^{(N),i}(s-)=0)$, and hence ζ replaced by $\mathcal{R}^{(N),i}_{\varphi,1}$. Since φ is bounded and $X^{(N),i}$ is right-continuous and $\{\mathcal{F}^{(N)}_t\}$-adapted (see proposition 5.21), θ is bounded, $\{\mathcal{F}^{(N)}_t\}$-adapted , and left-continuous, and thus, $\{\mathcal{F}^{(N)}_t\}$-predictable. Therefore, using the fact that (5.3.26) is an intensity of $E^{(N),i}$, and by the identity

(5.3.37)
$$\frac{1}{N}\varphi(0,t)\mathbb{1}\left(X^{(N),i}(t-)=0\right) h^{(N)}_E(R^{(N)}(t-))$$
$$\times \sum_{\ell'=1}^{\infty} \mathbb{1}\left(X^{(N),i}(t-)=\ell'-1\right)\mathfrak{P}_d\left(\overline{S}^{(N)}_{\ell'-1}(t-),\overline{S}^{(N)}_{\ell'}(t-)\right)$$
$$= \frac{1}{N}\varphi(0,t)\mathbb{1}\left(X^{(N),i}(t-)=0\right) h^{(N)}_E(R^{(N)}(t-))\mathfrak{P}_d\left(1,\overline{S}^{(N)}_1(t-)\right),$$

it follows from Lemma 5.14 that the process $\{\mathcal{N}^{(N),i}_{\varphi,1}(t); t \geq 0\}$ defined as

$$(5.3.38)\quad \mathcal{N}^{(N),i}_{\varphi,1}(t) \doteq \mathcal{R}^{(N),i}_{\varphi,1}(t) - \frac{1}{N}\int_0^t \varphi(0,s)\mathbb{1}\left(X^{(N),i}(s)=0\right) \times h^{(N)}_E(R^{(N)}(s))\mathfrak{P}_d\left(1,\overline{S}^{(N)}_1(s)\right) ds,$$

is a local $\{\mathcal{F}^{(N)}_t\}$-martingale (note that the right-hand side of (5.3.37) and the integrand of the display above only differ on a zero-measure subset of $[0,t]$). Moreover, since the number of idle servers at time t is $N - S^{(N)}_1(t)$, and by definition (6.1.1) of $\mathfrak{B}$ and definition (5.3.1) of $B^{(N)}_{\varphi,1}$, we have

(5.3.39)

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N \int_0^t \varphi(0,s)\mathbb{1}\left(X^{(N),i}(s)=0\right) h^{(N)}_E(R^{(N)}(s))\mathfrak{P}_d\left(1,\overline{S}^{(N)}_1(s)\right) ds \\
&\quad = \frac{1}{N}\int_0^t \varphi(0,s)h^{(N)}_E(R^{(N)}(s))\mathfrak{P}_d\left(1,\overline{S}^{(N)}_1(s)\right)\sum_{i=1}^N \mathbb{1}\left(X^{(N),i}(s)=0\right) ds \\
&\quad = \int_0^t \varphi(0,s)h^{(N)}_E(R^{(N)}(s))\left(1-\left(\overline{S}^{(N)}_1(s)\right)^d\right) ds \\
&\quad = B^{(N)}_{\varphi,1}(t).
\end{aligned}$$

Therefore, summing up both sides of (5.3.38) over $i = 1, ..., N$ and using (5.3.35) and (5.3.39), we have

$$\sum_{i=1}^N \mathcal{N}^{(N),i}_{\varphi,1} = \mathcal{R}^{(N)}_{\varphi,1} - B^{(N)}_{\varphi,1} = \mathcal{N}^{(N)}_{\varphi,1},$$

and hence, $\mathcal{N}^{(N)}_{\varphi,1}$ is a local $\{\mathcal{F}^{(N)}_t\}$-martingale. Moreover, by equation (5.3.12) of Lemma 5.14 in the setup described above,

$$[\mathcal{R}^{(N),i}_{\varphi,1}](t) = \int_0^t \varphi^2(0,s)\mathbb{1}\left(X^{(N),i}(u-)=0\right) dE^{(N),i}(u) = \mathcal{R}^{(N),i}_{\varphi^2,1}(t).$$

Also, for $i = 1, ..., N$ and $i' \neq i$, $\mathcal{R}^{(N),i}_{\varphi,1}$ and $\mathcal{R}^{(N),i'}_{\varphi,1}$ are pure jump processes with no common jump times almost surely (see Lemma 5.15), and hence, $[\mathcal{R}^{(N),i}_{\varphi,\ell}]$ is locally bounded and $[\mathcal{R}^{(N),i}_{\varphi,1}, \mathcal{R}^{(N),i'}_{\varphi,1}] \equiv 0$, almost surely. Together with the fact that

$B^{(N)}_{\varphi,1} = \mathcal{R}^{(N)}_{\varphi,1} - \mathcal{N}^{(N)}_{\varphi,1}$ is a continuous function with finite variation, this implies

$$[\mathcal{N}^{(N)}_{\varphi,1}] = [\mathcal{R}^{(N)}_{\varphi,1}] = [\sum_{i=1}^{N} \mathcal{R}^{(N),i}_{\varphi,1}] = \sum_{i=1}^{N} [\mathcal{R}^{(N),i}_{\varphi,1}] = \sum_{i=1}^{N} \mathcal{R}^{(N),i}_{\varphi^2,\ell} = \mathcal{R}^{(N)}_{\varphi^2,1}.$$

Similarly, for $\ell \geq 2$, consider the setup of Lemma 5.14 with the same ξ and $\{\mathcal{G}_t\}$ as above, but with $\theta(s)$ replaced by

$$\langle \varphi(\cdot, s), \nu^{(N),i}_{\ell-1}(s-) - \nu^{(N),i}_{\ell}(s-) \rangle = \varphi(a^{(N),i}(s-), s) \mathbb{1}\left(X^{(N),i}(s-) = \ell - 1\right)$$

and hence, ζ replaced by $\mathcal{R}^{(N),i}_{\varphi,\ell}$. Since $a^{(N),i}$ and $X^{(N),i}$ are $\{\mathcal{F}^{(N)}_t\}$-adapted (see Proposition 5.21), θ is bounded, $\{\mathcal{F}^{(N)}_t\}$-adapted, and left-continuous. Thus, using the fact that (5.3.26) is an intensity of $E^{(N),i}$, and by the identity

$$\begin{aligned}
&\frac{1}{N}\varphi\left(a^{(N),i}(t-), t\right) \mathbb{1}\left(X^{(N),i}(t-) = \ell - 1\right) h^{(N)}_E(R^{(N)}(t-)) \\
&\qquad\qquad \times \sum_{\ell'=1}^{\infty} \mathbb{1}\left(X^{(N),i}(t-) = \ell' - 1\right) \mathfrak{P}_d\left(\overline{S}^{(N)}_{\ell'-1}(t-), \overline{S}^{(N)}_{\ell'}(t-)\right) \\
&\qquad = \frac{1}{N}\varphi\left(a^{(N),i}(t-), t\right) \mathbb{1}\left(X^{(N),i}(t-) = \ell - 1\right) h^{(N)}_E(R^{(N)}(t-)) \\
&\qquad\qquad \times \mathfrak{P}_d\left(\overline{S}^{(N)}_{\ell-1}(t-), \overline{S}^{(N)}_{\ell}(t-)\right),
\end{aligned}$$

it follows from Lemma 5.14 that the process $\{\mathcal{N}^{(N),i}_{\varphi,\ell}(t); t \geq 0\}$ defined as

$$\begin{aligned}
\mathcal{N}^{(N),i}_{\varphi,\ell}(t) \doteq \mathcal{R}^{(N),i}_{\varphi,\ell}(t) - \frac{1}{N}\int_0^t \varphi\left(a^{(N),i}(s), s\right) \mathbb{1}\left(X^{(N),i}(s) = \ell - 1\right) \\
\times h^{(N)}_E(R^{(N)}(s)) \mathfrak{P}_d\left(\overline{S}^{(N)}_{\ell-1}(s), \overline{S}^{(N)}_{\ell}(s)\right) ds,
\end{aligned}$$

is a local $\{\mathcal{F}_t^{(N)}\}$-martingale. Again, by the representation (5.3.18) of $\nu_\ell^{(N),i}$,

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\int_0^t \varphi\left(a^{(N),i}(s),s\right)\mathbb{1}\left(X^{(N),i}(s)=\ell-1\right)h_E^{(N)}(R^{(N)}(s))\mathfrak{P}_d\left(\overline{S}_{\ell-1}^{(N)}(s),\overline{S}_\ell^{(N)}(s)\right)ds\\
&\quad=\frac{1}{N}\int_0^t h_E^{(N)}(R^{(N)}(s))\mathfrak{P}_d\left(\overline{S}_{\ell-1}^{(N)}(s),\overline{S}_\ell^{(N)}(s)\right)\sum_{i=1}^{N}\varphi(a^{(N),i}(s),s)\mathbb{1}\left(X^{(N),i}(s)=\ell-1\right)ds\\
&\quad=\frac{1}{N}\int_0^t h_E^{(N)}(R^{(N)}(s))\mathfrak{P}_d\left(\overline{S}_{\ell-1}^{(N)}(s),\overline{S}_\ell^{(N)}(s)\right)\sum_{i=1}^{N}\langle\varphi(\cdot,s),\nu_{\ell-1}^{(N),i}(s)-\nu_\ell^{(N),i}(s)\rangle ds\\
&\quad=\int_0^t h_E^{(N)}(R^{(N)}(s))\mathfrak{P}_d\left(\overline{S}_{\ell-1}^{(N)}(s),\overline{S}_\ell^{(N)}(s)\right)\langle\varphi(\cdot,s),\overline{\nu}_{\ell-1}^{(N)}(s)-\overline{\nu}_\ell^{(N)}(s)\rangle ds\\
&\quad=B_{\varphi,\ell}^{(N)}(t),
\end{aligned}
$$

and using (5.3.35), it follows that

$$
\sum_{i=1}^{N}\mathcal{N}_{\varphi,\ell}^{(N),i}=\mathcal{R}_{\varphi,\ell}^{(N)}-B_{\varphi,\ell}^{(N)}=\mathcal{N}_{\varphi,\ell}^{(N)},
$$

and hence, $\mathcal{N}_{\varphi,\ell}^{(N)}$ is also a local $\{\mathcal{F}_t^{(N)}\}$-martingale. Also, by equation (5.3.12) of Lemma 5.14 in the setup above, and since

$$
\theta^2(s)=\varphi^2(a^{(N),i}(s-),s)\mathbb{1}\left(X^{(N),i}(s-)=\ell-1\right)=\langle\varphi^2(\cdot,s),\nu_{\ell-1}^{(N),i}(s-)-\nu_\ell^{(N),i}(s-)\rangle
$$

we have,

$$
[\mathcal{R}_{\varphi,\ell}^{(N),i}](t)=\int_0^t\langle\varphi^2(\cdot,s),\nu_{\ell-1}^{(N),i}(s-)-\nu_\ell^{(N),i}(s-)\rangle dE^{(N),i}(s)=\mathcal{R}_{\varphi^2,\ell}^{(N),i}.
$$

Finally, with the same reasoning as the case $\ell=1$, for $\ell\geq 2$ we have

$$
[\mathcal{N}_{\varphi,\ell}^{(N)}]=[\mathcal{R}_{\varphi,\ell}^{(N)}]=\mathcal{R}_{\varphi^2,\ell}^{(N)}.
$$

This completes the proof. $\square$

5.3.4. Compensator of the weighted departure process. In this section, we prove Proposition 5.12. We start by computing the intensity of the process $D^{(N),i}$, defined in (5.1.25).

LEMMA 5.18. Suppose Assumptions I, II.a, and III.a hold. Then, for $i = 1, ..., N$, the process

$$\left\{ \mathbb{1}\left(X^{(N),i}(t-) \geq 1 \right) h\left(a^{(N),i}(t-) \right); t \geq 0 \right\}. \tag{5.3.40}$$

is an $\{\mathcal{F}_t^{(N)}\}$-intensity of $D^{(N),i}$.

PROOF. In this proof, we fix N and suppress the superscript (N) for ease of notation. Recall that although $a^i(t)$ is not defined if server i is idle at time t, the quantity $\mathbb{1}(X^i(t) \geq 1)h(a^i(t))$ is always well defined.

Suppose that for every $k \geq 0$, the conditional density $f_{k+1}^{\mathfrak{D},i}$ defined by

(5.3.41)

$$\mathbb{P}\left\{ \tau_{k+1} - \tau_k \in A, z_{k+1} = (\mathfrak{D}, i) \middle| \mathcal{F}_{\tau_k} \right\} = \int_A f_{k+1}^{\mathfrak{D},i}(\omega, r)dr, \qquad \omega \in \Omega,\ A \in \mathcal{B}[0, \infty),$$

exist. Then, by representation (5.3.16) of D^i in terms of the marked point process $\mathcal{T}^{(N)}$ and since the sequence $\{\tau_k\}$ is non-explosive by Corollary 5.16, it follows from Theorem III.T7 in **[14]**, comment (β) below the theorem, and equation (2.10) of the following remark that the process

$$\sum_{k=0}^{\infty} \frac{f_{k+1}^{\mathfrak{D},i}(t - \tau_k)}{\mathbb{P}\left\{ \tau_{k+1} > t | \mathcal{F}_{\tau_k} \right\}} \mathbb{1}\left(\tau_k < t \leq \tau_{k+1} \right). \tag{5.3.42}$$

is an $\{\mathcal{F}_t\}$-intensity of D^i.

Now we compute the conditional probability in (5.3.41) and show that its corresponding conditional density exists. Recall that ξ_{k+1} and σ_{k+1}^i are, respectively, the first arrival time and the first departure time from queue i after the even time τ_k. Using (5.3.25), the next event after τ_k is a departure from queue i (i.e. $z_{k+1} = (\mathfrak{D}, i)$) if the next departure from queue i occurs before the next arrival and all other departures from other queues (i.e. $\sigma_{k+1}^i \leq \xi_{k+1} \wedge \sigma_{k+1}^{i'}, i' = 1, ..., N, i' \neq i$). If the server i is idle at time τ_k (i.e., $i \notin \mathfrak{B}(\tau_k)$), $\sigma_{k+1}^i = \infty$ by definition, and the probability that the next event is a departure from queue i is zero, and therefore,

$$f_{k+1}^{\mathfrak{D},i} \equiv 0, \qquad \forall i \notin \mathfrak{B}(\tau_k). \tag{5.3.43}$$

However, for $i \in \mathfrak{B}(\tau_k)$, using equation (5.3.24) and the conditional independence, given $\mathcal{F}_{\tau_k}$, of ξ_{k+1} and $\sigma^i_{k+1}, i = 1, ..., N$, established in Lemma 5.15, we have

(5.3.44)

$$\begin{aligned}
&\mathbb{1}\left(i \in \mathfrak{B}(\tau_k)\right) \mathbb{P}\left\{\tau_{k+1} - \tau_k > t, z_{k+1} = (\mathfrak{D}, i) | \mathcal{F}_{\tau_k}\right\} \\
&\qquad = \mathbb{1}\left(i \in \mathfrak{B}(\tau_k)\right) \mathbb{P}\left\{\sigma^i_{k+1} - \tau_k > t, \xi_{k+1} \wedge \min_{i' \neq i} \sigma^{i'}_{k+1} > \sigma^i_{k+1} | \mathcal{F}_{\tau_k}\right\} \\
&\qquad = \mathbb{1}\left(i \in \mathfrak{B}(\tau_k)\right) \frac{1}{\overline{G}(a^i(\tau_k))} \int_t^\infty \mathbb{P}\left\{\xi_{k+1} \wedge \min_{i' \neq i} \sigma^{i'}_{k+1} > \tau_k + s | \mathcal{F}_{\tau_k}\right\} \\
&\qquad\qquad \times g(s + a^i(\tau_k)) ds.
\end{aligned}$$

Thus, the conditional density $f^{\mathfrak{D},i}_{k+1}$ exists and for every $t \geq \tau_k$,

(5.3.45)

$$f^{\mathfrak{D},i}_{k+1}(t - \tau_k) = \mathbb{1}\left(i \in \mathfrak{B}(\tau_k)\right) \frac{1}{\overline{G}(a^i(\tau_k))} \mathbb{P}\left\{\xi_{k+1} \wedge \min_{i' \neq i} \sigma^{i'}_{k+1} > t | \mathcal{F}_{\tau_k}\right\} g(t - \tau_k + a^i(\tau_k)).$$

Moreover, by another use of (5.3.25), (5.3.24) and the conditional independence established in Lemma 5.15, for $i \in \mathfrak{B}_k$ we have

$$\begin{aligned}
\mathbb{P}\left\{\tau_{k+1} > t | \mathcal{F}_{\tau_k}\right\} &= \mathbb{P}\left\{\xi_{k+1} \wedge \sigma^i_{k+1} \wedge \min_{i' \neq i} \sigma^{i'}_{k+1} > t | \mathcal{F}_{\tau_k}\right\} \\
&= \mathbb{P}\left\{\xi_{k+1} \wedge \min_{i' \neq i} \sigma^{i'}_{k+1} > t | \mathcal{F}_{\tau_k}\right\} \frac{\overline{G}(t - \tau_k + a^i(\tau_k)).}{\overline{G}(a^i(\tau_k))}. \qquad (5.3.46)
\end{aligned}$$

Substituting (5.3.43), (5.3.45) and (5.3.46), and definition (5.3.20) of $\mathfrak{B}_k$, the $\{\mathcal{F}_t\}$-intensity (5.3.42) of D^i is equal to

$$\begin{aligned}
&\sum_{k=0}^\infty \mathbb{1}\left(i \in \mathfrak{B}_k\right) h(t - \tau_k + a^i(\tau_k)) \mathbb{1}\left(\tau_k < t \leq \tau_{k+1}\right) \\
&\qquad = \sum_{k=0}^\infty \mathbb{1}\left(X^i(\tau_k) \geq 1\right) h(t - \tau_k + a^i(\tau_k)) \mathbb{1}\left(\tau_k < t \leq \tau_{k+1}\right).
\end{aligned}$$

Using the fact that for $t \in (\tau_k, \tau_{k+1}]$ and $i \in \mathfrak{B}_k$, $X^i(t-) = X^i(\tau_k)$ and $a^i(t-) = t - \tau_k + a^i(\tau_k)$, the display above can be simplified as

$$\sum_{k=0}^{\infty} \mathbb{1}\left(X^i(\tau_k) \geq 1\right) h(t - \tau_k + a^i(\tau_k))\mathbb{1}\left(\tau_k < t \leq \tau_{k+1}\right)$$
$$= \sum_{k=0}^{\infty} \mathbb{1}\left(X^i(t-) \geq 1\right) h\left(a^i(t-)\right) \mathbb{1}\left(\tau_k < t \leq \tau_{k+1}\right)$$
$$= \mathbb{1}\left(X^i(t-) \geq 1\right) h\left(a^i(t-)\right),$$

which completes the proof. □

PROOF OF PROPOSITION 5.12. Fix $N \in \mathbb{N}$, $\varphi \in \mathbb{C}_b([0, L) \times \mathbb{R}_+)$ and $\ell \geq 1$, and recall that $a_j^{(N)}(\cdot)$, v_j and $\gamma_j^{(N)}$ are the age process, service time and the departure time of job j, and that $a_j^{(N)}(\gamma_j^{(N)}) = v_j$. Since the server $\kappa_j^{(N)}$, at which the job j is receiving service is busy right before the departure time of j, the quantity $a^{(N),\kappa_j^{(N)}}(\gamma_j^{(N)}-)$ is well defined and satisfies

$$a^{(N),\kappa_j^{(N)}}(\gamma_j^{(N)}-) = a_j^{(N)}(\gamma_j^{(N)}) = v_j.$$

Therefore, using the identity above and definition (5.1.20) of $\chi_j^{D,(N)}$, for $t \geq 0$ the quantity $\mathcal{D}_{\varphi,\ell}^{(N)}(t)$ defined in (5.1.23) can be written as

$$\mathcal{D}_{\varphi,\ell}^{(N)}(t) = \sum_{j=j_0}^{\infty} \varphi\Big(a^{(N),\kappa_j^{(N)}}(\gamma_j^{(N)}-), \gamma_j^{(N)}\Big)\mathbb{1}\left(X^{(N),\kappa_j^{(N)}}(\gamma_j^{(N)}-) \geq \ell\right) \mathbb{1}\left(\gamma_j^{(N)} \leq t\right)$$
$$= \sum_{i=1}^{N} \mathcal{D}_{\varphi,\ell}^{(N),i}(t), \tag{5.3.47}$$

where

$$\mathcal{D}_{\varphi,\ell}^{(N),i}(t) \doteq \sum_{j=j_0}^{\infty} \varphi\Big(a^{(N),i}(\gamma_j^{(N)}-), \gamma_j^{(N)}\Big)\mathbb{1}\left(X^{(N),i}(\gamma_j^{(N)}-) \geq \ell\right)$$
$$\times \mathbb{1}\left(\kappa_j^{(N)} = i\right) \mathbb{1}\left(\gamma_j^{(N)} \leq t\right).$$

By definition (5.1.25) of $D^{(N),i}$, the process $\mathcal{D}_{\varphi,\ell}^{(N),i}$ can be represented as the following integral with respect to the departure process $D^{(N),i}$:

$$\mathcal{D}_{\varphi,\ell}^{(N),i}(t) = \int_0^t \varphi\left(a^{(N),i}(s-), s\right) \mathbb{1}\left(X^{(N),i}(s-) \geq \ell\right) dD^{(N),i}(s), \qquad t \geq 0. \tag{5.3.48}$$

Consider the setup of Lemma 5.14 with ξ replaced by $D^{(N),i}$, the filtration $\{\mathcal{G}_t\}$ replaced by $\{\mathcal{F}_t^{(N)}\}$, $\theta(s)$ replaced by $\varphi(a^{(N),i}(s-),s)\mathbb{1}(X^{(N),i}(s-))$, and hence ζ replaced by $\mathcal{D}_{\varphi,\ell}^{(N),i}$. Since $a^{(N),i}$ and $X^{(N),i}$ are right-continuous and $\{\mathcal{F}_t^{(N)}\}$-adapted (see Proposition 5.21), θ is bounded, $\{\mathcal{F}_t^{(N)}\}$-adapted, and left-continuous. Hence, using the fact that (5.3.40) is an intensity of $D^{(N),i}$, it follows from Lemma 5.14 that the process $\{\mathcal{M}_{\varphi,\ell}^{(N),i}(t); t \geq 0\}$ defined by

$$\mathcal{M}_{\varphi,\ell}^{(N),i}(t) \doteq \mathcal{D}_{\varphi,\ell}^{(N),i}(t) - \int_0^t \varphi\left(a^{(N),i}(s-),s\right) h(a^{(N),i}(s-))\mathbb{1}\left(X^{(N),i}(s-) \geq \ell\right) ds,$$

$$= \mathcal{D}_{\varphi,\ell}^{(N),i}(t) - \int_0^t \varphi\left(a^{(N),i}(s),s\right) h(a^{(N),i}(s))\mathbb{1}\left(X^{(N),i}(s) \geq \ell\right) ds, \tag{5.3.49}$$

is a local $\{\mathcal{F}_t^{(N)}\}$-martingale. Moreover, note that by the representation (5.3.19) of $\nu^{(N)}$ and definition (5.3.5) of $A_{\varphi,\ell}^{(N)}$, we have

$$\sum_{i=1}^N \int_0^t \varphi\left(a^{(N),i}(s),s\right) h(a^{(N),i}(s))\mathbb{1}\left(X^{(N),i}(s) \geq \ell\right) ds = \int_0^t \langle \varphi(\cdot,s)h(\cdot), \nu_\ell^{(N)}(s)\rangle ds$$

$$= A_{\varphi,\ell}^{(N)}(t). \tag{5.3.50}$$

Summing up both sides of (5.3.49) over $i = 1, ..., N$, and using (5.3.47) and (5.3.50), we have

$$\sum_{i=1}^N \mathcal{M}_{\varphi,\ell}^{(N),i} = \mathcal{D}_{\varphi,\ell}^{(N)} - A_{\varphi,\ell}^{(N)} = \mathcal{M}_{\varphi,\ell}^{(N)},$$

and hence $\mathcal{M}_{\varphi,\ell}^{(N)}$ is also a local $\{\mathcal{F}_t^{(N)}\}$-martingale.

Moreover, by equation (5.3.12) of Lemma 5.14 with the setup described above, for all $t \geq 0$,

$$[\mathcal{D}_{\varphi,\ell}^{(N),i}](t) = \int_0^t \varphi^2\left(a^{(N),i}(u-),u\right)\mathbb{1}\left(X^{(N),i}(u-) \geq \ell\right) dD^{(N),i}(u) = \mathcal{D}_{\varphi^2,\ell}^{(N),i}.$$

Furthermore, for every $i = 1, ..., N$ and $i' \neq i$, $\mathcal{D}_{\varphi,\ell}^{(N),i}$ and $\mathcal{D}_{\varphi,\ell}^{(N),i'}$ are pure jump processes with no common jump times almost surely (see Lemma 5.15), and hence $[\mathcal{D}_{\varphi,\ell}^{(N),i}, \mathcal{D}_{\varphi,\ell}^{(N),i'}] \equiv 0$, almost surely. Together with the fact that $A_{\varphi,\ell}^{(N)} = \mathcal{D}_{\varphi,\ell}^{(N)} - \mathcal{M}_{\varphi,\ell}^{(N)}$ is a continuous function with finite variation, this implies

$$[\mathcal{M}_{\varphi,\ell}^{(N)}] = [\mathcal{D}_{\varphi,\ell}^{(N)}] = [\sum_{i=1}^N \mathcal{D}_{\varphi,\ell}^{(N),i}] = \sum_{i=1}^N [\mathcal{D}_{\varphi,\ell}^{(N),i}] = \sum_{i=1}^N \mathcal{D}_{\varphi^2,\ell}^{(N),i} = \mathcal{D}_{\varphi^2,\ell}^{(N)},$$

which completes the proof. $\square$

5.A. On a Marked Point Process Representation

Throughout this section, we fix N and suppress the superscript (N) for conciseness. Moreover, for conciseness, we assume that the arrival process $E^{(N)}$ is a pure renewal process with inter-arrivals $u_n^{(N)}; n \geq 1$. The extension to delayed renewal process is straightforward.

5.A.1. Construction of State and Auxiliary Processes. Recall that I_0 defined in (5.1.14) is the initial state of the network, $u_j, j = 1, 2, ...$ are inter-arrival times, $\iota_j = (\iota_j(1), ..., \iota_j(d)), j = 1, 2, ...$ are queue indices randomly chosen by job j upon arrival and $v_j; j \in \mathbb{Z}$ are service times, and define *Data* to be the vector of all input data,

$$Data \doteq \left(I_0, u_j, \iota_j; j \geq 1, v_j, j \in \mathbb{Z}\right). \tag{5.A.1}$$

In Lemma 5.19 below, we show that given *Data*, we construct all state variables and auxiliary process for all times. In particular, we can construct E^i and D^i as a measurable function of *Data*, and hence the filtration $\{\mathcal{F}_t\}$ defined in (5.1.27) is a sub σ-algebra of $\mathcal{F}^{\text{Data}}$ generated by *Data*, defined as

$$\mathcal{F}^{\text{Data}} \doteq \sigma(I_0) \vee \sigma(u_j, \iota_j; j = 1, 2, ...) \vee \sigma(v_j; j \in \mathbb{Z}). \tag{5.A.2}$$

LEMMA 5.19. Suppose all inter-arrival times and service times are almost surely positive and arrival process is non-explosive. Then, we can construct α_j, β_j and γ_j for all jobs j, and the process E^i and D^i on $[0, \infty)$ for all queues $i = 1, ..., N$, as measurable functions of *Data*. In particular, $\mathcal{F}_t \subset \mathcal{F}^{\text{Data}}$ for all $t \geq 0$.

PROOF. First, define $\alpha_0 = 0$ and note that the arrival time α_j of a job $j \geq 0$ satisfies $\alpha_j = \sum_{j'=1}^{j} u_{j'}$. We define κ_j, β_j, γ_j and processes X^i, E^i and D^i on $(\alpha_j, \alpha_{j+1}]$ inductively. First, define $E^i(0) = 0$ and $D^i(0) = 0$ for $i = 1, ..., N$, and recall that $X^i(0)$ for $i = 1, ..., N$ and κ_j, β_j, γ_j for jobs $j \leq 0$ (initially in service) are measurable functions of the initial conditions I_0. Now for a $j_0 \geq 0$, assume κ_j, β_j, γ_j, $j \leq j_0$, and X^i, E^i and D^i on $[0, \alpha_{j_0}]$ are defined as a measurable function of I_0. Define

$$D^i_{j_0}(t) = \sum_{j=-X(0)+1}^{j_0} \mathbb{1}\left(\gamma_j \leq t\right) \mathbb{1}\left(\kappa_j = i\right), \qquad t \geq 0,$$

to be the departure process of jobs up to j_0 from queue i, which is a measurable function of *Data* by using induction hypothesis. Since service times are positive almost surely, departure time γ_j of any job $j \geq j_0 + 1$ satisfies $\gamma_j = \alpha_j + v_j > \alpha_{j_0+1}$, and hence, $D^i_{j_0} = D^i$ on $[0, \alpha_{j_0+1}]$. Therefore, D^i has been extended to $[0, \alpha_{j_0+1}]$ as a measurable function of *Data*. Moreover, since E and hence E^i, $i = 1, ..., N$, are piece-wise constant on $(\alpha_{j_0}, \alpha_{j_0+1})$, the length of a queue i right before the arrival of $(j_0 + 1)^{\text{th}}$ job can be obtained from the mass balance equation (5.1.26):

$$X^i(\alpha_{j_0+1}-) = X^i(0) + E^i(\alpha_{j_0+1}-) + D^i(\alpha_{j_0+1}-) = X^i(0) + E^i(\alpha_{j_0}) + D^i(\alpha_{j_0+1}-),$$

and is also a measurable function of *Data*. Therefore, the queue index κ_{j_0+1} to which the job $j_0 + 1$ is routed defined by (5.1.6) and

$$E^i(t) = \sum_{j=1}^{j_0+1} \mathbb{1}\left(\alpha_j \leq t\right) \mathbb{1}\left(\kappa_j = i\right) \qquad \text{on } [0, \alpha_{j_0+1}],$$

are both measurable functions of *Data*. The job $j_0 + 1$ will join the end of the queue (if the queue is not empty), enters service when the service requirement of all the jobs ahead of it in the same queue is completed, that is at time

$$\beta_{j_0+1} = \alpha_{j_0+1} + \sum_{j=-X(0)+1}^{j_0} \left(v_j - a_j(\alpha_{j_0+1})\right) \mathbb{1}\left(\kappa_j = \kappa_{j_0+1}\right),$$

and departs at time $\gamma_{j_0+1} = \beta_{j_0+1} + v_{j_0+1}$. Using these relations and since the age process a_j defined by (5.1.5) is a measurable function of *Data* for all $j \leq j_0$ by induction hypothesis, β_{j_0+1} and γ_{j_0+1} are also measurable functions of *Data*. This completes the induction. Finally, note that since the arrival process is non-explosive, $\lim_{j\to\infty} \alpha_j = \infty$, and hence, the construction is indeed on the whole $[0, \infty)$. $\square$

REMARK 5.20. Assumptions of Lemma 5.19 hold in particular under our Assumptions I and II.a because inter-arrival and service time distributions have densities g_E and g, respectively, and arrival process is non-explosive by Corollary 5.16.

Next, we prove that the state and auxiliary processes defined in Section 5.1 are adapted to $\{\mathcal{F}_t\}$.

PROPOSITION 5.21. Under Assumptions of Lemma 5.19, for every job j, arrival time α_j, service entry time β_j and departure time γ_j are $\{\mathcal{F}_t\}$-stopping times, and age process a_j and queue index process κ_j are $\{\mathcal{F}_t\}$-adapted. Also, the processes X^i, D_ℓ, and $\nu_\ell, \ell \geq 1$ are $\{\mathcal{F}_t\}$-adapted.

PROOF. First, note that $E(t) = \sum_{i=1}^{N} E^i(t)$ is $\mathcal{F}_t$-adapted, and since $X^i(0)$ is $\mathcal{F}_0$-measurable, the mass balance equation (5.1.26) for queue i implies that X^i is also adapted. Moreover, E^i, D^i and X^i are all right-continuous, and hence $\{\mathcal{F}_t\}$-progressive.

A job $j \geq 1$ has arrived at or before a time t if and only if at least j jobs have arrived to the system by time t, that is, $\{\alpha_j \leq t\} = \{E(t) \geq j\}$. Since $E(t)$ is $\mathcal{F}_t$-measurable, α_j is an $\{\mathcal{F}_t\}$-stopping time. Also, the queue κ_j to which job j is routed is the unique queue whose arrival process increases at α_j, that is,

$$\kappa_j = \text{argmax}_{i\in\{1,...,N\}} \left(E^i(\alpha_j) - E^i(\alpha_j-)\right)$$

Since $E^i, i = 1, ..., N$, are $\{\mathcal{F}_t\}$-progressive, each argument of the argmax above is $\mathcal{F}_{\alpha_j}$-measurable, and hence, so is κ_j. Therefore, $\kappa_j(t) = \mathbb{1}(\alpha_j \leq t)\kappa_j$ is $\{\mathcal{F}_t\}$-adapted for $j \geq 1$. For jobs initially in network ($j \leq 0$), $\kappa_j(t)$ is equal for all $t \geq 0$ to the $\mathcal{F}_0$-measurable quantity κ_j, and hence is also $\{\mathcal{F}_t\}$-adapted.

When a job $j \geq 1$ arrives at α_j and joins the queue κ_j, there are total number of $X^{\kappa_j}(\alpha_j) - 1$ jobs ahead of it in that queue. Hence, job j has entered service by the time t if it has already arrived by t and all the jobs ahead it in the same queue are already departed, that is,

$$\{\beta_j \leq t\} = \{\alpha_j \leq t\} \cap \{D^{\kappa_j}(t) - D^{\kappa_j}(\alpha_j) \geq X^{\kappa_j}(\alpha_j) - 1\}, \qquad j \geq 1.$$

Since α_j is an $\{\mathcal{F}_t\}$-stopping time, D^i and X^i are $\mathcal{F}_t$-progressive for all $i = 1, ..., N$ and κ_j is $\mathcal{F}_{\alpha_j}$-measurable, $D^{\kappa_j}(\alpha_j)$ and $X^{\kappa_j}(\alpha_j)$ are $\mathcal{F}_{\alpha_j}$-measurable, and hence $D^{\kappa_j}(t)$, $D^{\kappa_j}(\alpha_j)$ and $X^{\kappa_j}(\alpha_j)$ are $\mathcal{F}_t$-measurable on $\{\alpha_j \leq t\}$. Therefore, $\{\beta_j \leq t\}$ is $\mathcal{F}_t$-measurable and β_j is an $\{\mathcal{F}_t\}$-stopping time. Similarly,

$$\{\gamma_j \leq t\} = \{\alpha_j \leq t\} \cap \{D^{\kappa_j}(t) - D^{\kappa_j}(\alpha_j) \geq X^{\kappa_j}(\alpha_j)\}, \qquad j \geq 1,$$

and thus γ_j is also a stopping time. For jobs initially in network ($j \leq 0$), the queue index κ_j and the number of jobs p_j ahead of j at time 0 are $\mathcal{F}_0$-measurable, and since $\{\beta_j \geq t\} = \{D^{\kappa_j}(t) \geq p_j\}$ and $\{\gamma_j \geq t\} = \{D^{\kappa_j}(t) \geq p_j + 1\}$, β_j and γ_j for $j \leq 0$ are also $\{\mathcal{F}_t\}$-stopping times. Consequently, the age process $a_j^{(N)}$ defined in (5.1.5) is $\mathcal{F}_t$-adapted for all j.

Finally, recall that $\chi_j(t)$ is the queue size observed by job j at time $t \geq \alpha_j$, defined in (5.1.10). Since κ_j is $\mathcal{F}_{\alpha_j}$-measurable, $\mathbb{1}\left(\alpha_j \leq t\right)\chi_j(t)$ is $\{\mathcal{F}_t\}$-adapted, and also $\mathcal{F}_t$-progressive, as it is right-continuous. Therefore, since $\alpha_j \leq \gamma_j$, $\chi_j^D = X^{\kappa_j}(\gamma_j-)$ is $\mathcal{F}_{\gamma_j}$-measurable and hence D_ℓ defined in (5.1.21) is $\{\mathcal{F}_t\}$-adapted. Moreover, since γ_j and a_j are $\{\mathcal{F}_t\}$-stopping times, $a_j(\cdot)$ is $\{\mathcal{F}_t\}$-adapted and $\{\beta_j \leq t\} \subset \{\alpha_j \leq t\}$, ν_ℓ defined in (5.1.12) is $\{\mathcal{F}_t\}$-adapted. $\square$

5.A.2. Preliminary Independent Results. Recall that by non-idling condition, when $X^i(t) \geq 1$ for some $t \geq 0$, there exists a job receiving service at queue i at time t, which we denote by $J^i(t)$. Note that $\mathbb{1}(X^i(t) \geq 1)J^i(t)$ is well-defined for all $t \geq 0$.

LEMMA 5.22. Suppose Assumptions I, II.a, and III.a hold, and fix $k \geq 1$ and $i \in \{1, ..., N\}$. Then, on $\{z_k = (\mathfrak{E}, i)\}$, $u_{E(\tau_k)+1}$ and $v_{E(\tau_k)}$ are independent of $\mathcal{F}_{\tau_k}$, and for every $b \geq 0$,

$$\mathbb{1}\left(z_k = (\mathfrak{E}, i)\right)\mathbb{P}\left\{u_{E(\tau_k)+1} > b \middle| \mathcal{F}_{\tau_k}\right\} = \mathbb{1}\left(z_k = (\mathfrak{E}, i)\right)\overline{G}_E(b), \tag{5.A.3}$$

and

$$\mathbb{1}\left(z_k = (\mathfrak{E}, i)\right)\mathbb{P}\left\{v_{E(\tau_k)} > b \middle| \mathcal{F}_{\tau_k}\right\} = \mathbb{1}\left(z_k = (\mathfrak{E}, i)\right)\overline{G}(b). \tag{5.A.4}$$

Moreover, on $\{z_k = (\mathfrak{D}, i), X^i(\tau_k) \geq 1\}$, $v_{J^i(\tau_k)}$ is independent of $\mathcal{F}_{\tau_k}$ and for every $b \geq 0$,

(5.A.5)
$$\mathbb{1}\left(z_k = (\mathfrak{D}, i), X^i(\tau_k) \geq 1\right) \mathbb{P}\left\{v_{J^i(\tau_k)} > b \middle| \mathcal{F}_{\tau_k}\right\} = \mathbb{1}\left(z_k = (\mathfrak{D}, i), X^i(\tau_k) \geq 1\right) \overline{G}(b).$$

We prove Lemma 5.22 at end of this section, after we prove the following auxiliary result: we claim that on set

$$A_{k,i,n} \doteq \{z_k = (\mathfrak{E}, i), E(\tau_k) = n\}, \tag{5.A.6}$$

where the k^{th} event is the arrival of job n, the state of the network up to time τ_k only depends on the arrival times and random queue choices of the first n jobs, and service time of the first $n-1$ jobs, or equivalently, is $\mathcal{F}^{\text{Data}}_{\leq n}$-measurable, where

$$\mathcal{F}^{\text{Data}}_{\leq n} \doteq \sigma(I_0) \vee \sigma(u_j, \iota_j; j = 1, 2, ..., n) \vee \sigma(v_j; j \leq n-1). \tag{5.A.7}$$

Similarly, on the set

$$B_{k,i,m} \doteq \{z_k = (\mathfrak{D}, i), X^i(\tau_k) \geq 1, J^i(\tau_k) = m\}. \tag{5.A.8}$$

where a job has just departed queue i and job m has just entered service, we expect the state of the network up to time τ_k not to depend on the service time of job m, or equivalently, to be $\mathcal{F}^{\text{Data}}_{-m}$-measurable, where

$$\mathcal{F}^{\text{Data}}_{-m} \doteq \sigma(I_0) \vee \sigma(u_j, \iota_j; j \geq 1) \vee \sigma(v_j; j \in \mathbb{Z} \backslash \{m\}), \tag{5.A.9}$$

REMARK 5.23. Note that $A_{k,i,n}$ is clearly $\mathcal{F}_{\tau_k}$-measurable. Also, since $J^i(t)$ is the job with the largest index who has already entered service in queue i, we have

$$\mathbb{1}\left(X^i(t) \geq 1\right) J^i(t) = \mathbb{1}\left(X^i(t) \geq 1\right) \max\left\{j : \beta_j \leq t, \kappa_j = i\right\}.$$

Since β_j is an $\{\mathcal{F}_t\}$-stopping time and κ_j is $\mathcal{F}_{\beta_j}$-measurable by Lemma 5.19, the right-hand side above is $\{\mathcal{F}_t\}$-adapted, and hence $\{\mathcal{F}_t\}$-progressive because it is right-continuous as a function of t. Therefore, $\mathbb{1}\left(X^i(\tau_k) \geq 1\right) J^i(\tau_k)$ and hence, $B_{k,i,m}$ are τ_k-measurable.

Recall that for a sigma algebra $\mathcal{F}$ and a subset A, the trace of $\mathcal{F}$ on A is defined as $\mathcal{F} \wedge A = \{A \cap B : B \in \mathcal{F}\}$.

LEMMA 5.24. Suppose Assumptions of Lemma 5.19 holds. Then, for every $k, n \geq 1$, $i = 1, ..., N$ and $m \in \mathbb{Z}$,

(a) $A_{k,i,n}$ is $\mathcal{F}^{\text{Data}}_{\leq n}$-measurable and $\mathcal{F}_{\tau_k} \wedge A_{k,i,n} \subset \mathcal{F}^{\text{Data}}_{\leq n} \wedge A_{k,i,n}$.

(b) $B_{k,i,m}$ is $\mathcal{F}^{\text{Data}}_{-m}$-measurable and $\mathcal{F}_{\tau_k} \wedge B_{k,i,m} \subset \mathcal{F}^{\text{Data}}_{-m} \wedge B_{k,i,m}$.

PROOF. For part a, consider an identical queuing network driven by the alternative data

$$\widetilde{Data} = (\tilde{I}_0, \tilde{u}_j, \tilde{\iota}_j; j \geq 1, \tilde{v}_j; j \in \mathbb{Z}),$$

coupled with our original network so that $\tilde{I}_0 = I_0$, $\tilde{u}_j = u_j, \tilde{\iota}_j = \iota_j; j = 1, 2, ..., n$ and $\tilde{v}_j = v_j; j \leq n-1$, but $\tilde{u}_j = 1$ and $\tilde{\iota}_j(d') = \tilde{\iota}_j(d') = 1$ for $j \geq n+1$ and $d' = 1, ..., d$ and $\tilde{v}_j = 1$ for $j \geq n$. Hence,

$$\tilde{\mathcal{F}}^{\widetilde{\text{Data}}} \doteq \sigma(\tilde{I}_0, \tilde{u}_j, \tilde{\iota}_j; j \geq 1, \tilde{v}_j; j \in \mathbb{Z}) = \sigma(I_0, \iota_j, u_j; j = 1, ..., n, v_j; j \leq n-1) = \mathcal{F}^{\text{Data}}_{\leq n}.$$

Quantities associated to the alternative network are being distinguished from those of the original network by a tilde. Since the arrival time and random queue choices of the first n jobs and the service time of the first $n-1$ jobs are identical, two network are identical up to the arrival time $\alpha_n = \tilde{\alpha}_n$ of job n. In particular, $\tilde{A}_{k,i,n} \doteq \{\tilde{E}(\tilde{\tau}_k) = n, \tilde{z}_k = (\mathfrak{E}, i)\} = A_{k,i,n}$, and since $\tilde{A}_{k,i,n}$ is $\tilde{\mathcal{F}}_{\tilde{\tau}_k}$-measurable, so is $A_{k,i,n}$. But by Lemma 5.19, $\tilde{\mathcal{F}}_{\tilde{\tau}_k} \subset \tilde{\mathcal{F}}^{\widetilde{\text{Data}}} = \mathcal{F}^{\text{Data}}_{\leq n}$, and hence $A_{k,i,n}$ is $\mathcal{F}^{\text{Data}}_{\leq n}$-measurable.

Moreover, for all $t \geq 0$ and $i = 1, ..., n$,

$$E^i(t \wedge \alpha_n) = \tilde{E}^i(t \wedge \alpha_n), \qquad D^i(t \wedge \alpha_n) = \tilde{D}^i(t \wedge \alpha_n),$$

and hence, by [**14**, Theorem T3 of Section III.1] and definition (5.1.27) of the filtration $\{\mathcal{F}_t; t \geq 0\}$ and since α_j is a $\{\mathcal{F}_t\}$-stopping time by Lemma 5.19, we have $\mathcal{F}_{\alpha_n} = \tilde{\mathcal{F}}_{\alpha_n}$. In particular, since $\alpha_n = \tau_k$ on $A_{k,i,n}$ and by another use of Lemma 5.19,

$$\mathcal{F}_{\tau_k} \wedge A_{k,i,n} = \mathcal{F}_{\alpha_n} \wedge A_{k,i,n} = \tilde{\mathcal{F}}_{\alpha_n} \wedge A_{k,i,n} \subset \mathcal{F}^{\text{Data}}_{\leq n} \wedge A_{k,i,n}.$$

Finally for part b, consider an alternative network driven by another $\widetilde{Data}$ with $\tilde{I}_0 = I_0$, $\tilde{u}_j = u_j, \tilde{\iota}_j = \iota_j; j \geq 1$ and $\tilde{v}_j = v_j; j \in \mathbb{Z}\backslash\{m\}$, but with $\tilde{v}_m = 1$. Having in mind that two systems are identical up to the service entry time β_m of job m and that $B_{k,i,m}$ is $\mathcal{F}_{\tau_k}$-measurable by Remark 5.23 and β_m is an $\{\mathcal{F}_t\}$-stopping time, the result follows from exactly similar arguments as in part a. □

LEMMA 5.25. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, suppose sub σ-algebras $\mathcal{G}_c$, $\mathcal{G}_f \subset \mathcal{F}$ and a set $A \in \mathcal{G}_c \cap \mathcal{G}_f$ are such such that $\mathcal{G}_f$ is finer that $\mathcal{G}_c$ on A, that is, $\mathcal{G}_c \wedge A \subset \mathcal{G}_f \wedge A$. Then, for every integrable random variable X,

$$\mathbb{1}(A)\,\mathbb{E}\left[X|\mathcal{G}_c\right] = \mathbb{1}(A)\,\mathbb{E}\left[\mathbb{E}\left[X|\mathcal{G}_f\right]|\mathcal{G}_c\right]. \tag{5.A.10}$$

PROOF. For every $B \in \mathcal{G}_c$, by definition, $A \cap B \in \mathcal{G}_c \wedge A$ and since $\mathcal{G}_c \wedge A \subset \mathcal{G}_f \wedge A$, we have $A \cap B \in \mathcal{G}_f \wedge A$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}(B)\,\mathbb{1}(A)\,\mathbb{E}\left[X|\mathcal{G}_f\right]\right] &= \mathbb{E}\left[\mathbb{1}(B\cap A)\,\mathbb{E}\left[X|\mathcal{G}_f\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}(B\cap A)\,X|\mathcal{G}_f\right]\right] \\ &= \mathbb{E}\left[\mathbb{1}(B)\,\mathbb{1}(A)\,X\right],\end{aligned}$$

where the second equality holds because $A \cap B \in \mathcal{G}_f \wedge A$ and the last equality is by tower property of conditional expectation. Hence, we proved $\mathbb{E}\left[\mathbb{1}(A)\,X|\mathcal{G}_c\right] = \mathbb{E}\left[\mathbb{1}(A)\,\mathbb{E}\left[X|\mathcal{G}_f\right]|\mathcal{G}_c\right]$. Since $A \in \mathcal{G}_c$, (5.A.10) follows. □

PROOF OF LEMMA 5.22. To see why (5.A.3) holds, we partition the set $\{z_k = (\mathfrak{E}, i)\}$ based on the index of the job which arrived at time τ_k, i.e. the value of $E(\tau_k)$:

$$\{z_k = (\mathfrak{E}, i)\} = \bigcup_{n\in\mathbb{N}} A_{k,i,n}, \tag{5.A.11}$$

where $A_{k,i,n}$ is defined in (5.A.6). Conditions of Lemma 5.25 with $A = A_{k,i,n}$, $X = \mathbb{1}(u_{n+1} > b)$, $\mathcal{G}_c = \mathcal{F}_{\tau_k}$ and $\mathcal{G}_f = \mathcal{F}^{\text{Data}}_{\leq n}$ is satisfied by Lemma 5.24.a and the fact that $A_{k,i,n}$ is $\mathcal{F}_{\tau_k}$-measurable. Therefore, using Lemma 5.25, we have

$$\mathbb{1}(A_{k,i,n})\,\mathbb{P}\left\{u_{n+1} > b|\mathcal{F}_{\tau_k}\right\} = \mathbb{1}(A_{k,i,n})\,\mathbb{E}\left[\mathbb{P}\left\{u_{n+1} > b|\mathcal{F}^{\text{Data}}_{\leq n}\right\}\Big|\mathcal{F}_{\tau_k}\right]. \tag{5.A.12}$$

By Assumptions I and III.a, inter-arrival time u_{n+1} has complimentary CDF $\overline{G}_E$ and is independent of initial conditions, all other inter-arrival times and all service times and hence, independent of $\mathcal{F}^{\text{Data}}_{\leq n}$. Therefore, $\mathbb{P}\{u_{n+1} > b|\mathcal{F}^{\text{Data}}_{\leq n}\} = \overline{G}_E(b)$, and when combined with (5.A.11) and (5.A.12), we have

$$\mathbb{1}\left(z_k = (\mathfrak{E}, i)\right)\mathbb{P}\left\{u_{E(\tau_k)+1} > b|\mathcal{F}_{\tau_k}\right\} = \sum_{n\in\mathbb{N}} \mathbb{1}\left(A_{k,i,n}\right)\overline{G}_E(b) = \mathbb{1}\left(z_k = (\mathfrak{E}, i)\right)\overline{G}_E(b).$$

This proves (5.A.3). Similarly, by another use of Lemma 5.25, with the same A, $\mathcal{G}_c$, $\mathcal{G}_f$ as above and with $X = \mathbb{1}(v_n > b)$, in the second equality below, we have

$$\begin{aligned}\mathbb{1}\left(z_k = (\mathfrak{E}, i)\right)\mathbb{P}\left\{v_{E(\tau_k)} > b|\mathcal{F}_{\tau_k}\right\} &= \sum_{n\in\mathbb{N}} \mathbb{1}\left(A_{k,i,n}\right)\mathbb{P}\left\{v_n > b|\mathcal{F}_{\tau_k}\right\} \\ &= \sum_{n\in\mathbb{N}} \mathbb{1}\left(A_{k,i,n}\right)\mathbb{E}\left[\mathbb{P}\left\{v_n > b|\mathcal{F}^{\text{Data}}_{\leq n}\right\}\middle|\mathcal{F}_{\tau_k}\right].\end{aligned}$$

By Assumption II.a and III.a, v_n has the complimentary CDF $\overline{G}$ and is independent of I_0, the arrival process and the service times of jobs $j < n$, and hence, is independent of $\mathcal{F}^{\text{Data}}_{\leq n}$. This yields $\mathbb{P}\{v_n > b|\mathcal{F}^{\text{Data}}_{\leq n}\} = \overline{G}(b)$. When combined with the last display, (5.A.4) follows from another use of (5.A.11).

Finally to see (5.A.5), we partition the set $\{z_k = (\mathfrak{D}), X^i(\tau_k) \geq 1\}$ based on the value of $J^i(\tau_k)$, that is,

$$\{z_k = (\mathfrak{D}), X^i(\tau_k) \geq 1\} = \bigcup_{m\in\mathbb{Z}} B_{k,i,m}, \tag{5.A.13}$$

where $B_{k,i,m}$ is defined in (5.A.8) and is $\mathcal{F}_{\tau_k}$-measurable by Remark 5.23. For every $m \in \mathbb{Z}$, conditions of Lemma 5.25 with $A = B_{k,i,m}$, $X = \mathbb{1}(v_m > b)$, $\mathcal{G}_c = \mathcal{F}_{\tau_k}$ and $\mathcal{G}_f = \mathcal{F}^{\text{Data}}_{-m}$ hold by Lemma 5.24.b, and hence, using Lemma 5.25, we have

$$\mathbb{1}\left(B_{k,i,m}\right)\mathbb{P}\left\{v_m > b|\mathcal{F}_{\tau_k}\right\} = \mathbb{1}\left(B_{k,i,m}\right)\mathbb{E}\left[\mathbb{P}\left\{v_m > b|\mathcal{F}^{\text{Data}}_{-m}\right\}\middle|\mathcal{F}_{\tau_k}\right]. \tag{5.A.14}$$

By Assumption II.a, v_m has complimentary CDF $\overline{G}$ and is independent of I_0, arrival process and all other service times, and hence, is independent of $\mathcal{F}^{\text{Data}}_{-m}$, which implies $\mathbb{P}\left\{v_m > b|\mathcal{F}^{\text{Data}}_{-m}\right\} = \overline{G}(b)$. When combined with (5.A.13) and (5.A.14), we conclude

$$\mathbb{1}\left(z_k = (\mathfrak{D}, i), X^i(\tau_k \geq 1)\right) \mathbb{P}\left\{v_{J^i(\tau_k)} > b \middle| \mathcal{F}_{\tau_k}\right\} = \sum_{m\in\mathbb{Z}} \mathbb{1}\left(B_{k,i,m}\right) \mathbb{P}\left\{v_m > b | \mathcal{F}_{\tau_k}\right\}$$
$$= \sum_{m\in\mathbb{Z}} \mathbb{1}\left(B_{k,i,m}\right) \overline{G}(b),$$

and (5.A.5) follows from another use of (5.A.13). □

Next, define $z_{k,1} \in \{\mathfrak{E}, \mathfrak{D}\}$ and $z_{k,2} \in \{1, ..., N\}$ to be the two components of τ_k, that is, $z_k = (z_{k,1}, z_{k,2})$. The next lemma shows that when the next event is an arrival, σ_k^i is conditionally independent of the queue index to which the new arriving job is routed.

LEMMA 5.26. Suppose Assumption I and II.a hold. Then, for every $k \geq 1$,

$$\mathbb{1}\left(z_{k,1} = \mathfrak{E}\right) \mathbb{P}\left\{\sigma_k^i > \tau_k + b_i, i = 1, ..., N | \mathcal{F}_{\tau_k}\right\}$$
$$= \mathbb{1}\left(z_{k,1} = \mathfrak{E}\right) \mathbb{P}\left\{\sigma_k^i > \tau_k + b_i, i = 1, ..., N | \mathcal{F}_{\tau_{k-1}}, \tau_k, z_{k,1}\right\}. \tag{5.A.15}$$

PROOF. Define $A_{k,n} = \{z_{k,1} = \mathfrak{E}, E(\tau_k) = n\}$, and note that $\{z_{k,1} = \mathfrak{E}\} = \cup_{n\in\mathbb{N}} A_{k,n}$. On $A_{k,n}$, $z_{k,2}$ is the queue index of the newly arrived job n, that is, $z_{k,2} = \kappa_n$, where κ_n is defined by (5.1.6) and is a measurable function of ι_n and $X^i(\tau_k), i = 1, ..., N$. Since $X^i(\tau_{k-1}), i = 1, ..., N$, are all $\mathcal{F}_{\tau_{k-1}}$-measurable by Lemma 5.19, we have

$$\mathcal{F}_{\tau_k} \wedge A_{n,k} = \left(\mathcal{F}_{\tau_{k-1}} \vee \sigma(\tau_k, z_k)\right) \wedge A_{k,n} = \left(\mathcal{F}_{\tau_{k-1}} \vee \sigma(\tau_k, z_{k,1}, \iota_n)\right) \wedge A_{k,n},$$

and therefore,

$$\mathbb{1}\left(A_{k,n}\right) \mathbb{P}\left\{\sigma_k^i > \tau_k + b_i, i = 1, ..., N | \mathcal{F}_{\tau_k}\right\}$$
$$= \mathbb{1}\left(A_{k,n}\right) \mathbb{P}\left\{\sigma_k^i > \tau_k + b_i, i = 1, ..., N | \mathcal{F}_{\tau_{k-1}}, \tau_k, z_{k,1}, \iota_n\right\}. \tag{5.A.16}$$

Moreover, for $i \in \mathfrak{B}_{k-1}$, σ_k^i is the departure time of the job $J^i(\tau_{k-1})$ which is receiving service at queue i at time τ_{k-1}, and hence, $\sigma_k^i = \beta_{J^i(\tau_{k-1})} + v_{J^i(\tau_{k-1})}$. By Remark 5.23, $J^i(\tau_{k-1})$ is $\mathcal{F}_{\tau_{k-1}}$-measurable, and since β_j is a $\{\mathcal{F}_t\}$-stopping time for all $j \in \mathbb{Z}$

and $\beta_{J^i(\tau_{k-1})} \leq \tau_{k-1}$, $\beta_{J^i(\tau_{k-1})}$ is $\mathcal{F}_{\tau_{k-1}}$-measurable. Also, since service times are independent of the random queue choices, $v_{J^i(\tau_{k-1})}$ is conditionally independent of ι_n, given $\mathcal{F}_{\tau_{k-1}}$. On the other hand, if $i \notin \mathfrak{B}_{k-1}$, $\sigma_k^i = \infty$. Consequently, for every $i = 1, ..., N$, σ_k^i is conditionally independent of ι_n, given $\mathcal{F}_{\tau_{k-1}}$, and therefore, (5.A.16) yields

$$\begin{aligned}
&\mathbb{1}\left(A_{k,n}\right) \mathbb{P}\left\{\sigma_k^i > \tau_k + b_i, i = 1, ..., N | \mathcal{F}_{\tau_k}\right\} \\
&\qquad = \mathbb{1}\left(A_{k,n}\right) \mathbb{P}\left\{\sigma_k^i > \tau_k + b_i, i = 1, ..., N | \mathcal{F}_{\tau_{k-1}}, \tau_k, z_{k,1}\right\}.
\end{aligned}$$

Equation (5.A.15) follows from summation of both sides of the above display over $n \in \mathbb{Z}$. □

Finally, we state our last auxiliary independence lemma. Let $X_1, X_2, ..., X_n$ and Y be $\mathbb{R} \cup \{\infty\}$-valued random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra such that $X_i, i = 1, ..., n$, and Y are conditionally independent given $\mathcal{G}$. For $i = 1, ..., n$, let $\overline{F}_i$ be the conditional complimentary CDF of X_i given $\mathcal{G}$, that is

$$\mathbb{P}\left\{X_i > a | \mathcal{G}\right\} = \overline{F}_i(a), \qquad a > 0.$$

Define $T \doteq \min(X_1, ..., X_n, Y)$ and let Z be a discrete-valued random variable such that $\{Y < \min(X_1, ..., X_n)\} = \{Z = z_0\}$ for some value z_0.

LEMMA 5.27. Suppose $X_1, ..., X_n, Y, T$ and Z are as defined above. Then, on $\{Z = z_0\}$, $X_i - T; i = 1, ..., n$, are conditionally independent given $\mathcal{G}$, T and Z, that is, for every $b_1, ..., b_n \geq 0$,

$$\begin{aligned}
\text{(5.A.17)} \qquad &\mathbb{1}\left(Z = z_0\right) \mathbb{P}\left\{X_i > T + b_i; i = 1, ..., n \middle| \mathcal{G}, T, Z\right\} \\
&\qquad = \mathbb{1}\left(Z = z_0\right) \prod_{i=1}^{n} \mathbb{P}\left\{X_i > T + b_i \middle| \mathcal{G}, T, Z\right\},
\end{aligned}$$

and each have the following conditional marginal:

$$\text{(5.A.18)} \quad \mathbb{1}\left(Z = z_0\right) \mathbb{P}\left\{X_i > T + b \middle| \mathcal{G}, T, Z\right\} = \mathbb{1}\left(Z = z_0\right) \frac{\overline{F}_i(T + b)}{\overline{F}_i(T)}, \qquad \forall b \geq 0.$$

PROOF. First, we claim that for every $B \in \mathcal{B}[0,\infty)$,

$$\mathbb{E}\left[\mathbb{1}(T\in B)\,\mathbb{1}(Z=z_0)\prod_{i=1}^{n}\mathbb{1}(X_i>T+b_i)\Big|\mathcal{G}\right] = \mathbb{E}\left[\mathbb{1}(T\in B)\,\mathbb{1}(Z=z_0)\prod_{i=1}^{n}\frac{\overline{F}_i(T+b_i)}{\overline{F}_i(T)}\Big|\mathcal{G}\right]. \tag{5.A.19}$$

To prove the claim, note that since $T=Y$ on $\{Z=z_0\}$, we have

$$\{T\in B, Z=z_0, X_i>T+b_i; i=1,...,n\} = \{Y\in B, X_i>Y+b_i; i=1,...,n\},$$

Therefore, using the conditional independence of $X_1,...,X_n$ and Y, the left-hand side of (5.A.19) is equal to

$$\mathbb{P}\left\{Y\in B, X_i>Y+b_i, i=1,...,n\Big|\mathcal{G}\right\} = \mathbb{E}\left[\mathbb{1}(Y\in B)\prod_{i=1}^{n}\overline{F}_i(Y+b_i)\Big|\mathcal{G}\right].$$

Similarly, since $\{T\in B, Z=z_0\} = \{Y\in B, X_i>Y, i=1,...,n\}$, by conditionally independence of $X_1,...,X_n$ and Y given $\mathcal{G}$, the right-hand side of (5.A.19) is equal to

$$\mathbb{E}\left[\mathbb{1}(Y\in B, X_i>Y, i=1,...,n)\prod_{i=1}^{n}\frac{\overline{F}_i(Y+b_i)}{\overline{F}_i(Y),}\Big|\mathcal{G}\right] = \mathbb{E}\left[\mathbb{1}(Y\in B)\prod_{i=1}^{n}\overline{F}_i(Y+b_i)\Big|\mathcal{G}\right].$$

The claim (5.A.19) follows from the last two displays. Moreover, by definition of conditional expectation, (5.A.19) implies

$$\mathbb{1}(Z=z_0)\,\mathbb{P}\left\{X_i>T+b_i; i=1,...,n\Big|\mathcal{G},T,Z\right\} = \mathbb{1}(Z=z_0)\prod_{i=1}^{n}\frac{\overline{F}_i(T+b_i)}{\overline{F}_i(T)}. \tag{5.A.20}$$

Now, substituting $b_i=b$ and $b_{i'}=0, i'\neq i$, in (5.A.20) and observing that $X_i>T$ on $\{Z=z_0\}$ for all i, (5.A.18) follows. Finally, (5.A.17) is obtained by substituting each term of the product on the right-hand side of (5.A.20) from the equation (5.A.18). $\square$

5.A.3. Proofs of Lemma 5.15.

PROOF OF LEMMA 5.15. We prove the lemma using induction on k. Assume by induction hypothesis that $\mathbb{P}\{\Omega_{k-1}\} = 1$, ξ_k and σ_k^i, $i = 1, ..., N$, are conditionally independent given $\mathcal{F}_{\tau_{k-1}}$, and (5.3.23) and (5.3.24) hold with k replaced by $k-1$, that is for every $b > 0$,

$$\mathbb{P}\left\{\xi_k - \tau_{k-1} > b \middle| \mathcal{F}_{\tau_{k-1}}\right\} = \frac{\overline{G}_E(R_E(\tau_{k-1}) + b)}{\overline{G}_E(R_E(\tau_{k-1}))}, \tag{5.A.21}$$

and for every $i = 1, ..., N$,

$$\mathbb{1}\left(i \in \mathfrak{B}_{k-1}\right) \mathbb{P}\left\{\sigma_k^i - \tau_{k-1} > b \middle| \mathcal{F}_{\tau_{k-1}}\right\} = \mathbb{1}\left(i \in \mathfrak{B}_{k-1}\right) \frac{\overline{G}(a^i(\tau_{k-1}) + b)}{\overline{G}(a^i(\tau_{k-1}))}, \tag{5.A.22}$$

First, we prove $\mathbb{P}\{\tilde{\Omega}_k\} = 1$. Note that using (5.3.22) with k replaced by $k-1$ and since $\mathbb{P}\{\tilde{\Omega}_{k-1}\} = 1$, we have

$$\tau_k - \tau_{k-1} = \min\{\xi_k - \tau_{k-1}, \sigma_k^i - \tau_{k-1}; i = 1, ..., N\},$$

almost surely. By induction hypothesis, given $\mathcal{F}_{\tau_{k-1}}$, $\xi_k - \tau_{k-1}$ and $\sigma_k^i - \tau_{k-1}, i = 1, ..., N$, are conditionally independent, and by (5.A.21) and (5.A.22) and since G_E and $\overline{G}$ have densities g_E and g by Assumptions I and II.a, are absolutely continuous with respect to the Lebesgue measure on $(0, \infty)$. Therefore, they are all distinct and strictly positive, and in particular, $\mathbb{P}\left\{\tau_k > \tau_{k-1} | \mathcal{F}_{\tau_{k-1}}\right\} = 1$. Moreover, since inter-arrival time is almost surely positive by Assumption I, all future arrival times are strictly greater than ξ_k. Similarly, all service times are almost surely positive by Assumption II.a and hence, all future departure times from a queue $i \in \mathfrak{B}_{k-1}$ is strictly greater than σ_k^i, and all departure times from a queue $i \notin \mathfrak{B}_{k-1}$ is strictly greater than ξ_k, almost surely. Therefore, given $\mathcal{F}\tau_{k-1}$, τ_k is almost surely distinct, that is, $\mathbb{P}\left\{\tilde{\Omega}_k | \mathcal{F}_{\tau_{k-1}}\right\} = 1$, and by the tower property of conditional expectation,

$$\mathbb{P}\left\{\tilde{\Omega}_k\right\} = \mathbb{E}\left[\mathbb{P}\left\{\tilde{\Omega}_k = 1 | \mathcal{F}_{\tau_{k-1}}\right\}\right] = 1.$$

Next, we establish (5.3.23) and (5.3.24) and the conditional independence of ξ_{k+1} and σ_{k+1}^i given $\mathcal{F}_{\tau_k}$. Since $\mathbb{P}\{\tilde{\Omega}_k\} = 1$, by (5.3.22) we have

$$\tau_k = \min(\xi_k, \sigma_k^i; i = 1, ..., N), \tag{5.A.23}$$

almost surely. Also, by representation (5.3.17) for the filtration $\{\mathcal{F}_t\}$ and by Theorem T30 of Section A.2 in **[14]**, $\mathcal{F}_{\tau_k} = \sigma(I_0) \vee \sigma(\tau_{k'}, z_{k'}; 1 \leq k' \leq k)$, and therefore $\mathcal{F}_{\tau_k} = \mathcal{F}_{\tau_{k-1}} \vee \sigma(\tau_k, z_k)$. We partition Ω based on the value of z_k and queue lengths right before the event time τ_k and show the conditional independence on each subset of Ω. Then conclude the proof.

First, for every $i_* \in \{1, ..., N\}$, let $A_{i_*} \doteq \{z_k = (\mathfrak{E}, i_*)\}$ be the set of realizations on which the k^{th} event (which is already proved to be distinct) is an arrival to the queue i_*. On A_{i*}, $\tau_k = \xi_k$ and $\xi_{k+1} = \tau_k + u_{E(\tau_k)+1}$, and hence using (5.A.3) of Lemma 5.22, for every $b_0, b_1, ..., b_N \geq 0$ we have,

$$
\begin{aligned}
&\mathbb{1}(A_{i*})\,\mathbb{P}\left\{\xi_{k+1} > \tau_k + b_0, \sigma^i_{k+1} > \tau_k + b_i, i = 1, ..., N | \mathcal{F}_{\tau_k}\right\} \\
&\qquad = \mathbb{1}(A_{i*})\,\mathbb{P}\left\{u_{E(\tau_k)+1} > b_0, \sigma^i_{k+1} > \tau_k + b_i, i = 1, ..., N | \mathcal{F}_{\tau_k}\right\} \\
&\qquad = \mathbb{1}(A_{i*})\,\overline{G}_E(b_0)\mathbb{P}\left\{\sigma^i_{k+1} > \tau_k + b_i, i = 1, ..., N | \mathcal{F}_{\tau_k}\right\}.
\end{aligned} \tag{5.A.24}
$$

In particular, substituting $b_0 = b$ and $b_i = 0$, $i = 1, ..., N$, in (5.A.24) and using the fact that $R_E(\tau_k) = 0$ on A_{i_*}, we have

$$
\mathbb{1}(A_{i_*})\,\mathbb{P}\left\{\xi_{k+1} > \tau_k + b \middle| \mathcal{F}_{\tau_k}\right\} = \mathbb{1}(A_{i_*})\,\overline{G}_E(b) = \mathbb{1}(A_{i_*})\,\frac{\overline{G}_E(R_E(\tau_k) + b)}{\overline{G}_E(R_E(\tau_k))}. \tag{5.A.25}
$$

Next, we look at the joint conditional distribution of $\sigma^i_k, i = 1, ..., N$, given $\mathcal{F}_{\tau_k}$, using Lemma 5.27. By induction hypothesis, ξ_k and $\sigma^i_k, i = 1, ..., N$, are conditionally independent given $\mathcal{F}_{\tau_{k-1}}$. Consider the setup of Lemma 5.27 with $Y = \xi_k$ and $X_i = \sigma^i, i = 1, ..., N$, and recall the notation $z_k = (z_{k,1}, z_{k,2})$. Using (5.A.23) and since $\{\xi_k \leq \min(\sigma^i_k; i = 1, ..., N)\} = \{z_k = (\mathfrak{E}, i) \text{ for some } i = 1, ..., N\} = \{z_{k,1} = \mathfrak{E}\}$, we have $T = \tau_k$, $Z = z_{k,1}$ and $z_0 = \mathfrak{E}$. Therefore, and since $A_{i_*} \subset \{z_{k,1} = \mathfrak{E}\}$ and using (5.A.15) of Lemma 7.30 in the second and the last equality and (5.A.17) of Lemma

5.27 in the third equality, we have

$$
\begin{aligned}
&\mathbb{1}(A_{i_*})\,\mathbb{P}\left\{\sigma_k^i > \tau_k + b_i, i = 1, ..., N | \mathcal{F}_{\tau_k}\right\} \\
&\qquad = \mathbb{1}(A_{i_*})\,\mathbb{1}(z_{k,1} = \mathfrak{E})\,\mathbb{P}\left\{\sigma_k^i > \tau_k + b_i, i = 1, ..., N | \mathcal{F}_{\tau_{k-1}}, \tau_k, z_k\right\} \\
&\qquad = \mathbb{1}(A_{i_*})\,\mathbb{1}(z_{k,1} = \mathfrak{E})\,\mathbb{P}\left\{\sigma_k^i > \tau_k + b_i, i = 1, ..., N | \mathcal{F}_{\tau_{k-1}}, \tau_k, z_{k,1}\right\} \\
&\qquad = \mathbb{1}(A_{i_*}) \prod_{i=1}^{N} \mathbb{P}\left\{\sigma_k^i > \tau_k + b_i | \mathcal{F}_{\tau_{k-1}}, \tau_k, z_{k,1}\right\} \\
&\qquad = \mathbb{1}(A_{i_*}) \prod_{i=1}^{N} \mathbb{P}\left\{\sigma_k^i > \tau_k + b_i | \mathcal{F}_{\tau_k}\right\}. \qquad (5.A.26)
\end{aligned}
$$

Moreover, if $i \in \mathfrak{B}_{k-1}$, the conditional distribution of σ_i^k is given by (5.A.22), and hence using (5.A.18) of Lemma 5.27 and the fact that on A_{i_*}, $a^i(\cdot)$ grows linearly on $(\tau_{k-1}, \tau_k]$, we have

$$
\begin{aligned}
&\mathbb{1}(A_{i_*})\,\mathbb{1}(i \in \mathfrak{B}_{k-1})\,\mathbb{P}\left\{\sigma_k^i > \tau_k + b | \mathcal{F}_{\tau_k}\right\} \\
&\qquad = \mathbb{1}(A_{i_*})\,\mathbb{1}(i \in \mathfrak{B}_{k-1})\,\mathbb{P}\left\{\sigma_k^i - \tau_{k-1} > \tau_k - \tau_{k-1} + b \big| \mathcal{F}_{\tau_{k-1}}, \tau_k, z_k\right\} \\
&\qquad = \mathbb{1}(A_{i_*})\,\mathbb{1}(i \in \mathfrak{B}_{k-1})\,\frac{\overline{G}(a^i(\tau_{k-1}) + \tau_k - \tau_{k-1} + b)}{\overline{G}(a^i(\tau_{k-1}) + \tau_k - \tau_{k-1})} \\
&\qquad = \mathbb{1}(A_{i_*})\,\mathbb{1}(i \in \mathfrak{B}_{k-1})\,\frac{\overline{G}(a^i(\tau_k) + b)}{\overline{G}(a^i(\tau_k))}. \qquad (5.A.27)
\end{aligned}
$$

Now we further partition A_{i_*} into two subsets, based on the length of queue i_* right before the arrival. First, consider the set $A_{i_*,1} \doteq \{z_k = (\mathfrak{E}, i_*), X^{i_*}(\tau_{k-1}) \geq 1\}$ on which queue i_* is non-empty right before the arrival. Since an arrival to a non-empty queue does not change the set of busy servers $\mathfrak{B}_k = \mathfrak{B}_{k-1}$, and also $\sigma_{k+1}^i = \sigma_k^i$ for all $i = 1, ..., N$, on $A_{i_*,1}$. Substituting these identities in (5.A.26) and (5.A.27), respectively, we have

$$
\begin{aligned}
&\mathbb{1}(A_{i_*,1})\,\mathbb{P}\left\{\sigma_{k+1}^i > \tau_k + b_i, i = 1, ..., N | \mathcal{F}_{\tau_k}\right\} \\
&\qquad = \mathbb{1}(A_{i_*,1}) \prod_{i=1}^{N} \mathbb{P}\left\{\sigma_{k+1}^i > \tau_k + b_i | \mathcal{F}_{\tau_k}\right\}, \qquad (5.A.28)
\end{aligned}
$$

and for every $i = 1, ..., N$,

(5.A.29)

$$\mathbb{1}(A_{i_*,1})\,\mathbb{1}(i \in \mathfrak{B}_k)\,\mathbb{P}\left\{\sigma^i_{k+1} > \tau_k + b \middle| \mathcal{F}_{\tau_k}\right\} = \mathbb{1}(A_{i_*,1})\,\mathbb{1}(i \in \mathfrak{B}_k)\,\frac{\overline{G}(a^i(\tau_k)+b)}{\overline{G}(a^i(\tau_k))}.$$

On the other hand, let $A_{i_*,2} \doteq \{z_k = (\mathfrak{E}, i_*), X^{i_*}(\tau_{k-1}) = 0\}$ be the set of realizations on which the k^{th} event (which is already proved to be distinct) is an arrival to an empty queue i_*. This arrival makes the previously idle server i_* busy, that is, $\mathfrak{B}_k = \mathfrak{B}_{k-1} \cup \{i_*\}$, and also $\sigma^i_{k+1} = \sigma^i_k$ for all $i \neq i_*$ and $\sigma^{i_*}_k = \tau_{k+1} + v_{J^{i_*}(\tau_k)}$. Therefore, using (5.A.4) of Lemma 5.22 in the second equality and (5.A.26) in the third equality, we have

$$\begin{aligned}(5.A.30)\qquad \mathbb{1}(A_{i_*,2})\,\mathbb{P}\left\{\sigma^i_{k+1} > \tau_k + b_i; i = 1, ..., N | \mathcal{F}_{\tau_k}\right\} &\\ = \mathbb{1}(A_{i_*,2})\,\mathbb{P}\left\{v_{J^{i_*}(\tau_k)} > b_{i_*}, \sigma^i_k > \tau_k + b_i; i \neq i_* | \mathcal{F}_{\tau_k}\right\} &\\ = \mathbb{1}(A_{i_*,2})\,\overline{G}(b_{i_*})\mathbb{P}\left\{\sigma^i_k > \tau_k + b_i; i \neq i_* | \mathcal{F}_{\tau_k}\right\} &\\ = \mathbb{1}(A_{i_*,2})\,\overline{G}(b_{i_*}) \prod_{i=1, i\neq i_*}^{N} \mathbb{P}\left\{\sigma^i_k > \tau_k + b_i | \mathcal{F}_{\tau_k}\right\}. &\end{aligned}$$

In particular, by substituting $b_{i_*} = b$ and $b_i = 0$ for $i = 1, ..., N, i \neq i_*$ in (5.A.30) and since $i_* \in \mathfrak{B}_k$ on $A_{i_*,2}$ and the age of the job $J^{i_*}(\tau_k)$ which has just begun receiving service in queue i_* is zero, we have

$$\begin{aligned}\mathbb{1}(A_{i_*,2})\,\mathbb{1}(i_* \in \mathfrak{B}_k)\,\mathbb{P}\left\{\sigma^{i*}_{k+1} > \tau_k + b | \mathcal{F}_{\tau_k}\right\} &= \mathbb{1}(A_{i_*,2})\,\mathbb{1}(i_* \in \mathfrak{B}_k)\,\overline{G}(b) \\ (5.A.31)\qquad &= \mathbb{1}(A_{i_*,2})\,\mathbb{1}(i_* \in \mathfrak{B}_k)\,\frac{\overline{G}(a^{i_*}(\tau_k)+b)}{\overline{G}(a^{i_*}(\tau_k))}.\end{aligned}$$

Finally, recall again that $\mathfrak{B}_k = \mathfrak{B}_{k-1} \cup \{i_*\}$ on $A_{i_*,2}$. For every $i \neq i_*$, substituting $\mathbb{1}(i \in \mathfrak{B}_k) = \mathbb{1}(i \in \mathfrak{B}_{k-1})$ and $\sigma^i_{k+1} = \sigma^i_k$ in (5.A.27), we have

(5.A.32)

$$\mathbb{1}(A_{i_*,2})\,\mathbb{1}(i \in \mathfrak{B}_k)\,\mathbb{P}\left\{\sigma^i_{k+1} > \tau_k + b \middle| \mathcal{F}_{\tau_k}\right\} = \mathbb{1}(A_{i_*,2})\,\mathbb{1}(i \in \mathfrak{B}_k)\,\frac{\overline{G}(a^i(\tau_k)+b)}{\overline{G}(a^i(\tau_k))}.$$

Now consider the set $B_{i_*} = \{z_k = (\mathfrak{D}, i_*)\}$ of realizations on which the k^{th} event (which has already proved to be distinct) is a departure from queue i_*. By induction hypothesis, ξ_k and $\sigma^i_k, i = 1, ..., N$, are conditionally independent given

$\mathcal{F}_{\tau_{k-1}}$. Consider the setup of Lemma 5.27 with $Y = \sigma_k^{i^*}$, $X_{i_*} = \xi_k$ and $X_i = \sigma_k^i$ for $i = 1, ..., N, i \neq i_*$. Using (5.A.23) and since $\{\sigma_k^{i_*} < \min(\xi_k, \sigma_k^i; i = 1, ..., N, i \neq i_*)\} = B_{i_*}$, we have $T = \tau_k$, $Z = z_k$ and $z_0 = (\mathfrak{D}, i_*)$. Therefore, using (5.A.17) of Lemma 5.27 in the second equality below, we have

$$\begin{aligned}
&\mathbb{1}(B_{i_*})\,\mathbb{P}\left\{\xi_k \geq \tau_k + b_0, \sigma_k^i > \tau_k + b_i, i = 1, ..., N, i \neq i_* | \mathcal{F}_{\tau_k}\right\} \\
&\qquad = \mathbb{1}(B_{i_*})\,\mathbb{1}(\tau_k = (\mathfrak{D}, i)) \\
&\qquad\quad \times \mathbb{P}\left\{\xi_k \geq \tau_k + b_0, \sigma_k^i > \tau_k + b_i, i = 1, ..., N, i \neq i_* | \mathcal{F}_{\tau_{k-1}}, \tau_k, z_k\right\} \\
&\qquad = \mathbb{1}(B_{i_*})\,\mathbb{P}\left\{\xi_k > \tau_k + b_i | \mathcal{F}_{\tau_k}\right\} \prod_{i=1, i\neq i_*}^{N} \mathbb{P}\left\{\sigma_k^i > \tau_k + b_i | \mathcal{F}_{\tau_k}\right\}.
\end{aligned} \tag{5.A.33}$$

Moreover, given $\mathcal{F}_{\tau_{k-1}}$, the conditional distribution of ξ_k and σ_k^i for $i \in \mathfrak{B}_{k-1} \backslash \{i_*\}$ are given by (5.A.21) and (5.A.22), respectively. Therefore, using (5.A.18) of Lemma 5.27, we have

$$\begin{aligned}
&\mathbb{1}(B_{i_*})\,\mathbb{P}\left\{\xi_k > \tau_k + b \middle| \mathcal{F}_{\tau_k}\right\} \\
&\qquad = \mathbb{1}(B_{i_*})\,\mathbb{P}\left\{\xi_k - \tau_{k-1} > \tau_k - \tau_{k-1} + b \middle| \mathcal{F}_{\tau_k}\right\} \\
&\qquad = \mathbb{1}(B_{i_*})\,\frac{\overline{G}_E(R_E(\tau_{k-1}) + \tau_k - \tau_{k-1} + b)}{\overline{G}_E(R_E(\tau_{k-1}) + \tau_k - \tau_{k-1})},
\end{aligned} \tag{5.A.34}$$

and for every $i \neq i_*$,

$$\begin{aligned}
&\mathbb{1}(B_{i_*})\,\mathbb{1}(i \in \mathfrak{B}_{k-1})\,\mathbb{P}\left\{\sigma_k^i > \tau_k + b \middle| \mathcal{F}_{\tau_k}\right\} \\
&\qquad = \mathbb{1}(B_{i_*})\,\mathbb{1}(i \in \mathfrak{B}_{k-1})\,\frac{\overline{G}(a^i(\tau_{k-1}) + \tau_k - \tau_{k-1} + b)}{\overline{G}(a^i(\tau_{k-1}) + \tau_k - \tau_{k-1})}.
\end{aligned} \tag{5.A.35}$$

On B_{i_*}, $\xi_{k+1} = \xi_k$ and $R_E(\cdot)$ grows linearly on $(\tau_{k-1}, \tau_k]$. Hence, it follows from (5.A.34) that

$$\mathbb{1}(B_{i_*})\,\mathbb{P}\left\{\xi_{k+1} > \tau_k + b \middle| \mathcal{F}_{\tau_k}\right\} = \mathbb{1}(B_{i_*})\,\frac{\overline{G}_E(R_E(\tau_k) + b)}{\overline{G}_E(R_E(\tau_k))}. \tag{5.A.36}$$

Similarly, on B_{i_*}, for every $i = 1, ..., N$, $i \neq i_*$, $\sigma_{k+1}^i = \sigma_k^i$, $i \in \mathfrak{B}_k$ if and only if $i \in \mathfrak{B}_{k-1}$ and $a^i(\cdot)$ grows linearly on $(\tau_{k-1}, \tau_k]$. Hence, it follows from (5.A.35) that for every $i = 1, ..., N$, and $i \neq i_*$,

(5.A.37)
$$\mathbb{1}\left(B_{i_*}\right)\mathbb{1}\left(i \in \mathfrak{B}_k\right)\mathbb{P}\left\{\sigma^i_{k+1} > \tau_k + b \middle| \mathcal{F}_{\tau_k}\right\} = \mathbb{1}\left(B_{i_*}\right)\mathbb{1}\left(i \in \mathfrak{B}_k\right)\frac{\overline{G}(a^i(\tau_k)+b)}{\overline{G}(a^i(\tau_k))}.$$

The the value of $\sigma^{i_*}_{k+1}$ depends on the length of queue i_* right before the departure, and hence we further partition B_{i_*} into two parts. On

$$B_{i_*,1} = \{z_k = (\mathfrak{D}, i_*), X^{i_*}(\tau_{k-1}) = 1\},$$

queue i_* becomes empty after the departure at τ_k, and hence $\sigma^{i_*}_{k+1} = \infty$ by definition. Therefore, using (5.A.33) and trivial identity $\mathbb{P}\{\sigma^{i_*}_{k+1} > b_{i_*}\}$=1 in the last equality, we have

$$\begin{aligned}\mathbb{1}\left(B_{i_*,1}\right)&\mathbb{P}\left\{\xi_{k+1} \geq \tau_k + b_0, \sigma^i_{k+1} > \tau_k + b_i; i = 1, ..., N | \mathcal{F}_{\tau_k}\right\}\\ &= \mathbb{1}\left(B_{i_*,1}\right)\mathbb{P}\left\{\xi_k > \tau_k + b_0, \sigma^i_k > \tau_k + b_i; i = 1, ..., N, i \neq i_* | \mathcal{F}_{\tau_k}\right\}\\ &= \mathbb{1}\left(B_{i_*,1}\right)\mathbb{P}\left\{\xi_k > \tau_k + b_0 | \mathcal{F}_{\tau_k}\right\}\mathbb{P}\left\{\sigma^{i_*}_{k+1} > b_{i_*}\right\}\\ &\qquad \times \prod_{i=1, i\neq i_*}^{N}\mathbb{P}\left\{\sigma^i_k > \tau_k + b_i | \mathcal{F}_{\tau_k}\right\}.\end{aligned} \tag{5.A.38}$$

On the other hand, on $B_{i_*,2} = \{z_k = (\mathfrak{D}, i_*), X^{i_*}(\tau_{k-1}) \geq 2\}$, a new job, namely $J^{i_*}(\tau_k)$, enters service i_* right after departure at τ_k, and hence $\sigma^{i_*}_{k+1} = \tau_k + v_{J^{i_*}(\tau_k)}$ and $a^{i_*}(\tau_k) = 0$. Therefore, using (5.A.5) of Lemma 5.22 in the second equality and (5.A.33) in the third equality, we have

(5.A.39)
$$\begin{aligned}\mathbb{1}\left(B_{i_*,2}\right)&\mathbb{P}\left\{\xi_{k+1} \geq \tau_k + b_0, \sigma^i_{k+1} > \tau_k + b_i, i = 1, ..., N | \mathcal{F}_{\tau_k}\right\}\\ &= \mathbb{1}\left(B_{i_*,2}\right)\mathbb{P}\left\{v_{J^{i_*}(\tau_k)} > b_{i_*}, \xi_k \geq \tau_k + b_0, \sigma^i_k > \tau_k + b_i, i = 1, ..., N, i \neq i_* | \mathcal{F}_{\tau_k}\right\}\\ &= \mathbb{1}\left(B_{i_*,2}\right)\overline{G}(b_{i_*})\mathbb{P}\left\{\xi_k \geq \tau_k + b_0, \sigma^i_k > \tau_k + b_i, i = 1, ..., N, i \neq i_* | \mathcal{F}_{\tau_k}\right\}\\ &= \mathbb{1}\left(B_{i_*,2}\right)\overline{G}(b_{i_*})\mathbb{P}\left\{\xi_k > \tau_k + b_0 | \mathcal{F}_{\tau_k}\right\}\prod_{i=1, i\neq i_*}^{N}\mathbb{P}\left\{\sigma^i_k > \tau_k + b_i | \mathcal{F}_{\tau_k}\right\}.\end{aligned}$$

In particular, substituting $b_{i_*} = b$ and $b_i = 0$ for $i = 0, 1, ..., N, i \neq i_*$ in (5.A.39) and using the facts that $i_* \in \mathfrak{B}_k$ and $a^{i_*}(\tau_k) = 0$ on $B_{i_*,2}$, we have

$$\begin{aligned} \mathbb{1}\left(B_{i_*,2}\right)\mathbb{1}\left(i_* \in \mathfrak{B}_k\right)\mathbb{P}\left\{\sigma_{k+1}^{i_*} > \tau_k + b|\mathcal{F}_{\tau_k}\right\} &= \mathbb{1}\left(B_{i_*,2}\right)\mathbb{1}\left(i_* \in \mathfrak{B}_k\right)\overline{G}(b) \\ &= \mathbb{1}\left(B_{i_*,2}\right)\mathbb{1}\left(i_* \in \mathfrak{B}_k\right)\frac{\overline{G}(a^{i_*}(\tau_k) + b)}{\overline{G}(a^{i_*}(\tau_k))}. \end{aligned} \tag{5.A.40}$$

Now we can conclude the lemma, observing that

$$\Omega = \bigcup_{i_*=1}^{N}\left(A_{i_*} \cup B_{i_*}\right), \quad A_{i_*} = A_{i_*,1} \cup A_{i_*,2}, \quad B_{i_*} = B_{i_*,1} \cup B_{i_*,2}, \quad i_* = 1, ..., N. \tag{5.A.41}$$

Independence of ξ_{k+1} and $\sigma_{k+1}^i, i = 1, ..., N$ follows from (5.A.41) and conditional independence results on each partition, i.e. (5.A.24), (5.A.28), (5.A.30), (5.A.38) and (5.A.39). Moreover, the form (5.3.23) for conditional distribution of ξ_{k+1} given $\mathcal{F}_{\tau_k}$ follows from (5.A.41) and this conditional distribution on each partition, i.e. (5.A.25) and (5.A.36). Finally, for every $i \in \mathfrak{B}_k$, the form (5.3.24) for conditional distribution of σ_{k+1}^i given $\mathcal{F}_{\tau_k}$ follows from (5.A.41) and this conditional distribution on each partition, i.e. (5.A.29), (5.A.31), (5.A.32), (5.A.37) and (5.A.40).

□

CHAPTER 6

Characterization of the Limit: Hydrodynamic Equations

In this chapter, we present a system of coupled measure-valued equations. The solution to these equations characterizes the hydrodynamic limit of the Markovian state process of the load-balancing network introduced in Chapter 5, so we call these equations "measure-valued hydrodynamic equations". Despite being measure-valued, this countable system of coupled equations can be partially solved in an explicit fashion, leading to a tractable and intuitive description of the limit. This characterization enables the study of (approximate) transient behavior of large-scale load-balancing networks with general service distributions. We study these equations and prove the uniqueness of the solution to these equations is this chapter; the hydrodynamic limit theorem is established in the next chapter.

The measure-valued Markovian state processes and its corresponding limit contain a lot of information about the state the network. However, one may only be interested in simple performance measures such as queue lengths. Here, we introduce a reduced version of measure-valued hydrodynamic equations, which is a system of coupled partial integro-differential equations (PDEs) that we call hydrodynamic PDEs. As we show in Chapter 7 (Section 7.3), the hydrodynamic limit of performance measures such as queue length distribution and virtual waiting time can be captured by the solution to the hydrodynamic PDEs, thus we provide a direct generalization of the ODE description obtained in [**96, 81**] when the service time distribution is exponential.

First we introduce and study measure-valued hydrodynamic equations in Section 6.1. Then in Section 6.2, we introduce the Hydrodynamic PDEs and show the connection to the measure-valued counterpart (see Section 6.2.2). We study these PDEs under an additional smoothness assumption on the service time distribution (see Assumption IV in Section 6.2), and in particular, we prove the uniqueness of the solution to these PDEs. The extra assumption is not necessary, and only helps us define the solution of the PDEs in a nicer functional space.

6.1. Measure-Valued Hydrodynamic Equations

In this section, we study the measure-valued hydrodynamic equation. First we introduce these equations in Section 6.1.1. Then in Section 6.1.2, we show that the hydrodynamic equations admit at most one solution (Theorem 6.3). Finally in Section 6.1.3, we provide a dynamical characterization of the hydrodynamic equations as solutions to a system of coupled integro-differential equations, which will be used in Section 7.2.3 to establish Theorem 7.1.

6.1.1. Definition of Measure-Valued Hydrodynamic Equations. We now introduce the hydrodynamic equations which provides a "fluid limit" approximation for the network. First, for every $d \geq 2$, define the polynomial

$$\mathfrak{P}_d(x,y) \doteq \frac{x^d - y^d}{x-y} = \sum_{m=0}^{d-1} x^m \, y^{d-1-m}. \tag{6.1.1}$$

Note that for $x, y, \tilde{x}, \tilde{y} \leq 1$, $\mathfrak{P}_d(x,y) \leq d$ and

$$\mathfrak{P}_d(x,y) - \mathfrak{P}_d(\tilde{x},\tilde{y}) \leq d^2\Big((x-\tilde{x}) + (y-\tilde{y})\Big). \tag{6.1.2}$$

Moreover, in the case $d = 2$, we have the simple from $\mathfrak{P}_2(x,y) = x + y$.

DEFINITION 6.1. **(Hydrodynamic Equations)** Given an initial condition $\nu(0) \in \mathbb{S}$ and a non-negative function $\lambda \in \mathbb{L}^1_{\text{loc}}(0,\infty)$, $\{\nu(t) = (\nu_\ell(t); \ell \geq 1); t \geq 0\}$ in $\mathbb{C}_{\mathbb{S}}[0,\infty)$ is said to solve the *hydrodynamic equations* associated with $(\lambda, \nu(0))$ if and only if for every $t \in [0,\infty)$,

$$\int_0^t \langle h, \nu_1(s)\rangle ds < \infty, \tag{6.1.3}$$

and for every $t \geq 0$ and $\ell \geq 1$,

$$\langle \mathbf{1}, \nu_\ell(t)\rangle - \langle \mathbf{1}, \nu_\ell(0)\rangle = D_{\ell+1}(t) + \int_0^t \langle \mathbf{1}, \eta_\ell(s)\rangle ds - D_\ell(t), \tag{6.1.4}$$

and for every $f \in \mathbb{C}_b[0,\infty)$,

$$\langle f, \nu_\ell(t)\rangle = \langle f(\cdot + t)\frac{\overline{G}(\cdot + t)}{\overline{G}(\cdot)}, \nu_\ell(0)\rangle + \int_{[0,t]} f(t-s)\overline{G}(t-s)dD_{\ell+1}(s) + \int_0^t \langle f(\cdot + t - s)\frac{\overline{G}(\cdot + t - s)}{\overline{G}(\cdot)}, \eta_\ell(s)\rangle ds. \tag{6.1.5}$$

where

$$D_\ell(t) \doteq \int_0^t \langle h, \nu_\ell(s)\rangle ds, \qquad \forall \ell \geq 1, \tag{6.1.6}$$

and

$$\eta_\ell(t) \doteq \begin{cases} \lambda(t)\left(1 - \langle \mathbf{1}, \nu_1(t)\rangle^d\right)\delta_0 & \text{if } \ell = 1, \\ \lambda(t)\mathfrak{P}_d\Big(\langle \mathbf{1}, \nu_{\ell-1}(t)\rangle, \langle \mathbf{1}, \nu_\ell(t)\rangle\Big)(\nu_{\ell-1}(t) - \nu_\ell(t)) & \text{if } \ell \geq 2. \end{cases} \tag{6.1.7}$$

For a solution ν of the hydrodynamic equations, we define $S_\ell(t) \doteq \langle \mathbf{1}, \nu_\ell(t)\rangle$, for all $t \geq 0$ and $\ell \geq 1$.

REMARK 6.2. The bound (6.1.3) implies that for every $\ell \geq 1$, the process D_ℓ is well-defined.

To motivate the form of these equations we provide an intuitive description of the N-server dynamics; a more precise characterization of the N-server dynamics is deferred to Section 5.2. Given the age $a(s)$ of a job in service at time s, the conditional probability that the job will complete service in the time interval $(s, s+ds)$ is roughly $h(a(s))ds$. Summing over the ages of all jobs in service at queues of length no less than ℓ, we see that the conditional mean departure rate from such queues at time s is $\langle h, \nu_\ell(s)\rangle$. In the deterministic limit, the departure process coincides with its mean, thus giving rise to the equality in (6.1.6), where D_ℓ is the limit of the departure process form queues with queues of length at least ℓ just before departure.

Next, note that the quantity $\langle \mathbf{1}, \nu_\ell\rangle$ represents the number of jobs receiving service in servers with queues of length no less than ℓ. In the interval $[0,t]$, this quantity decreases only due to the scaled limit of departures caused by service completions at a queue of length ℓ, which is equal to $D_\ell(t) - D_{\ell+1}(t)$, and increases only due to the scaled limit of arrivals of jobs that are routed to a queue that has length $\ell - 1$ prior to the job joining the queue (and causes the job receiving service in that server contribute to $\langle \mathbf{1}, \nu_\ell \rangle$.) The scaled mean arrival rate of jobs to the network at any time s is given by λ (or $\lambda(s)$ in more general time-varying settings) and for $\ell \geq 1$, the probability that an arriving job is routed to a queue of length ℓ at time s is approximately equal to $(\langle \mathbf{1}, \nu_{\ell-1}(s) \rangle)^d - (\langle \mathbf{1}, \nu_\ell(s) \rangle)^d$ (with the convention $\langle \mathbf{1}, \nu_0(s) \rangle \doteq 1$.) Thus, the total arrival rate to a queue of length $\ell - 1$ is given by $\langle \mathbf{1}, \eta_\ell \rangle$, where η_ℓ is defined by (6.1.7). These observations, when combined, justify the form of equation (6.1.4).

We now consider equation (6.1.5). There are three terms that contribute to the mass in the measure $\nu_\ell(t)$ for any given $t > 0$. The first is due to jobs that were already in service at time 0 in a queue of length at least ℓ, and have still not yet completed service by time t. Given that a job had age $a(0)$ at time 0, the conditional probability that it has still not completed service at time t is $\overline{G}(a(0) + t)/\overline{G}(a(0))$, and in that case, its age at time t would be $a(0) + t$. Thus, summing over all possible initial jobs, or equivalently integrating over ν_ℓ, we obtain the first term.

The second term captures service completions at queues of length $k \geq \ell + 1$. Any such queue that has a job that completes service at time s will become of length $k - 1 \geq \ell$ at time s, and thus still be represented in $\nu_\ell(s)$, but now with a delta mass at zero to represent the new job that has just entered service. The second term is then obtained by summing over all such possible job completions at time s, which is represented by $dD_{\ell+1}(s)$, noting that the conditional probability that this job is still in the system at time t is $\overline{G}(t - s)$, and its age at time t will be $t - s$, and integrating s from 0 to t.

The third term represents the contribution due to new arrivals into the system that are routed to a queue of length $\ell - 1$ in the interval $[0, t]$. The conditional probability (given $\nu(s)$) of routing a new arrival to any such queue at time s is $(\langle \mathbf{1}, \nu_{\ell-1}(s) \rangle)^d - (\langle \mathbf{1}, \nu_\ell(s) \rangle)^d / (\langle \mathbf{1}, \nu_{\ell-1} \rangle - \langle \mathbf{1}, \nu_\ell \rangle)$. Moreover, if the job in service at that queue had age a at time s, it now belongs to a queue of length ℓ, and therefore also appears as a new mass at position a in $\nu_\ell^{(N)}(s)$. However, the conditional probability of such a job not having completed service by time t is $\overline{G}(a+t-s)/\overline{G}(a)$, and its age at time t would be $a+t-s$. Summing over all queues of length exactly ℓ at time s, and then integrating s from 0 to t, we obtain the last term in (6.1.5).

Note that $\int_0^t \langle \mathbf{1}, \eta_1(s)ds$ is the number of arrivals to idle servers up to time t and hence $D_2(t) + \int_0^t \langle \mathbf{1}, \eta_1(s)ds\rangle$ is the cumulative service entry process (see Section 5.2). Therefore, for $\ell = 1$, this equation is analogous to the evolution of a single measure-valued process that describes the limit dynamics of a single many-server queue (see (3.11) of **[62]**). [1]

6.1.2. Uniqueness of Solutions to the Hydrodynamic Equations. In this section, we prove that the measure-valued hydrodynamic equations has at most one solution. The existence of a solution follows later in Chapter 7 (see Corollary 7.2.)

THEOREM 6.3. Suppose Assumptions I and II.a hold. Then for every non-negative function $\lambda \in \mathbb{L}^1_{\text{loc}}(0,\infty)$ and $\nu(0) \in \mathbb{S}$, the hydrodynamic equations associated with $(\lambda, \nu(0))$ have at most one solution.

PROOF. Fix a locally integrable function $\lambda(\cdot)$ on $[0,\infty)$ and $\nu(0) \in \mathbb{S}$, and let ν and $\tilde{\nu}$ both be solutions to the hydrodynamic equation associated with $(\lambda, \nu(0))$. For $\ell \geq 1$, define $\Delta H_\ell \doteq H_\ell - \tilde{H}_\ell$ for $H = \nu, D, \eta, S$. Consider the parameterized family of continuous bounded functions

$$\mathbb{F} \doteq \{\vartheta^r; r \geq 0\} \subset \mathbb{C}_b[0,L),$$

where

$$\vartheta^r \doteq \frac{\overline{G}(\cdot + r)}{\overline{G}(\cdot)} \tag{6.1.8}$$

and note that $\mathbf{1} = \vartheta^0 \in \mathbb{F}$. For every $\ell \geq 1$ and $t \geq 0$, define

$$V_\ell(t) = \sup_{f \in \mathbb{F}} |\langle f, \Delta\nu_\ell(t)\rangle|.$$

---

[1] where in this Tez.

By (6.1.7) with $\ell = 1$ and (6.1.1), for $s \geq 0$ and every $f \in \mathbb{C}_b[0, L)$ we have,

$$\begin{aligned} \langle f, \Delta\eta_1(s)\rangle =& \lambda(s)f(0)\left((\tilde{S}_1(s))^d - (S_1(s))^d\right) \\ =& \lambda(s)f(0)\mathfrak{P}_d\left(\tilde{S}_1(s), S_1(s)\right)\langle \mathbf{1}, \Delta\nu_1(s)\rangle. \end{aligned} \tag{6.1.9}$$

Therefore, for every $f \in \mathbb{F}$, since $\mathfrak{P}_d(x, y) \leq d$ and $\|f\|_\infty \leq 1$,

$$|\langle f, \Delta\eta_1(s)\rangle| \leq \lambda(s)d|\langle \mathbf{1}, \Delta\nu_1\rangle| \leq \lambda(s)dV_1(s), \qquad f \in \mathbb{F}. \tag{6.1.10}$$

Next, for every $\ell \geq 2$, $s \geq 0$ and $f \in \mathbb{C}_b[0, L)$,

(6.1.11)

$$\begin{aligned} \langle f, \Delta\eta_\ell(s)\rangle =& \lambda(s)\mathfrak{P}_d\big(S_{\ell-1}(s), S_\ell(s)\big)\langle f, \Delta\nu_{\ell-1}(s) - \Delta\nu_\ell(s)\rangle \\ &+ \lambda(s)\Big(\mathfrak{P}_d\big(S_{\ell-1}(s), S_\ell(s)\big) - \mathfrak{P}_d\big(\tilde{S}_{\ell-1}(s), \tilde{S}_\ell(s)\big)\Big)\langle f, \tilde{\nu}_{\ell-1}(s) - \tilde{\nu}_\ell(s)\rangle. \end{aligned}$$

In particular for every $f \in \mathbb{F}$, using $\|f\|_\infty \leq 1$, $\nu_{\ell-1} - \nu_\ell \geq 0$, $\tilde{\nu}_{\ell-1} - \tilde{\nu}_\ell \geq 0$, $\mathfrak{P}_d(x, y) \leq d$ and property (6.1.2) of $\mathfrak{P}_d$, we have

$$|\langle f, \Delta\eta_\ell(s)\rangle| \leq (d + d^2)\lambda(s)\big(V_{\ell-1}(s) + V_\ell(s)\big), \qquad f \in \mathbb{F}. \tag{6.1.12}$$

Now, first note for $f \in \mathbb{C}_b^1[0, \infty)$ that applying integration by parts to (6.1.5), for every $f \in \mathbb{C}_b^1[0, \infty)$,

$$\begin{aligned} \langle f, \nu_\ell(t)\rangle =& \langle f(\cdot + t)\frac{\overline{G}(\cdot + t)}{\overline{G}(\cdot)}, \nu_\ell(0)\rangle + f(0)\overline{G}(0)D_{\ell+1}(t) \\ &+ \int_{[0,t]}(f\overline{G})'(t - s)D_{\ell+1}(s)ds + \int_0^t \langle f(\cdot + t - s)\frac{\overline{G}(\cdot + t - s)}{\overline{G}(\cdot)}, \eta_\ell(s)\rangle ds. \end{aligned} \tag{6.1.13}$$

Also, note that by definition of ϑ^r, for every $t \geq s \geq 0$ we have

$$\vartheta^r(t - s)\overline{G}(t - s) = \overline{G}(r + t - s),$$

and

$$\vartheta^r(x + t - s)\frac{\overline{G}(x + t - s)}{\overline{G}(x)} = \frac{\overline{G}(r + x + t - s)}{\overline{G}(x)} = \vartheta^{r+t-s}(x), \qquad x \in [0, L).$$

Thus, substituting $f = \vartheta^r$ in both (6.1.13) and in (6.1.13) with $\nu_\ell, D_\ell, \eta_\ell$ replaced by $\tilde{\nu}_\ell, \tilde{D}, \tilde{\eta}_\ell$, respectively, using the last two relations and the facts that $\vartheta^r(0) = \overline{G}(r)$ and $\overline{(\vartheta^r)}'(t-s) = -g(r+t-s)$, we have

(6.1.14)

$$\langle \vartheta^r, \Delta\nu_\ell(t)\rangle = \overline{G}(r)\Delta D_{\ell+1}(t) - \int_0^t g(r+t-s)\Delta D_{\ell+1}(s)ds + \int_0^t \langle \vartheta^{t+r-s}, \Delta\eta_\ell(s)\rangle ds.$$

Since $\vartheta^0 = \mathbf{1}$, equation (6.1.14) for $r = 0$ gives

$$\langle \mathbf{1}, \Delta\nu_\ell(t)\rangle = \Delta D_{\ell+1}(t) - \int_0^t g(t-s)\Delta D_{\ell+1}(s)ds + \int_0^t \langle \vartheta^{t-s}, \Delta\eta_\ell(s)\rangle ds. \tag{6.1.15}$$

On the other hand, using the fact that $(\nu_\ell\ ; \ell \geq 1)$ satisfies (6.1.4) and $(\tilde{\nu}_\ell\ ; \ell \geq 1)$ satisfies the analogous equation and that $\Delta\nu_\ell(0) = 0$, we also have

$$\Delta D_{\ell+1}(t) = \langle \mathbf{1}, \Delta\nu_\ell(t)\rangle + \Delta D_\ell(t) - \int_0^t \langle \mathbf{1}, \Delta\eta_\ell(s)\rangle ds. \tag{6.1.16}$$

Substituting $\Delta D_{\ell+1}(t)$ from (6.1.16) into (6.1.15), it follows that for every $\ell \geq 2$, ΔD_ℓ satisfies the renewal equation

$$\Delta D_\ell(t) = g * \Delta D_\ell(t) + F_\ell(t), \tag{6.1.17}$$

with

$$F_\ell(t) \doteq \int_0^t \langle \mathbf{1}, \Delta\eta_\ell(s)\rangle ds + g * \langle \mathbf{1}, \Delta\nu_\ell\rangle(t) - (g * \int_0^\cdot \langle \mathbf{1}, \Delta\eta_\ell(s)\rangle ds)(t)$$
$$- \int_0^t \langle \vartheta^{t-s}, \Delta\eta_\ell(s)\rangle ds.$$

Since $\Delta\nu_\ell$ is the difference of two measures in $\mathbb{D}_{\mathbb{M}_F[0,L)}[0,\infty)$ and λ is locally integrable, $\langle \mathbf{1}, \Delta\nu_\ell\rangle$ and $\langle f, \Delta\eta_\ell\rangle$ for $f \in \mathbb{C}_b[0,\infty)$ are also locally integrable, and hence F_ℓ is uniformly bounded on finite intervals (i.e., $\|F\|_t < \infty$ for all $t \geq 0$). ΔD_ℓ is also bounded on finite intervals because of the bound (6.1.3) and therefore, by Theorem 2.4 of Chapter V in [**3**], ΔD_ℓ has the representation

$$\Delta D_\ell(t) = F_\ell(t) + u * F_\ell(t),$$

where u is the renewal density on $[0, L)$. Note that u exists because G has density g, see Proposition 2.7 of Chapter V in [**3**], which also shows that u is uniformly bounded on every finite interval and satisfies the equation $u = g * u + g$. Substituting the definition of F_ℓ into equation (6.1.17) and using the equality $u * g = u - g$, we have

$$\begin{aligned} \Delta D_\ell(t) = & \int_0^t \langle \mathbf{1}, \Delta\eta_\ell(s)\rangle ds + u * \langle \mathbf{1}, \Delta\nu_\ell\rangle(t) - \int_0^t \langle \vartheta^{t-s}, \Delta\eta_\ell(s)\rangle ds \\ & + \left(u * \int_0^\cdot \langle \vartheta^{\cdot - s}, \Delta\eta_\ell(s)\rangle ds \right)(t). \end{aligned} \tag{6.1.18}$$

Now we bound each term in the term in the right-hand side of equation (6.1.18). Fix $T \geq 0$. The bound in (6.1.12) implies

$$\left| \int_0^t \langle \mathbf{1}, \Delta\eta_\ell(s)\rangle ds \right| \leq (d + d^2) \int_0^t \lambda(s) \left(V_{\ell-1}(s) + V_\ell(s) \right) ds, \quad t \leq T. \tag{6.1.19}$$

Moreover, recall that for a function f, $\|f\|_T = \sup_{t\in[0,T]} |f(t)|$, we also have

$$|u * \langle \mathbf{1}, \Delta\nu_\ell\rangle(t)| \leq \int_0^t u(t-s) \left| \langle \mathbf{1}, \Delta\nu_\ell(s)\rangle \right| ds \leq \|u\|_T \int_0^t V_\ell(s) ds, \quad t \leq T. \tag{6.1.20}$$

Furthermore, for the function $\zeta_\ell(t) \doteq \int_0^t \langle \vartheta^{t-s}, \Delta\eta_\ell(s)\rangle ds$, the bound (6.1.12) with $f = \vartheta^{t-s}$ implies

$$|\zeta_\ell(s)| \leq (d + d^2) \int_0^s \lambda(v) \left(V_{\ell-1}(v) + V_\ell(v) \right) dv, \qquad s \geq 0, \tag{6.1.21}$$

and hence, applying Tonelli's theorem in the third inequality below, we obtain

$$\begin{aligned} |u * \zeta_\ell(t)| \leq & \int_0^t u(t-s) |\zeta_\ell(s)| ds \\ \leq & (d + d^2) \int_0^t \int_0^s u(t-s) \lambda(v) \left(V_{\ell-1}(v) + V_\ell(v) \right) dv \, ds \\ \leq & (d + d^2) \int_0^t \lambda(s) \left(V_{\ell-1}(s) + V_\ell(s) \right) \left(\int_s^t u(t-v) dv \right) ds \\ \leq & (d + d^2) U(T) \int_0^t \lambda(s) \left(V_{\ell-1}(s) + V_\ell(s) \right) ds, \qquad \forall t \leq T, \end{aligned} \tag{6.1.22}$$

where $U(\cdot) = 1 + \int_0^\cdot u(s) ds$ is the renewal function associated with G. From equation (6.1.18) and the bounds (6.1.19)-(6.1.22), for $\ell \geq 2$ we then have

$$\|\Delta D_\ell\|_t \leq C(T) \int_0^t (1 + \lambda(s)) \left(V_{\ell-1}(s) + V_\ell(s) \right) ds, \quad \forall t \leq T,$$

with $C(T) \doteq \|u\|_T + 2(d+d^2)(1+U(T)) < \infty$, or equivalently, for $\ell \geq 1$,

$$\|\Delta D_{\ell+1}\|_t \leq C(T) \int_0^t (1+\lambda(s)) \left(V_\ell(s) + V_{\ell+1}(s)\right) ds, \quad \forall t \leq T. \tag{6.1.23}$$

Finally, incorporating the bounds (6.1.10), (6.1.12) and (6.1.23) in the equation (6.1.14), we have

$$|\langle f, \Delta\nu_1(t)\rangle| \leq 3C(T) \int_0^t (1+\lambda(s))(V_1(s)+V_2(s))ds, \qquad \forall f \in \mathbb{F},$$

and for every $\ell \geq 2$,

$$|\langle f, \Delta\nu_\ell(t)\rangle| \leq 3C(T) \int_0^t (1+\lambda(s))(V_{\ell-1}(s)+V_\ell(s)+V_{\ell+1}(s))ds, \qquad \forall f \in \mathbb{F}.$$

Taking the supremum over all $f \in \mathbb{F}$ on the left-hand sides of the last two equations, for all $t \leq T$ we have

$$V_\ell(t) \leq \begin{cases} 3C(T) \int_0^t (1+\lambda(s))(V_1(s)+V_2(s))ds, & \text{if } \ell = 1, \\ 3C(T) \int_0^t (1+\lambda(s))(V_{\ell-1}(s)+V_\ell(s)+V_{\ell+1}(s))ds, & \text{if } \ell \geq 2. \end{cases} \tag{6.1.24}$$

Define $V(t) \doteq \sum_{\ell=1}^\infty 2^{-\ell} V_\ell(t)$. Since for every $f \in \mathbb{F}$, $|\langle f, \nu_\ell(t)\rangle| \leq 1$, $V_\ell(t) \leq 2$ for all $\ell \geq 1$, and hence $V(t) \leq 2 < \infty$ for all $t \geq 0$. Also multiplying both sides of the bound (6.1.24) by $2^{-\ell}$ and summing up over $\ell \geq 1$, we obtain

$$V(t) \leq 12C(T) \int_0^t (1+\lambda(s))V(s)ds. \tag{6.1.25}$$

Since $V(0) = 0$, an application of Gronwall's inequality then shows that $V(t) = 0$ for all $t \geq 0$, and thus $V_\ell(t) = 0$ for all $t \geq 0$ and all $\ell \geq 1$. In particular, this shows that

$$\langle \mathbf{1}, \nu_k \rangle = 0, \qquad k \geq 1. \tag{6.1.26}$$

Moreover, by inequality (6.1.23), $\Delta D_\ell \equiv 0$ for all $\ell \geq 2$. Substituting $\Delta_{\ell+1} = 0$ in (6.1.13) and using the fact that $\Delta\nu_\ell(0) = 0$, we see that for every $f \in \mathbb{C}_b^1[0,\infty)$,

For $\ell \geq 1$, we can then rewrite (6.1.14) as

$$\langle f, \Delta\nu_\ell(t)\rangle = \int_0^t \langle f(\cdot + t - s)\vartheta^{t-s}(\cdot), \Delta\eta_\ell(s)\rangle ds. \tag{6.1.27}$$

We finish the proof by using induction on ℓ to show that $\Delta\nu_\ell \equiv 0$ for all $\ell \geq 1$. We first note that $f(\cdot + t - s)\vartheta^{t-s}(\cdot) \in \mathbb{C}_b[0,L)$ for all $f \in \mathbb{C}_b[0,L)$ and $t, s \geq 0$. For $\ell =$

1, (6.1.9) and the equality (6.1.26) with $k = 1$ imply $\langle f(\cdot + t - s)\vartheta^{t-s}, \Delta\eta_1(s)\rangle = 0$ for all $t, s \geq 0$, and therefore, $\Delta\nu_1 \equiv 0$ using (6.1.27). Furthermore, if $\Delta\nu_{\ell-1} \equiv 0$ for some $\ell \geq 2$, it follows from (6.1.11), (6.1.2) and the equality (6.1.26) for $k = \ell - 1$ and $k = \ell$ that

$$
\begin{aligned}
&\left|\langle f(\cdot + t - s)\vartheta^{t-s}(\cdot), \Delta\eta_\ell(s)\rangle\right| \\
&\qquad\qquad \leq \lambda(s)\mathfrak{P}_d\big(\langle \mathbf{1}, \nu_{\ell-1}(s)\rangle, \langle \mathbf{1}, \nu_\ell(s)\rangle\big)\langle f(\cdot + t - s)\vartheta^{t-s}(\cdot), \Delta\nu_\ell(s)\rangle \\
&\qquad\qquad \leq \lambda(s)\|f\|_\infty \|\Delta\nu(s)\|_{TV}\, d,
\end{aligned}
$$

where the last inequality uses $\mathfrak{P}_d(x, y) \leq d$ and $\|f(\cdot + t - s)\vartheta^{t-s}(\cdot)\|_\infty \leq \|f\|_\infty$. Incorporating this bound into (6.1.27) and taking the supremum over all $f \in \mathbb{C}_b[0, L)$ with $\|f\|_\infty \leq 1$, we obtain

$$
\|\Delta\nu_\ell(t)\|_{TV} \leq d \int_0^t \lambda(s)\|\Delta\nu_\ell(s)\|_{TV} ds, \qquad \forall \geq 0. \tag{6.1.28}
$$

Since $\Delta\nu_\ell(0) = 0$, another application of Gronwall's inequality shows $\Delta\nu_\ell(t) = 0$ for all $t \geq 0$. $\square$

.

6.1.3. A Dynamical Characterization of the Hydrodynamic Limit. In this section we provide an alternative, dynamical characterization of solutions to the hydrodynamic equations, which will be used in Section 7.2.3 to establish convergence of $\nu^{(N)}$ as $N \to \infty$. The idea is that although the hydrodynamic equations are highly non-linear, the dependence of ν_ℓ on $D_{\ell+1}$ and η_ℓ is linear, and we will show that the ν_ℓ in (6.1.5) is simply the unique solution to a linear (measure-valued) PDE with inputs $D_{\ell+1}$ and η_ℓ. A similar representation was used in **[61]** and **[62]**, and indeed, our result below generalizes Theorem 4.1 of **[61]**. Here is the main result of this section.

PROPOSITION 6.4. Suppose $\nu(0) = (\nu_\ell(0); \ell \geq 1) \in \mathbb{S}$ and nonnegative function $\lambda(\cdot) \in \mathbb{L}^1_{\text{loc}}[0, \infty)$ are given and a process $\nu = (\nu_\ell)_{\ell \geq 1} \in \mathbb{D}_\mathbb{S}[0, \infty)$, satisfies the following: for every $\ell \geq 1$ and $t \geq 0$, the bound

$$D_\ell(t) = \int_0^t \langle h, \nu_\ell(s)\rangle ds < \infty, \tag{6.1.29}$$

and the equation

$$\langle \mathbf{1}, \nu_\ell(t)\rangle - \langle \mathbf{1}, \nu_\ell(0)\rangle = D_{\ell+1}(t) + \int_0^t \langle \mathbf{1}, \eta_\ell(s)\rangle ds - D_\ell(s), \tag{6.1.30}$$

hold, and for every $\varphi \in \mathbb{C}_c^{1,1}([0,L) \times \mathbb{R}_+)$, the equation

$$\begin{aligned}\langle \varphi(\cdot,t), \nu_\ell(t)\rangle =& \langle \varphi(\cdot,0), \nu_\ell(0)\rangle + \int_0^t \langle \varphi_x(\cdot,s) + \varphi_s(\cdot,s) - \varphi(\cdot,s)h(\cdot), \nu_\ell(s)\rangle ds \\ &+ \int_0^t \varphi(0,s) dD_{\ell+1}(s) + \int_0^t \langle \varphi(\cdot,s), \eta_\ell(s)\rangle ds,\end{aligned} \tag{6.1.31}$$

also holds where η_ℓ is defined in (6.1.7). Then, ν is a solution to the Hydrodynamic Equations associated to $(\lambda, \nu(0))$.

Throughout this section, we fix the non-negative $\lambda(\cdot) \in \mathbb{L}^1_{\text{loc}}$ in Proposition 6.4. We also denote by $\tilde{\mathbb{M}}$ the space of Radon measures on $\mathbb{R}^2$ whose supports lie in $[0,L) \times \mathbb{R}_+$, and the integral with respect to any Radon measure Θ on $\mathbb{R}^2$ is denoted by

$$\Theta(\varphi) = \int\int_{\mathbb{R}^2} \varphi(x,s)\Theta(dx,ds).$$

DEFINITION 6.5. **(Abstract Age Equation)** Suppose a Radon measure $\Theta \in \tilde{\mathbb{M}}$ is given. A measure-valued function $\mu = \{\mu(t); t \geq 0\}$ in $\mathbb{D}_{\mathbb{M}_F[0,L)}[0,\infty)$ is said to solve the *abstract age equation* associated with Θ if and only if for every $t \in [0,\infty)$,

$$\int_0^t \langle h, \mu(s)\rangle ds < \infty, \tag{6.1.32}$$

and for every $\varphi \in \mathbb{C}_c^{1,1}([0,L) \times \mathbb{R}_+)$,

$$-\int_0^\infty \langle \varphi_x(\cdot,s) + \varphi_s(\cdot,s) - \varphi(\cdot,s)h(\cdot), \mu(s)\rangle ds = \Theta(\varphi). \tag{6.1.33}$$

The following Lemma gives an explicit solution to the abstract age equation.

LEMMA 6.6. Suppose $\mu = \{\mu(t); t \geq 0\}$ in $\mathbb{D}_{\mathbb{M}_F[0,L)}[0,\infty)$ solves the abstract age equation associated with some $\Theta \in \tilde{\mathbb{M}}$. Then for every $f \in \mathbb{C}_c[0,L)$ and $t \geq 0$,

$$\langle f, \mu(t)\rangle = \int_0^t \int_{[0,L)} f(x+t-s)\frac{\overline{G}(x+t-s)}{\overline{G}(x)}\Theta(dx,ds). \tag{6.1.34}$$

PROOF. Define the measure the measure $h\mu$ on $\mathbb{R}^2$ as

$$(h\mu)(\varphi) = \int_0^\infty \langle \varphi(\cdot,s)h(\cdot), \mu(s)\rangle ds, \qquad \varphi \in \mathbb{C}_c(\mathbb{R}^2_+).$$

Since $\mu(s) \in \mathbb{M}_F[0,L)$ for every $s \geq 0$, we have $(h\mu)(\varphi) = 0$ for all φ with support outside of $[0,L) \times \mathbb{R}_+$. Also, for every $\varphi \in \mathbb{C}_c(\mathbb{R}^2_+)$, using the bound (6.1.32), we have

$$|(h\mu)(\varphi)| \leq \|\varphi\|_\infty \int_0^\infty \langle h, \mu(s)\rangle ds < \infty.$$

Therefore, $h\mu$ is a finite Radon measure on $\mathbb{R}^2$ with support in $[0,L) \times \mathbb{R}_+$, that is $h\nu \in \tilde{\mathbb{M}}$. Therefore, by Corollary 4.17 in **[61]**, for every $t \geq 0$ and $f \in \mathbb{C}_c[0,L)$,

$$\langle f, \mu(t)\rangle = \Theta\big(\psi_{-h}\Lambda^t_{f(\cdot)\psi_h(\cdot,t)}\big), \tag{6.1.35}$$

where

$$\psi_h(x,t) = \begin{cases} \frac{\overline{G}(x)}{\overline{G}(x-t)} & \text{if } 0 \leq t \leq x < M, \\ \overline{G}(x) & \text{if } 0 \leq x \leq t < \infty, \\ 0 & \text{otherwise.} \end{cases} \tag{6.1.36}$$

and for every $t \geq 0$, $\Lambda^t : \mathbb{C}_c(\mathbb{R}) \mapsto \mathbb{C}_c(R^2)$ is defined by

$$\Lambda^t_f(x,s) = f(x+t-s)\mathbf{1}_{[0,t]}(s). \tag{6.1.37}$$

Combining definitions (6.1.36)-(6.1.37), we have

$$\psi_{-h}\Lambda^t_{f(\cdot)\psi_h(\cdot,t)}(x,s) = f(x+t-s)\frac{1-G(x+t-s)}{1-G(x)}\mathbf{1}_{[0,t]}(s),$$

and hence, (6.1.34) follows from substituting $\psi_{-h}\Lambda^t_{f(\cdot)\psi_h(\cdot,t)}$ from the equation above in (6.1.35). $\square$

PROOF OF PROPOSITION 6.4. Recall the definition (6.1) of the Hydrodynamic equations. Since the equations (6.1.3) and (6.1.4) are readily given in (6.1.29) and (6.1.30), respectively, we only need to show that (6.1.5) holds for all $t \geq 0$, $\ell \geq 1$ and $f \in \mathbb{C}_b[0, L)$.

Fix $\ell \geq 1$, and consider the linear functional ξ_ℓ on $\mathbb{C}_c^{1,1}([0, L) \times \mathbb{R}_+)$, where for every $\varphi \in \mathbb{C}_c^{1,1}([0, L) \times \mathbb{R}_+)$, $\xi_\ell(\varphi)$ is defined by

$$\xi_\ell(\varphi) \doteq \langle \varphi(\cdot, 0), \nu_\ell(0) \rangle + \int_{[0,\infty)} \varphi(0, s) dD_{\ell+1}(s) + \int_0^\infty \langle \varphi(\cdot, s), \eta_\ell(s) \rangle ds. \tag{6.1.38}$$

By definition (6.1.7) of η_ℓ, for all $m \in [0, L)$ and $T \in [0, \infty)$ and every $\varphi \in \mathbb{C}_c(R^2)$ with $\text{supp}(\varphi) \subset [0, m] \times [0, T]$,

$$\big| \langle \varphi(\cdot, s), \eta_\ell(s) \rangle \big| \leq \|\varphi\|_\infty \lambda(s) d, \tag{6.1.39}$$

and hence, since $D_{\ell+1}$ is non-decreasing,

$$|\xi_\ell(\varphi)| \leq \|\varphi\|_\infty \big(|\nu_\ell(0)|_{TV} + D_{\ell+1}(T) + C(T) \big), \tag{6.1.40}$$

with $C(T) \doteq d \int_0^T \lambda(s) ds < \infty$ by local integrability of λ and $D_{\ell+1}(T) < \infty$ by (6.1.29). Moreover, $\xi_\ell(\varphi)$=0 for all φ such that $\text{supp}(\varphi) \cap [0, L) \times \mathbb{R}_+ = \emptyset$. Therefor, ξ is a Radon measure on $\mathbb{R}^2$ with support in $[0, L) \times \mathbb{R}_+$, i.e. $\xi \in \tilde{\mathbb{M}}$. Moreover, since any $\varphi \in \mathbb{C}_c^{1,1}([0, L) \times \mathbb{R}_+)$, has compact support, the left-hand side of (6.1.31) equals to 0 for sufficiently large t. Therefore, sending $t \to \infty$ in (6.1.31), we have

$$-\int_0^t \langle \varphi_x(\cdot, s) + \varphi_s(\cdot, s) - \varphi(\cdot, s) h(\cdot), \nu_\ell(s) \rangle ds = \xi_\ell(\varphi). \tag{6.1.41}$$

Hence, ν_ℓ satisfies the abstract age equation associated with $\xi_\ell \in \tilde{\mathbb{M}}$, and therefore by Lemma 6.6 and definition (6.1.38) of ξ_ℓ, for all $f \in \mathbb{C}_c[0, L)$ and $t \geq 0$,

$$\begin{aligned}
\langle f, \nu_\ell(t) \rangle = & \int_0^t \int_{[0,L)} f(x + t - s) \frac{\overline{G}(x + t - s)}{\overline{G}(x)} \xi_\ell(dx, ds) \\
= & \langle f(\cdot + t) \frac{\overline{G}(\cdot + t)}{\overline{G}(\cdot)}, \nu_\ell(0) \rangle + \int_{[0,t]} f(t - s) \overline{G}(t - s) dD_{\ell+1}(s) \\
& + \int_0^t \langle f(\cdot + t - s) \frac{\overline{G}(\cdot + t - s)}{\overline{G}(\cdot)}, \eta_\ell(s) \rangle ds.
\end{aligned} \tag{6.1.42}$$

Since for ever $t \geq 0$, the right-hand side above is finite for every $f \in \mathbb{C}_b[0, L)$, (6.1.42) can be extended for all $f \in \mathbb{C}_b[0, L)$ by an application of dominated convergence theorem, and (6.1.5) follows. This completes the proof. $\square$

6.2. Hydrodynamic PDEs

In this section, we introduce and study the Hydrodynamic PDEs. The equations are first introduced in Section 6.2.1. Then in Section 6.2.2, we show that under an additional smoothness assumption on the service time distribution (see Assumption IV below), the solution to the these PDEs is closely related to the solution to the measure-valued hydrodynamic equations defined in Section 6.1.1. Finally, we prove the uniqueness of the solution to the hydrodynamic PDEs in Section 6.2.3.

The following extra smoothness assumption on the service time distribution helps us define the solution of the PDEs in a nicer functional space.

ASSUMPTION IV. (Service Distribution) The hazard rate h is continuous and bounded, i.e. $h \in \mathbb{C}^1_{\leq 1}[0, \infty)$. In particular,

$$H \doteq \sup_{x \in [0,L)} h(x) < \infty.$$

6.2.1. Description of the Hydrodynamic PDE. Define $\mathbb{C}^1_{\leq 1}[0, \infty)$ to be the space of continuously differentiable functions f with bounded derivative, such that $|f|$ is bounded by 1. Also, define $\mathbb{X}$ to be the set of functions φ on $[0, \infty) \times [0, \infty)$ such that for every $t \geq 0$, $\varphi(t, \cdot) \in \mathbb{C}^1_{\leq 1}[0, \infty)$ with $\partial_r \varphi(t, \cdot)$ denoting the derivative with respect to the second variable, and for every $r \geq 0$, the mapping $t \mapsto \varphi(t, r)$ is measurable. Now we define the PDE.

DEFINITION 6.7. For every function λ and set of functions $Z^0_\ell \in \mathbb{C}^1_{\leq 1}[0, \infty)$, $\ell \geq 1$, a set of functions $(Z_\ell; \ell \geq 1)$ with $Z_\ell \in \mathbb{X}$ for all $\ell \geq 1$ is said to solve the hydrodynamic PDE associated to $(\lambda, Z^0_\ell; \ell \geq 1)$ if for every $t, r \geq 0$, Z_1 satisfies

$$Z_1(t,r) = Z_1^0(t+r) - \int_0^t \overline{G}(t+r-u)\partial_r Z_2(u,0)du + \int_0^t \lambda(u)\overline{G}(t+r-u)(1-Z_1(u,0)^d)du, \tag{6.2.1}$$

with the boundary condition

$$Z_1(t,0) = Z_1^0(0) + \int_0^t \left(\partial_r Z_1(u,0) - \partial_r Z_2(u,0)\right)du + \int_0^t \lambda(u)\left(1 - Z_1(u,0)^d\right)du, \tag{6.2.2}$$

and for $\ell \geq 2$, Z_ℓ satisfies

$$\begin{aligned} Z_\ell(t,r) = {} & Z_\ell^0(t+r) - \int_0^t \overline{G}(t+r-u)\partial_r Z_{\ell+1}(u,0)du \\ & + \int_0^t \lambda(u)\mathfrak{P}_d\left(Z_{\ell-1}(u,0), Z_\ell(u,0)\right)\left(Z_{\ell-1}(u,t+r-u) - Z_\ell(u,t+r-u)\right)du, \end{aligned} \tag{6.2.3}$$

with the boundary conditions

$$Z_\ell(t,0) = Z_\ell^0(0) + \int_0^t \left(\partial_r Z_\ell(u,0) - \partial_r Z_{\ell+1}(u,0)\right)du + \int_0^t \lambda(u)\left(Z_{\ell-1}(u,0)^d - Z_\ell(u,0)^d\right)du. \tag{6.2.4}$$

Note that by equations (6.2.1) and (6.2.3), Z^0_ℓ is the initial condition for Z_ℓ, that is $Z_\ell(0,r) = Z^0_\ell(r)$, for all $\ell \geq 1$ and $r \geq 0$.

REMARK 6.8. Although the equations (6.2.1)-(6.2.4) are partial integro-differential equations and not partial differential equations in the classic sense, we still refer to them as hydrodynamic PDE for conciseness.

6.2.2. Connection to the Measure-Valued Equation. Here, we establish the connection between the solution of the hydrodynamic PDEs and their measure-valued counterpart defined in Section 6.1.1.

Let λ be a non-negative locally integrable function, and fix any $\nu(0) = (\nu_\ell(0); \ell \geq 0) \in \mathbb{S}$, and suppose $\nu = \{(\nu_\ell(t); \ell \geq 0); t \geq 0\}$ satisfies the measure-valued hydrodynamic equations (6.1.4)-(6.1.7) associated with $(\lambda, \nu(0))$. For every $\ell \in N$, we define the function $Z_\ell : [0,\infty) \times [0,\infty) \mapsto \mathbb{R}$ as

$$Z_\ell(t,r) \doteq \langle \vartheta^r, \nu_\ell(t)\rangle, \qquad \forall t,r \geq 0. \tag{6.2.5}$$

We first show in lemma 6.9 below that the functions Z_ℓ defined above lie in $\mathbb{X}$, and in particular, the functions Z_ℓ^0 defined as

$$Z_\ell^0 \doteq \langle \vartheta^r, \nu_\ell(0)\rangle, \qquad r \geq 0. \tag{6.2.6}$$

lie in $\mathbb{C}^1_{\leq 1}[0,\infty)$ for all $\ell \geq 1$. Then we show in Proposition 6.10 that $Z = (Z_\ell; \ell \geq 1)$ satisfies the hydrodynamic PDEs associated to (λ, Z^0).

LEMMA 6.9. Suppose Assumptions I, II and IV hold. For every $\ell \geq 1$, $Z_\ell \in \mathbb{X}$, and for every $t \geq 0$,

$$\langle h, \nu_\ell(t)\rangle = -\partial_r Z_\ell(t,0). \tag{6.2.7}$$

PROOF. Fix $\ell \geq 1$. For every $x, r \geq 0$, and $\delta > 0$, by mean value theorem

$$\frac{\overline{G}(x+r) - \overline{G}(x+r+\delta)}{\delta\, \overline{G}(x)} = \frac{g(x+r+\delta^*(x,r))}{\overline{G}(x)} \leq h(x+r+\delta^*(x,r)) \leq \|h\|_\infty, \tag{6.2.8}$$

for some $\delta^*(x,r) \in [0,\delta]$, and $\|h\|_\infty$ is finite due to Assumption IV. Therefore,

$$\begin{aligned}
\lim_{\delta\to 0} \frac{1}{\delta}\left(Z_\ell(t,r+\delta) - Z_\ell(t,r)\right) &= \lim_{\delta\to 0} \Big\langle \frac{\overline{G}(\cdot+r+\delta) - \overline{G}(\cdot+r)}{\delta\, \overline{G}(\cdot)}, \nu_\ell(t)\Big\rangle \\
&= \Big\langle \lim_{\delta\to 0} \frac{\overline{G}(\cdot+r+\delta) - \overline{G}(\cdot+r)}{\delta\, \overline{G}(\cdot)}, \nu_\ell(t)\Big\rangle \\
&= -\Big\langle \frac{g(\cdot+r)}{\overline{G}(\cdot)}, \nu_\ell(t)\Big\rangle. \qquad (6.2.9)
\end{aligned}$$

The second equality is by (6.2.8) and an application of bounded convergence theorem. Continuity of $\partial_r Z_\ell(t,r)$ follows similarly by another application of bounded convergence theorem. Therefore, for every $t \geq 0$, $Z_\ell(t,\cdot) \in \mathbb{C}^1_{\leq 1}[0,\infty)$.

Also, for every $r \geq 0$, the function $\overline{G}(\cdot+r)/\overline{G}(\cdot)$ is bounded and continuous, and hence, the mapping $T : \mathbb{M}_F[0,\infty) \mapsto \mathbb{R}$ which takes a measure μ to $\langle \overline{G}(\cdot + r)/\overline{G}(\cdot), \mu\rangle$ is continuous. Together with the fact that the process $\{\nu_\ell(t); t \geq 0\}$ is right-continuous, if follows that for every $r \geq 0$, the mapping $t \mapsto Z_\ell(t,r)$ is right-continuous and hence, measurable. Therefore, $Z_\ell \in \mathbb{X}$. Finally, (6.2.7) holds by substituting $r = 0$ in (6.2.9). $\square$

Lemma 6.9 in particular shows that the initial conditions $Z^0_\ell(\cdot) = Z_\ell(0, \cdot)$ lie in $\mathbb{C}^1_{\leq 1}[0, \infty)$ for all $\ell \geq 1$.

PROPOSITION 6.10. Suppose Assumptions I and IV hold, and let ν be a solution to the measure-valued hydrodynamic equations associated to $(\lambda, \nu(0))$ for some non-negative function λ and some $\nu(0) \in \mathbb{S}$. Then functions $Z = (Z_\ell; \ell \geq 1)$ defined in (6.2.5) satisfies the hydrodynamic PDEs (6.2.1)-(6.2.4) associated to $(\lambda, \tilde{Z}_\ell; \ell \geq 0)$, with Z^0_ℓ defined in (6.2.6).

PROOF. Recall the definition (6.1.8) of ϑ^r for $r \geq 0$. For every $r \geq 0$, $0 \leq u \leq t$ and $x \geq 0$

$$\vartheta^r(t-u)\overline{G}(t-u) = \overline{G}(t+r-u) \tag{6.2.10}$$

and

$$\vartheta^r(x+u)\frac{\overline{G}(x+u)}{\overline{G}(x)} = \vartheta^{r+u}(x). \tag{6.2.11}$$

Also, substituting $\langle h, \nu_\ell(t)\rangle$ from (6.2.7) in definition (6.1.6) of $D_\ell(t)$, we have

$$D_\ell(t) = -\int_0^t \partial_r Z_\ell(u, 0)du. \tag{6.2.12}$$

Equations (6.2.1)-(6.2.4) follow from substituting the function $f = \vartheta^r$ and the relations in equations (6.2.10), (6.2.11) and (6.2.12) in measure-valued hydrodynamic equation (6.1.4)-(6.1.7). $\square$

6.2.3. Uniqueness of the Solution to Hydrodynamic PDEs. In this section, we show that the hydrodynamic PDEs defined in Section 6.2.1 has at most one solution. We present the proof for the case $d = 2$; the generalization to arbitrary $d \geq 1$ is straightforward.

THEOREM 6.11. Suppose Assumptions I, II and IV hold. Then, for every non-negative function $\lambda \in \mathbb{L}^1_{\text{loc}}(0,\infty)$ and $Z^0_\ell \in \mathbb{C}^1_{\leq 1}[0,\infty), \ell \geq 1$, then the hydrodynamic PDEs (6.2.1)-(6.2.4) with $d = 2$ has at most one solution.

PROOF. Fix a non-negative function $\lambda \in \mathbb{L}^1_{\text{loc}}[0,\infty)$ and $Z^0 = (Z^0_\ell; \ell \geq 1)$, and let $Z = (Z_\ell; \ell \geq 1)$ and $\widetilde{Z} = (\widetilde{Z}_\ell; \ell \geq 1)$ both be solutions to the hydrodynamic PDE (6.2.1)-(6.2.4) associated to (λ, Z^0). Defining the functions

$$D_\ell(t) \doteq -\int_0^t \partial_r Z_\ell(s,0)ds, \qquad t \geq 0,$$

and using integration by parts, we can rewrite

(6.2.13)

$$\int_0^t \overline{G}(t+r-s)\partial_r Z_{\ell+1}(s,0)ds = -\overline{G}(r)D_{\ell+1}(t) + \int_0^t D_{\ell+1}(s)g(t+r-s)ds.$$

Substituting (6.2.13) into (6.2.1) and (6.2.3), and the definition of D_ℓ into (6.2.2) and (6.2.4), we see that for every $\ell \geq 1$, Z_ℓ satisfies for $t, r \geq 0$,

$$Z_\ell(t,r) = Z^0_\ell(r+t) + \overline{G}(r)D_{\ell+1}(t) + R_\ell(t,r) - \int_0^t D_{\ell+1}(s)g(t+r-s)ds$$

with boundary condition

$$Z_\ell(t,0) = Z^0_\ell(0) - D_\ell(t) + D_{\ell+1}(t) + \Lambda_\ell(t),$$

where

$$R_1(t) = \int_0^t \lambda(s)\overline{G}(t+r-s)\big(1 - Z_1^2(s,0)\big)ds, \tag{6.2.14}$$

and, for $\ell \geq 2$,

(6.2.15)

$$R_\ell(t) \int_0^t \lambda(s)\big(Z_{\ell-1}(s,0) - Z_\ell(s,0)\big)\big(Z_{\ell-1}(s,t+r-s) - Z_\ell(s,t+r-s)\big)ds,$$

and

$$\Lambda_\ell(t) \doteq \begin{cases} \int_0^t \lambda(s)\big(1 - Z_1^2(s,0)\big)ds & \text{if } \ell = 1, \\ \int_0^t \lambda(s)\big(Z^2_{\ell-1}(s,0) - Z^2_\ell(s,0)\big)ds & \text{if } \ell \geq 2. \end{cases} \tag{6.2.16}$$

Similarly, $\widetilde{Z}$ satisfies analogous equations. Defining $\Delta H_\ell \doteq H_\ell - \widetilde{H}_\ell$ for $H = Z, D, R, \Lambda$ and $\ell \geq 1$, for all $\ell \geq 1$ and $t, r \geq 0$ we have

$$\Delta Z_\ell(t,r) = \overline{G}(r)\Delta D_{\ell+1}(t) + \Delta R_\ell(t,r) - \int_0^t \Delta D_{\ell+1}(s) g(t+r-s)ds, \tag{6.2.17}$$

with boundary condition

$$\Delta Z_\ell(t,0) = -\Delta D_\ell(t) + \Delta D_{\ell+1}(t) + \Delta\Lambda_\ell(t). \tag{6.2.18}$$

Now for every $t \geq 0$ and $\ell \geq 1$, define

$$V_\ell(t) \doteq \sup_{r\geq 0} |\Delta Z_\ell(t,r)|. \tag{6.2.19}$$

Note that V_ℓ is measurable because it is the supremum over measurable functions, and is bounded by 2 because $Z_\ell(t,\cdot)$ and $\widetilde{Z}_\ell(t,\cdot)$ are both in $\mathbb{C}^1_{\leq 1}[0,\infty)$. By definition (6.2.15) of R_1, we have,

$$\Delta R_1(t,r) = -\int_0^t \lambda(s)\overline{G}(t+r-s)\big(Z_1(s,0) + \widetilde{Z}_1(s,0)\big)\Delta Z_1(s,0)ds,$$

and hence,

$$|\Delta R_1(t,r)| \leq 2\int_0^t \lambda(s) V_1(s)ds, \qquad \forall t, r \geq 0. \tag{6.2.20}$$

Similarly for $\ell \geq 2$, by definition (6.2.15) of R_ℓ,

$$\begin{aligned}\Delta R_\ell(t,r) = &\int_0^t \lambda(s)\big(Z_{\ell-1}(s,0) + Z_\ell(s,0)\big)\big(\Delta Z_{\ell-1}(s,t+r-s) - \Delta Z_\ell(s,t+r-s)\big)ds \\ &+ \int_0^t \lambda(s)\big(\Delta Z_{\ell-1}(s,0) + \Delta Z_\ell(s,0)\big)\big(\widetilde{Z}_{\ell-1}(s,t+r-s) - \widetilde{Z}_1(s,t+r-s)\big)ds,\end{aligned}$$

and hence, for all $t, r \geq 0$

$$|\Delta R_\ell(t,r)| \leq 4\int_0^t \lambda(s)\big(V_{\ell-1}(s) + V_\ell(s)\big)ds. \tag{6.2.21}$$

Furthermore, to bound ΔD_ℓ for $\ell \geq 2$, we substitute $\Delta D_{\ell+1}$

from (6.2.18) in (6.2.17) and set $r = 0$ to conclude that ΔD_ℓ satisfies the renewal equation

$$\Delta D_\ell(t) = g * \Delta D_\ell(t) + F_\ell(t), \qquad t \geq 0,$$

with

$$F_\ell(t) \doteq \Delta\Lambda_\ell(t) - g * \Delta\Lambda_\ell(t) + g * \Delta Z_\ell(t,0) - \Delta R_\ell(t,0).$$

Fix $\ell \geq 2$ and note that by definition (6.2.16) of Λ_ℓ,

$$\begin{aligned}\Delta\Lambda_\ell(t) = &\int_0^t \lambda(s)\big(Z_{\ell-1}(s,0) + \widetilde{Z}_{\ell-1}(s,0)\big)\Delta Z_{\ell-1}(s,0)ds \\ &- \int_0^t \lambda(s)\big(Z_\ell(s,0) + \widetilde{Z}_\ell(s,0)\big)\Delta Z_\ell(s,0)ds,\end{aligned}$$

and hence,

$$|\Delta\Lambda_\ell(t)| \leq 4\int_0^t \lambda(s)\big(V_{\ell-1}(s) + V_\ell(s)\big)ds, \qquad \forall t \geq 0. \tag{6.2.22}$$

Recall that $V_\ell(t) \leq 2$ for all $t \geq 0$ and $\ell \geq 1$, and due to the local integrability of λ, (6.2.22) implies that $\Delta\Lambda_\ell(t)$ is uniformly bounded on any finite interval $t \in [0,T]$. Hence,

$$g * \Delta\Lambda_\ell(t) \leq \|\Delta\Lambda\|_T \int_0^t g(s)ds < \infty, \qquad t \in [0,T].$$

Similarly, the bound (6.2.21) shows that $\Delta R_\ell(t)$ is also bounded for finite t. Therefore, F_ℓ is bounded on finite intervals, and hence by the renewal theorem (see Theorem V.2.4 in [**3**])

$$\Delta D_\ell(t) = \int_0^t F_\ell(t-s)dU_G(s),$$

where U_G is the renewal measure corresponding to the service distribution G. Since G has a density g, U_G satisfies

$$U_G(x) = 1 + \int_0^x u_G(y)dy, \qquad x \geq 0,$$

and the density u_G is bounded on finite intervals and satisfies the equation $u_G = g * u_G + g$ due to Proposition V.2.7 in [**3**]. Therefore, $\Delta D_\ell(t)$ can be written as

(6.2.23)

$$\Delta D_\ell(t) = F_\ell(t) + u_G * F_\ell(t) = \Delta\Lambda_\ell(t) + u_G * \Delta Z_\ell(t,0) - \Delta R_\ell(t,0) - u_G * \Delta R_\ell(t,0).$$

For every fixed $T \geq 0$ and all $0 \leq t \leq T$, by the definition of the convolution operator and (6.2.19),

$$|u_G * \Delta Z_\ell(t,0)| \leq \int_0^t u_G(s)|\Delta Z_\ell(t-s,0)|ds \leq \|u_G\|_T \int_0^t V_\ell(s)ds. \tag{6.2.24}$$

Also, using the bound (6.2.21) we have

$$
\begin{aligned}
|u_G * \Delta R_\ell(t,0)| \leq & 4\int_0^t u_G(t-s)\int_0^s \lambda(v)\big(V_{\ell-1}(v)+V_\ell(v)\big)dv\, ds \\
= & 4\int_0^t \lambda(v)\big(V_{\ell-1}(v)+V_\ell(v)\big)\int_v^t u_G(t-s)ds\, dv \\
\leq & 4U_G(T)\int_0^t \lambda(s)\big(V_{\ell-1}(s)+V_\ell(s)\big)ds. \qquad (6.2.25)
\end{aligned}
$$

Bounding the terms on the right-hand side of (6.2.23) using inequalities (6.2.21), (6.2.22), (6.2.24) and (6.2.25), for every $\ell \geq 2$ and $t \leq T$, we have

$$
\|\Delta D_\ell\|_t \leq C_T \int_0^t (1+\lambda(s))(V_{\ell-1}(s)+V_\ell(s))ds, \qquad (6.2.26)
$$

with $C_T \doteq 8 + 4U_G(T) + \|u_G\|_T$. Next, substituting the bound (6.2.26), but with ℓ replaced by $\ell+1$, (6.2.20) and (6.2.21) into (6.2.17), we obtain that for all $0 \leq t \leq T$ and $r \geq 0$,

$$
|\Delta Z_1(t,r)| \leq 4(C_T\overline{G}(r)+1)\int_0^t (1+\lambda(s))(V_1(s)+V_2(s))ds, \qquad (6.2.27)
$$

and for $\ell \geq 2$,

$$
|\Delta Z_\ell(t,r)| \leq 4(C_T\overline{G}(r)+1)\int_0^t (1+\lambda(s))(V_{\ell-1}+V_\ell(s)+V_{\ell+1}(s))ds. \qquad (6.2.28)
$$

Taking supremum over $r \geq 0$ on both sides of (6.2.27) and (6.2.28), we have

$$
V_1(t) \leq 8C_T \int_0^t (1+\lambda(s))(V_1(s)+V_2(s))ds, \qquad (6.2.29)
$$

and for $\ell \geq 2$,

$$
V_\ell(t) \leq 8C_T \int_0^t (1+\lambda(s))(V_{\ell-1}(s)+V_\ell(s)+V_{\ell+1}(s))ds. \qquad (6.2.30)
$$

Now define

$$
V(t) \doteq \sum_{\ell \geq 1} 2^{-\ell} V_\ell(s). \qquad (6.2.31)
$$

Note that V is measurable, and $V(t) = 0$ if and only if $V_\ell(t) = 0$ for all $\ell \geq 1$. Considering the weighted sums of both sides of (6.2.29) and (6.2.30) over $\ell \geq 2$, we obtain

$$V(t) \leq 26C_T \int_0^t (1+\lambda(s))V(s)ds, \qquad 0 \leq t \leq T. \tag{6.2.32}$$

An application of Gronwall's inequality shows that $V \equiv 0$, and hence, $Z_\ell = \widetilde{Z}_\ell$ for all $\ell \geq 0$. This completes the proof. $\square$

CHAPTER 7

Hydrodynamic Limit Theorems

In this Chapter, we establish the main result of Part II, that is a hydrodynamic limit theorem for the state representation of the load balancing network described in Chapter 5. Specifically, we prove that the scaled interacting measure-valued representation $(\nu_\ell^{(N)}; \ell \geq 1)$ defined in Section 5.1.2, converges to the unique solution of the measure-valued hydrodynamic equations of Section 6.1.1 (see Theorem 7.1 below).

The measure-valued hydrodynamic limit theorem implies, in particular, the convergence of the fraction of queues with length greater than or equal to ℓ, to a functional $(S_\ell; \ell \geq 1)$ of the measure-valued limit process $(\nu_\ell; \ell)$, although $(\nu_\ell; \ell \geq 1)$ contains more information than just S_ℓ. In this chapter, we show that if one is only interested the limit of queue length distribution, the limit can be characterized by the hydrodynamic PDEs. As depicted in the next chapter, this limit process can then be analyzed using numerical approximations for the solutions to these PDEs. Moreover, we show in this chapter that the PDEs can be used to approximate not only the queue length distribution but also other quality of service (QoS) parameters such as the virtual waiting time.

Here we state the main result of this chapter. For $H = E, D, \nu_\ell, \nu, S_\ell$, we define the scaled version $\overline{H}^{(N)}$ of the process $H^{(N)}$ as follows:

$$\overline{H}^{(N)}(t) = \frac{H^{(N)}(t)}{N}, \quad N \in \mathbb{N}, t \geq 0. \tag{7.0.33}$$

The following additional conditions are imposed on the sequence of initial conditions.

ASSUMPTION III. (cont'd.) The sequence of initial conditions satisfies the followings.

(b) there exists $\nu(0) = (\nu_\ell(0); \ell \geq 1) \in \mathbb{S}$ such that $\overline{\nu}^{(N)}(0) \to \nu(0)$ in $\mathbb{S}$, $\mathbb{P}$-almost surely, as $N \to \infty$.

(c) $\limsup_N \mathbb{E}[\overline{X}^{(N)}(0)] < \infty$, and $\overline{X}^{(N)}(0) \to X(0)$ as $N \to \infty$, where

$$X(0) \doteq \sum_{\ell \geq 1} \langle \mathbf{1}, \nu_\ell(0) \rangle.$$

THEOREM 7.1. Suppose Assumptions I - III hold. Then the sequence $\{\overline{\nu}^{(N)}\}$ converges in distribution to a function ν in $\mathbb{D}_{\mathbb{S}}[0,\infty)$ which satisfies the hydrodynamic equations associated with $(\lambda \mathbf{Id}, \nu(0))$.

PROOF. See Section 7.2.3. □

The following corollary is a direct implication of Theorem 6.3 and 7.1.

COROLLARY 7.2. Suppose Assumptions I-III hold. Then Given $\nu(0) \in \mathbb{S}$ and λ, the hydrodynamic equations associated with $(\nu(0), \lambda \mathbf{Id})$ has a unique solution.

We first prove the tightness of state processes as well as other auxiliary processes in Section 7.1, which is summarized in Theorem 7.20. Then in Section 7.2, we characterize every subsequential limit of the state processes as a solution to the hydrodynamic equations defined in Section 6.1.1, and prove Theorem 7.1 (see Section 7.2.3). Finally, we establish the convergence of queue length distribution and virtual waiting time, and characterize the corresponding limits as functionals of the solution to the hydrodynamic PDEs.

7.1. Tightness

As the first step towards establishing the hydrodynamic limit theorem 7.1, we prove that the sequence $\{\overline{\nu}^{(N)}\}_{N \in \mathbb{N}}$ of state processes is tight in its state space $\mathbb{D}_{\mathbb{S}}[0,\infty)$ (see Theorem 7.20).

First in Section 7.1.1, we recall two tightness criteria that are used throughout this section. The tightness of arrival process and weighed routing process are proved in Section 7.1.2. Then in Section 7.1.3, we recall some results from the G/G/N model in **[61]**, that continue to hold in our setting. Section 7.1.4 focuses on establishing tightness of departure processes. Finally in Section 7.1.5, we establish tightness for state variables and other measure-valued processes, and summarize all these results in Theorem 7.20.

7.1.1. Review of Tightness Criteria. We first recall the Kurtz's criteria to prove tightness for a sequence of random elements in $\mathbb{D}_{\mathbb{R}}[0,\infty)$. Recall from Section 1.4 that for a function f in $\mathbb{D}_{\mathbb{R}}[0,\infty)$, $w'(f,\cdot,\cdot)$ denotes the the modulus of continuity of f in $\mathbb{D}_{\mathbb{R}}[0,\infty)$.

PROPOSITION 7.3 (Kurtz's criteria). A sequence of processes $\{Y^{(N)}\}_{N\in\mathbb{N}}$ with sample paths in $\mathbb{D}_{\mathbb{R}}[0,\infty)$ is relatively compact (and hence, tight) if and only if $\{Y^{(N)}\}_{N\in\mathbb{N}}$ satisfies

K1.: For every rational $t \geq 0$,

$$\lim_{r\to\infty}\sup_N \mathbb{P}\left\{|Y^{(N)}(t)| > r\right\} = 0, \tag{7.1.1}$$

and

K2a.: For every $\eta > 0$ and $T > 0$, there exists $\delta > 0$ such that

$$\sup_N \mathbb{P}\left\{w'(Y^{(N)},\delta,T) \geq \eta\right\} \leq \eta. \tag{7.1.2}$$

Moreover, if $\{Y^{(N)}\}_{N\in\mathbb{N}}$ satisfies K1 and

K2b.: For each $T \geq 0$, there exists $\beta > 0$ such that

$$\lim_{\delta\to 0}\sup_N \mathbb{E}\left[\sup_{0\leq t\leq T}\left|Y^{(N)}(t+\delta) - Y^{(N)}(t)\right|^{\beta}\right] = 0, \tag{7.1.3}$$

then, $\{Y^{(N)}\}_{N\in\mathbb{N}}$ is relatively compact.

PROOF. The necessity and sufficiency of K1 and K2a follows from Theorem 3.7.2 in [**33**] and the sufficiency of K1 and K2b follows from Theorems 3.7.2 and 3.8.6 and Remark 3.8.7 in [**33**]. Since $\mathbb{D}_{\mathbb{R}}[0,\infty)$ is a Polish space, tightness follows from relative compactness by Prohorov's theorem. □

Moreover, recall from the Section 1.4 that for $\mathcal{Z} = [0, L)$ or $\mathcal{Z} = [0, L) \times [0, \infty)$, the space $\mathbb{M}_F(\mathcal{Z})$ of finite measures on $\mathcal{Z}$, equipped with the topology of weak convergence is a complete separable metric space. To prove the tightness of measure-valued processes, we will invoke Jakubowski's criteria. In what follows, recall that a family $\mathbb{F}$ of real continuous functionals on $\mathbb{M}_F(\mathcal{Z})$ is called to separate the points in $\mathbb{M}_F(\mathcal{Z})$ if for every distinct $\mu, \tilde{\mu} \in \mathbb{M}_F(\mathcal{Z})$, there exists a function $f \in \mathbb{F}$ such that $f(\mu) \neq f(\tilde{\mu})$.

PROPOSITION 7.4 (Jakubowski's criteria). Let $\mathcal{Z}$ be either $[0, L)$ or $[0, L) \times [0, \infty)$. A sequence $\{\pi^{(N)}\}_{N\in\mathbb{N}}$ of $\mathbb{D}_{\mathbb{M}_F(\mathcal{Z})}[0, \infty)$-valued random elements is tight if and only if

J1.: (Compact containment condition) For each $T > 0$ and $\eta > 0$ there exists a compact set $\mathcal{K}_{T,\eta} \subset \mathbb{M}_F(\mathcal{Z})$ such that

$$\liminf_N \mathbb{P}\left\{\pi_t^{(N)} \in \mathcal{K}_{T,\eta} \text{ for all } t \in [0, T]\right\} > 1 - \eta.$$

J2.: There exists a family $\mathbb{F}$ of real continuous functionals on $\mathbb{M}_F(\mathcal{Z})$ that (i) is closed under addition, and (ii) separates points in $\mathbb{M}_F(\mathcal{Z})$, such that $\{\pi^{(N)}\}$ is $\mathbb{F}$-weakly tight, that is for every $F \in \mathbb{F}$, the sequence $\{F(\pi_s^{(N)}); s \geq 0\}_{N\in\mathbb{N}}$ is tight in $\mathbb{D}_{\mathbb{R}}[0, \infty)$.

PROOF. See Theorem 4.6 of **[55]**. □

REMARK 7.5. It can be easily verified that the set

$$\mathbb{F} = \{F : \exists f \in \mathbb{C}_c^1(\mathcal{Z}) \text{ such that } F(\mu) = \langle f, \mu \rangle, \ \ \forall \mu \in \mathbb{M}_F(\mathcal{Z})\},$$

satisfies the properties (i) and (ii) in condition J2.

7.1.2. Tightness of Arrival Processes. Recall from Assumption I that for every $N \in \mathbb{N}$, the arrival process $E^{(N)}$ is a renewal process with inter-arrival times $\{u_n^{(N)}, n \geq 1\}$ with $\mathbb{E}[u^{(N)}] = (N\lambda)^{-1}$. Also, recall that **Id** denotes the identity function on $[0, \infty)$.

LEMMA 7.6. Suppose Assumption I holds. Then, as $N \to \infty$, $\overline{E}^{(N)} \to \lambda\mathbf{Id}$ in $\mathbb{D}_{\mathbb{R}}[0,\infty)$, $\mathbb{P}$-almost surely, and in particular, $\{\overline{E}^{(N)}\}_{N\in\mathbb{N}}$ is tight in $\mathbb{D}_{\mathbb{R}}[0,\infty)$. Moreover, for all $t \geq 0$, $\mathbb{E}[\overline{E}^{(N)}(t)] \to \mathbb{E}[\lambda t]$ as $N \to \infty$, and hence,

$$\limsup_{N\to\infty} \mathbb{E}\left[\overline{E}^{(N)}(t)\right] < \infty. \tag{7.1.4}$$

Furthermore,

$$\limsup_{N\to\infty} \mathbb{E}\left[\overline{E}^{(N)}(t)^2\right] < \infty. \tag{7.1.5}$$

PROOF. The almost sure convergence of $\overline{E}^{(N)}$ to $\lambda\mathbf{Id}$ in $\mathbb{D}_{\mathbb{R}}[0,\infty)$ is by the functional Low of Large Numbers for renewal processes (see e.g. [**20**, Theorem 5.10].) Also, for every $t \geq 0$, by elementary renewal theorem (see e.g. [**3**, Proposition V.1.4]),

$$\lim_{N\to\infty} \mathbb{E}\left[\overline{E}^{(N)}(t)\right] = \lim_{N\to\infty} \frac{1}{N}\widetilde{E}(Nt) = t.$$

Let $U_{\widetilde{E}}$ be the renewal measure associated with the inter-arrival distribution $G_{\widetilde{E}}$. Then, by elementary calculations (or see (2.3) in [**26**]), we have

$$\mathbb{E}\left[\widetilde{E}(t)^2\right] \leq U_{\widetilde{E}}(t) + \int_0^t U_{\widetilde{E}}(t-s)dU_{\widetilde{E}}(s) \leq 2U_{\widetilde{E}}(t)^2.$$

Hence, by another use of the elementary renewal theorem,

$$\limsup_{N\to\infty} \mathbb{E}\left[\overline{E}^{(N)}(t)^2\right] = \limsup_{N\to\infty} \frac{1}{N^2}\mathbb{E}\left[\widetilde{E}(Nt)^2\right] \leq 2t^2 \limsup_{N\to\infty} \left(\frac{U_{\widetilde{E}}(Nt)}{Nt}\right)^2 = 2\lambda^2 t^2,$$

which shows (7.1.5). $\square$

REMARK 7.7. Using Lemma 7.6, it follows from Assumptions I and III.c that for every $t \geq 0$,

$$\limsup_{N\to\infty} \mathbb{E}\left[\overline{E}^{(N)}(t) + \overline{X}^{(N)}(0)\right] < \infty. \tag{7.1.6}$$

Next, we prove tightness for the sequence of weighted routing process $\{\mathcal{R}^{(N)}_{\varphi,\ell}\}_{N\in\mathbb{N}}$. For every $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$, $i = 1,...,N$, and $u \geq 0$, $\varphi(0,u)\mathbb{1}(X^{(N),i}(u-) = 0) \leq \|\varphi\|_\infty$, and for every $\ell \geq 2$, by representation (5.3.18) of $\nu^{(N),i}_\ell$,

$$\langle \varphi(\cdot,u), \nu^{(N),i}_{\ell-1}(u-) - \nu^{(N),i}_\ell(u-)\rangle = \varphi\left(a^{(N),i}(u-),u\right)\mathbb{1}\left(X^{(N),i}(u-) = \ell\right) \leq \|\varphi\|_\infty.$$

Hence, since $E^{(N),i}$ is non-decreasing, and by definition (5.1.19) of $\mathcal{R}^{(N)}_{\varphi,\ell}$, for every $\ell \geq 1$ and $0 \leq s \leq t$, the bound

$$
\begin{aligned}
(7.1.7) \qquad |\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}(t) - \overline{\mathcal{R}}^{(N)}_{\varphi,\ell}(s)| &\leq \frac{1}{N}\sum_{i=1}^{N} \|\varphi\|_\infty \left(E^{(N),i}(t) - E^{(N),i}(s) \right) \\
&= \|\varphi\|_\infty \left(\overline{E}^{(N)}(t) - \overline{E}^{(N)}(s) \right)
\end{aligned}
$$

holds.

LEMMA 7.8. Suppose Assumption I holds, and fix $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$ and $\ell \geq 1$. Then, for every $t \geq 0$ we have

$$
(7.1.8) \qquad \limsup_{N\to\infty} \mathbb{E}\left[\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}(t) \right] \leq \|\varphi\|_\infty \limsup_{N\to\infty} \mathbb{E}\left[\overline{E}^{(N)}(t) \right] < \infty.
$$

Moreover, the sequence $\{\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}\}_{N\in\mathbb{N}}$ is tight in $\mathbb{D}_{\mathbb{R}}[0,\infty)$.

PROOF. Inequality (7.1.8) follows by taking limit superior as $N \to \infty$ from both sides of the bound (7.1.7) with $s = 0$, and using (7.1.4). Moreover, (7.1.7) shows that the modulus of continuity w' for $\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}$ is bounded by that of $\overline{E}^{(N)}$. Tightness of $\{\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}\}_{N\in\mathbb{N}}$ then follows from tightness of $\overline{E}^{(N)}$, proved in Lemma 7.6, and the necessity and sufficiency of Kurtz's criteria K1 and K2a stated in Proposition 7.3. $\square$

Next, recall the definition of $\mathcal{N}^{(N)}_{\varphi,\ell}$ in (5.3.3).

LEMMA 7.9. Suppose Assumptions I, II.a, and III.a hold, and fix $N \in \mathbb{N}$, $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$ and $\ell \geq 1$. Then $\mathcal{N}^{(N)}_{\varphi,\ell}$ is a square integrable $\{\mathcal{F}^{(N)}_t\}$-martingale. Moreover, for every $t \geq 0$ and $\epsilon > 0$,

$$
(7.1.9) \qquad \limsup_{N\to\infty} \mathbb{P}\left\{ \left| \sup_{0\leq s\leq t} \overline{\mathcal{N}}^{(N)}_{\varphi,\ell}(s) \right| > \epsilon \right\} = 0.
$$

and $\overline{\mathcal{N}}^{(N)}_{\varphi,\ell} \Rightarrow 0$ in $\mathbb{D}_{\mathbb{R}}[0,\infty)$, as $N \to \infty$.

PROOF. By Proposition 5.11, $\mathcal{N}^{(N)}_{\varphi,\ell}$ is a local martingale with $[\mathcal{N}^{(N)}_{\varphi,\ell}] = \mathcal{R}^{(N)}_{\varphi^2,\ell}$. Since for every $t \geq 0$, $\mathbb{E}[\mathcal{R}^{(N)}_{\varphi^2,\ell}(t)] < \infty$ by (7.1.8) (with φ replaced by φ^2), by Theorem 7.35 in **[66]**, $\mathcal{N}^{(N)}_{\varphi,\ell}$ is a square integrable martingale and,

$$\mathbb{E}\left[\left(\mathcal{N}^{(N)}_{\varphi,\ell}(t)\right)^2\right] = \mathbb{E}\left[\mathcal{R}^{(N)}_{\varphi^2,\ell}(t)\right].$$

Therefore, an application of Doob's inequality shows that for every $T \geq 0$ and $\epsilon > 0$,

$$\mathbb{P}\left\{\left|\sup_{t\in[0,T]} \overline{\mathcal{N}}^{(N)}_{\varphi,\ell}(t)\right| > \epsilon\right\} \leq \frac{1}{N\epsilon^2}\mathbb{E}\left[\overline{\mathcal{R}}^{(N)}_{\varphi^2,\ell}(T)\right].$$

Taking the limit superior of both sides of above inequality as $N \to \infty$ and using the bound (7.1.8), (7.1.9) follows. This shows that $\overline{\mathcal{N}}^{(N)}_{\varphi,\ell}$ converges to zero in $\mathbb{D}_{\mathbb{R}}[0,\infty)$ in probability, and hence in distribution. $\square$

The following result is a direct corollary of Lemmas 7.8 and 7.9.

COROLLARY 7.10. Suppose Assumption I, II.a, and III.a hold. Then, for every $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$ and $\ell \geq 1$, the sequence $\{\overline{B}^{(N)}_{\varphi,\ell}\}_{N\in\mathbb{N}}$ is tight in $\mathbb{D}_{\mathbb{R}}[0,\infty)$.

7.1.3. Results From to the G/G/N Queue. In this section, we recall some of the results in **[61]**, which remain valid in our setting as well. Most of these result follow from bounds on departure process which hold under a fairly general network structures. Here, we state the results using our notation, and argue that they are valid in our setting.

Note that by definition (5.2.24) of the cumulative service entry process $K^{(N)}$ and Lemma 5.9,

$$\int_{[0,t]} \varphi(0,u)dK^{(N)}(u) = \int_{[0,t]} \varphi(0,u)dD_2^{(N)}(u) + \mathcal{R}^{(N)}_{\varphi,1}(t),$$

for every $\varphi \in \mathbb{C}_b([0,L)\times\mathbb{R}_+)$. Substituting the above equation in (5.2.2), the equation (5.4) in **[61]** is satisfied with $\nu^{(N)}$ and $\mathcal{Q}^{(N)}_\varphi$ replaced by $\nu_1^{(N)}$ and $\mathcal{D}^{(N)}_{\varphi,1}$, respectively, that is

$$
\begin{aligned}
\langle\varphi(\cdot,t),\nu_1^{(N)}(t)\rangle =& \langle\varphi(\cdot,0),\nu_1^{(N)}(0)\rangle + \int_0^t \langle\varphi_s(\cdot,s)+\varphi_x(\cdot,s),\nu_1^{(N)}(s)\rangle ds \\
&- \mathcal{D}^{(N)}_{\varphi,1}(t) + \int_{[0,t]}\varphi(0,s)dK^{(N)}(s).
\end{aligned}
\tag{7.1.10}
$$

Moreover, using Lemma 7.6, Assumption I and Assumption III imply Assumption I in **[61]**. The proofs of the results in this section very similar to the that of Theorem 5.14 in **[61]**, as most pasts of that proof does not depend on the particular routing algorithm there.

LEMMA 7.11. For every $N\in\mathbb{N}$, $\varphi\in\mathbb{C}_b([0,L)\times\mathbb{R}_+)$ and $T>0$,

$$
\lim_{\delta\to 0}\sup_N \mathbb{E}\left[\sup_{t\in[0,T]}\left|\overline{A}^{(N)}_{\varphi,1}(t+\delta)-\overline{A}^{(N)}_{\varphi,1}(t)\right|\right]=0.
\tag{7.1.11}
$$

PROOF. The proof is exactly analogous to the proof of Part (2) of Lemma 5.8 in **[61]** ($\overline{A}^{(N)}_\varphi$ there replaced by $\overline{A}^{(N)}_{\varphi,1}$), using the results from Lemmas 5.5 and 5.12 and Proposition 5.7 there. The proof only use Assumption I of **[61]**, the fact that the age of jobs in service have linear growth and the bound

$$
K^{(N)}(t)+\langle\mathbf{1},\nu^{(N)}(0)\rangle \le X^{(N)}(0)+E^{(N)}(t),
\tag{7.1.12}
$$

which continuous to hold in our model with $\nu^{(N)}$ replaced by $\nu_1^{(N)}$, using equations (5.1.22), (5.2.3) and (5.2.25). □

LEMMA 7.12. Suppose Assumptions I, II.a, and III holds. Then

$$
\lim_{m\to L}\sup_N \mathbb{E}\left[\overline{\nu}_1^{(N)}(0,(m,L))\right]=0,
\tag{7.1.13}
$$

and if $L<\infty$,

$$
\lim_{m\to L}\sup_N \mathbb{E}\left[\int_{[0,m)}\frac{1-G(m)}{1-G(x)}\overline{\nu}_1^{(N)}(0,dx)\right]=0 \text{ if } L<\infty.
\tag{7.1.14}
$$

Also, if $L = \infty$, there exist a sequence sequence $r(n) \uparrow \infty$ such that the sequence $\{\overline{\nu}_1^{(N)}\}_{N\in\mathbb{N}}$ satisfies the compact containment condition J1 for the compact set

$$\mathcal{K}_{\eta,T} = \{\mu \in \mathbb{M}_F(\mathbb{R}_+) : \mu(r(n)+T,\infty) \leq \frac{1}{n} \text{ for all } n < -\lceil \log \eta / \log 2 \rceil\}. \tag{7.1.15}$$

Otherwise, if $L < \infty$, there exist $m < L$ such that the sequence $\{\overline{\nu}_1^{(N)}\}_{N\in\mathbb{N}}$ satisfies the compact containment condition J1 for the compact set

$$\mathcal{K}_{\eta,T} = \{\mu \in \mathbb{M}_F([0,L)) : \nu([m,L)) \leq \eta\}. \tag{7.1.16}$$

PROOF. The proof is exactly similar to that of Lemma 5.12 in **[61]** (which uses the notation $\nu^{(N)}$ for our $\nu_1^{(N)}$). $\square$

As discussed in **[61]** (page 96) and we also show later in Section 7.1.5, for every $N \in \mathbb{N}$, $t \geq 0$ and $\ell \geq 1$, the linear functional $\mathcal{D}_{\cdot,1}^{(N)}(t) : \varphi \mapsto \mathcal{D}_{\varphi,1}^{(N)}(t)$ can be identified with a finite non-negative Radon measure on $[0,L) \times \mathbb{R}_+$ (the same quantity is denoted by $\mathcal{Q}^{(N)}$ in **[61]**).

LEMMA 7.13. Suppose Assumptions I, II.a, and III holds. Then, for every $\eta > 0$ and $T \geq 0$, there exist a constant $B(\eta) < \infty$ and a sequence $\{m(n,\eta)\}$ with $m(n,\eta) \to L$ as $n \to \infty$ such that

$$\mathbb{P}\left\{\overline{\mathcal{D}}_{\cdot,1}^{(N)}(t) \notin \mathcal{K}_\eta \text{ for some } t \in [0,T]\right\} \leq \eta, \tag{7.1.17}$$

for the compact subset $\mathcal{K}_\eta \subset \mathbb{M}_F([0,L) \times \mathbb{R}_+)$ defined by

$$\mathcal{K}_\eta = \left\{\mu \in \mathbb{M}_F([0,L) \times \mathbb{R}_+) \ : \langle \mathbf{1}, \mu \rangle \leq B(\eta),\ \mu((m(n,\eta),L) \times R_+) \leq \frac{1}{n}\ \forall n \in \mathbb{N}\right\}. \tag{7.1.18}$$

PROOF. The result follows from Lemma 5.13 in **[61]** (with $\overline{\mathcal{Q}}^{(N)}$ replaced by $\overline{\mathcal{D}}_{\cdot,1}^{(N)}$.) $\square$

Finally, we will use the following result later in Section 7.2.2 to characterize the sub-sequential limits if departure process.

LEMMA 7.14. Suppose Assumptions I-III hold. For every $\ell \geq 1$, suppose $\{\overline{\nu}_\ell^{(N)}\}_{N\in\mathbb{N}}$ converges almost surely to some $\nu_\ell \in \mathbb{D}_{M_F[0,L)}[0,\infty)$ along a subsequence. Then,

(a) for every $T \geq 0$,

$$\lim_{m\uparrow L}\sup_N \mathbb{E}\left[\int_0^T \left(\int_{[m,L)} h(x)\overline{\nu}_1^{(N)}(s,dx)\right) ds\right] = 0; \tag{7.1.19}$$

for every non-negative function $f \in \mathbb{L}^1[0,L)$, there exists a constant $C < \infty$,

$$\sup_N \mathbb{E}\left[\int_0^T \langle f, \overline{\nu}_1^{(N)}(s)\rangle ds\right] \leq C \sup_{u\in[0,L)} \int_u^{(u+T)\wedge L} f(x)dx; \tag{7.1.20}$$

and for every $m < L$ and every non-negative function $f \in \mathbb{L}^1_{\mathrm{loc}}[0,L)$ with support in $[0,m]$, there exists $\tilde{L}(m,T) < \infty$ such that

$$\left|\int_0^T \langle f, \nu_1(s)\rangle ds\right| \leq \tilde{L}(m,T)\int_{[0,L)} f(x)dx. \tag{7.1.21}$$

(b) For every $\ell \geq 1$, if the bounds (7.1.19)-(7.1.21) are satisfied for $\nu_1^{(N)}$ and ν_1 substituted by $\nu_\ell^{(N)}$ and ν_ℓ, respectively, then

$$\limsup_{N\to\infty} \mathbb{E}\left[\left|\overline{A}_{\varphi,\ell}^{(N)}(t) - \int_0^t \langle \varphi(\cdot,s)h(\cdot), \nu_\ell(s)\rangle ds\right|\right] = 0. \tag{7.1.22}$$

PROOF. Recall that our Assumption I and Assumption III imply Assumption I in **[61]**. For part a, (7.1.19), (7.1.20) and (7.1.21) follow respectively from Lemma 5.8(1), Proposition 5.7 and Lemma 5.16 in **[61]** (with $\nu^{(N)}$ replaced by $\nu_1^{(N)}$). Also, for Part b, (7.1.22) can be obtained in an exactly analogous way as in the proof of Proposition 5.17 in **[61]** with $\nu^{(N)}$ and ν there replaced by $\nu_\ell^{(N)}$ and ν_ℓ, respectively. $\square$

7.1.4. Tightness of Departure Processes. Next, we show the tightness of the departure processes. First, note that using (5.1.24) (with φ replaced by φ^2) and (7.1.6), we have

$$\limsup_{N\to\infty} \mathbb{E}\left[\overline{\mathcal{D}}_{\varphi^2,\ell}^{(N)}(t)\right] \leq \|\varphi^2\|_\infty \limsup_{N\to\infty} \mathbb{E}\left[\overline{X}^{(N)}(0) + \overline{E}^{(N)}(t)\right] < \infty. \tag{7.1.23}$$

Recall the definition 5.3.7 of $\overline{\mathcal{M}}^{(N)}_{\varphi,\ell}$.

LEMMA 7.15. Suppose Assumptions I, II.a, and III.a hold, and fix $N \in \mathbb{N}$, $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$, $\ell \geq 1$. Then, $\mathcal{M}^{(N)}_{\varphi,\ell}$ is a square integrable martingale. Moreover, $t \geq 0$ and $\epsilon > 0$,

$$\limsup_{N\to\infty} \mathbb{P}\left\{ \left| \sup_{0\leq s\leq t} \overline{\mathcal{M}}^{(N)}_{\varphi,\ell}(s) \right| > \epsilon \right\} = 0, \tag{7.1.24}$$

and $\overline{\mathcal{M}}^{(N)}_{\varphi,\ell} \Rightarrow 0$ in $\mathbb{D}_{\mathbb{R}}[0,\infty)$, as $N \to \infty$.

PROOF. By Proposition 5.11, $\mathcal{M}^{(N)}_{\varphi,\ell}$ is a local martingale with $[\mathcal{M}^{(N)}_{\varphi,\ell}] = \mathcal{D}^{(N)}_{\varphi^2,\ell}$. By (7.1.23) and Theorem 7.35 in **[66]**, $\mathcal{M}^{(N)}_{\varphi,\ell}$ is a square integrable $\{\mathcal{F}^{(N)}_t\}$-martingale and for all $t \geq 0$,

$$\mathbb{E}\left[\left(\mathcal{M}^{(N)}_{\varphi,\ell}(t)\right)^2\right] = \mathbb{E}\left[\mathcal{D}^{(N)}_{\varphi^2,\ell}(t)\right].$$

Therefore, an application of Doob's inequality shows that for every $T \geq 0$ and $\epsilon > 0$,

$$\mathbb{P}\left\{ \left| \sup_{t\in[0,T]} \overline{\mathcal{M}}^{(N)}_{\varphi,\ell}(t) \right| > \epsilon \right\} \leq \frac{1}{N\epsilon^2}\mathbb{E}\left[\overline{\mathcal{D}}^{(N)}_{\varphi^2,\ell}(T)\right].$$

Taking the limit superior of both sides of above inequality as $N \to \infty$ and using the bound (7.1.23), (7.1.24) follows. This shows that $\overline{\mathcal{M}}^{(N)}_{\varphi,\ell}$ converges to zero in $\mathbb{D}_{\mathbb{R}}[0,\infty)$ in probability, and hence in distribution. $\square$

LEMMA 7.16. Suppose Assumptions I, II.a, and III.a hold. Then, for every $\ell \geq 1$ and $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$, the sequences $\{\overline{D}^{(N)}_{\ell}\}_{N\in\mathbb{N}}$, $\{\overline{\mathcal{D}}^{(N)}_{\varphi,\ell}\}_{N\in\mathbb{N}}$, and $\{\overline{A}^{(N)}_{\varphi,\ell}\}_{N\in\mathbb{N}}$ are tight in $\mathbb{D}_{\mathbb{R}}[0,\infty)$.

PROOF. By definition (5.3.7) of $\mathcal{M}^{(N)}_{\varphi,\ell}$, and since $\mathcal{M}^{(N)}_{\varphi,\ell}$ is a martingale by Lemma 7.15, for every $t \geq 0$, $\mathbb{E}[\mathcal{M}^{(N)}_{\varphi_\ell}(t)] = 0$, and hence $\mathbb{E}[A^{(N)}_{\varphi,\ell}(t)] = \mathbb{E}[\mathcal{D}^{(N)}_{\varphi,\ell}(t)]$. Therefore, using the bound (7.1.23), we have

$$\limsup_{N\to\infty} \mathbb{E}\left[\left|\overline{A}^{(N)}_{\varphi,\ell}(t)\right|\right] \leq \limsup_{N\to\infty} \mathbb{E}\left[\overline{A}^{(N)}_{|\varphi|,\ell}(t)\right] < \infty. \tag{7.1.25}$$

The Kurtz's K1 and K2b conditions for the sequence $\{\overline{A}^{(N)}_{\varphi,1}\}_{N\in\mathbb{N}}$ follow from (7.1.25) (with $\ell = 1$) and (7.1.11) of Lemma 7.11, respectively, and therefore

$\{\overline{A}^{(N)}_{\varphi,1}\}_{N\in\mathbb{N}}$ is tight. Moreover, for $\ell \geq 1$, note that by definition of $A^{(N)}_{\varphi,\ell}$ and since for every non-negative function f $\langle f, \nu^{(N)}_1 \rangle \geq \langle f, \nu^{(N)}_\ell \rangle$, for all $0 \leq s \leq t$ we have

$$|A^{(N)}_{\varphi,\ell}(t)| \leq A^{(N)}_{|\varphi|,1}(t),$$

and

$$\left| A^{(N)}_{\varphi,\ell}(t) - A^{(N)}_{\varphi,\ell}(s) \right| \leq \left| A^{(N)}_{|\varphi|,1}(t) - A^{(N)}_{|\varphi|,1}(s) \right|.$$

Therefore, the Kurtz's conditions K1 and K2b for $\{\overline{A}^{(N)}_{\varphi,\ell}\}_{N\in\mathbb{N}}$ follows from those of $\{\overline{A}^{(N)}_{\varphi,1}\}_{N\in\mathbb{N}}$, and hence $\{\overline{A}^{(N)}_{\varphi,\ell}\}_{N\in\mathbb{N}}$ is also tight. Also, by Lemma 7.15, the sequence $\{\overline{\mathcal{M}}^{(N)}_{\varphi,\ell}\}_{N\in\mathbb{N}}$ converges weakly to zero and thus is tight, and hence, using Equation (5.3.7), the sequence $\{\overline{\mathcal{D}}^{(N)}_{\varphi,\ell}\}$ is also tight. In particular, substituting $\varphi = \mathbf{1}$ implies the tightness of $\{\overline{D}^{(N)}_\ell\}_{N\in\mathbb{N}}$ for every $\ell \geq 1$. $\square$

7.1.5. Tightness of State Variables and Other Measure-Valued Processes.

PROPOSITION 7.17. Suppose Assumptions I, II.a, and III.a hold. Then, for every $\ell \geq 1$, the sequence $\{\overline{\nu}^{(N)}_\ell\}_{N\in\mathbb{N}}$ is tight in $\mathbb{D}_{\mathbb{M}_F[0,L)}[0,\infty)$.

PROOF. We use the Jakubowski's criteria. It has been shown in Lemma 7.12 that the sequence $\{\overline{\nu}^{(N)}_1\}_{N\in\mathbb{N}}$ satisfies the compact containment condition J1 for the compact set $\mathcal{K}_{\eta,T}$ which is equal to either (7.1.15) or (7.1.16) depending on whether L is finite or infinite. Moreover, for every interval I, every $t \geq 0$ and every $\ell \geq 2$, $\overline{\nu}^{(N)}_\ell(t,I) \leq \overline{\nu}^{(N)}_1(t,I)$, and hence, the condition J1 also holds for $\{\overline{\nu}^{(N)}_\ell\}$ with the same compact set $\mathcal{K}_{\eta,T}$.

Now fix any $f \in \mathbb{C}^1_c[0,L)$. Regarding f as a function on $[0,L) \times \mathbb{R}_+$, it follows from Proposition 5.6 and Remark 5.10 that

$$\langle f, \overline{\nu}^{(N)}_\ell(t) \rangle = \langle f, \overline{\nu}^{(N)}_\ell(0) \rangle + \int_0^t \langle f', \overline{\nu}^{(N)}_\ell(s) \rangle ds - \overline{\mathcal{D}}^{(N)}_{f,\ell}(t) + f(0)\overline{D}^{(N)}_{\ell+1}(t) + \overline{\mathcal{R}}^{(N)}_{f,\ell}(t).$$

Tightness of $\{\langle f, \overline{\nu}^{(N)}_\ell(0) \rangle\}_{N\in\mathbb{N}}$ follows from Assumption III.b, tightness of the sequences $\{\overline{\mathcal{D}}^{(N)}_{f,\ell}\}_{N\in\mathbb{N}}$ and $\{\overline{D}^{(N)}_{\ell+1}\}_{N\in\mathbb{N}}$ follows from Lemma 7.16, and tightness of

$\{\overline{\mathcal{R}}^{(N)}_{f,\ell}\}_{N\in\mathbb{N}}$ follows from Lemma 7.8. Moreover, both criteria K1 and K2b, and hence the tightness of $\{\int_0^{\cdot}\langle f',\overline{\nu}^{(N)}_\ell(s)\rangle ds\}_{N\in\mathbb{N}}$ follows from the bound

$$\left|\int_t^{t+\delta}\langle f',\overline{\nu}^{(N)}_\ell(s)\rangle ds\right| \le \delta\|f'\|_\infty,$$

which uses the fact that $\overline{\nu}^{(N)}_\ell(s)$ is a sub-probability measure. Therefore, for all $\ell\ge 1$, $\{\langle f,\overline{\nu}^{(N)}_\ell(t)\rangle\}_{N\in\mathbb{N}}$ is tight in $\mathbb{D}_{\mathbb{R}}[0,\infty)$ for every $f\in\mathbb{C}^1_b[0,\infty)$, and hence, due to Remark 7.5, $\{\nu^{(N)}_\ell\}$ satisfies the criterion J2. This completes the proof. $\square$

For all $t\ge 0$, $\ell\ge 1$ and $N\in\mathbb{N}$, it follows, respectively, from the bounds (5.1.24) and (7.1.7) that the linear functionals $\mathcal{D}^{(N)}_{\cdot,\ell}:\varphi\mapsto\mathcal{D}^{(N)}_{\varphi,\ell}$ and $\mathcal{R}^{(N)}_{\cdot,\ell}:\varphi\mapsto\mathcal{R}^{(N)}_{\varphi,\ell}$ can be identified with finite nonnegative (Radon) measures on $[0,L)\times\mathbb{R}_+$ (see Section 1.4.) This means that for every $\varphi\in\mathbb{C}^{1,1}_c([0,L)\times\mathbb{R}_+)$, the integrals of φ with respect to the Radon measures $\mathcal{D}^{(N)}_{\cdot,\ell}(t)$ and $\mathcal{R}^{(N)}_{\cdot,\ell}(t)$ are $\mathcal{D}^{(N)}_{\varphi,\ell}(t)$ and $\mathcal{R}^{(N)}_{\varphi,\ell}(t)$, respectively, and thus $\{\mathcal{D}^{(N)}_{\cdot,\ell}(t),t\ge 0\}$ and $\{\mathcal{R}^{(N)}_{\cdot,\ell}(t),t\ge 0\}$ can be viewed as $\mathbb{M}_F([0,L)\times\mathbb{R}_+)$-valued càdlàg processes. The tightness of these sequences is established in the next two lemmas.

LEMMA 7.18. Suppose Assumptions I-III hold. Then for every $\ell\ge 1$, the sequence $\{\overline{\mathcal{R}}^{(N)}_{\cdot,\ell}\}_{N\in\mathbb{N}}$ is tight in $\mathbb{D}_{\mathbb{M}_F([0,L)\times\mathbb{R}_+)}[0,\infty)$.

PROOF. We have already shown in Lemma 7.8 that for every $\varphi\in\mathbb{C}_b([0,L)\times\mathbb{R}_+)$, $\{\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}\}_{N\in\mathbb{N}}$ is tight in $\mathbb{D}_{\mathbb{R}}[0,\infty)$. Therefore, it is just remained to show that $\{\overline{\mathcal{R}}^{(N)}_{\cdot,\ell}\}_{N\in\mathbb{N}}$ satisfies the compact containment condition J1. We first claim that for every $\ell\ge 1$

$$\lim_{m\to L}\sup_N \mathbb{E}\left[\overline{\mathcal{R}}^{(N)}_{\mathbf{1}_{(m,L)},\ell}(T)\right]=0. \tag{7.1.26}$$

For $m>0$, substituting $\varphi(x,s)=\mathbf{1}_{(m,L)}(x)$ in definition (5.1.19) of $\mathcal{R}^{(N)}_{\varphi,1}$, and since $\mathbf{1}_{(m,L)}(0)=0$, $\overline{\mathcal{R}}^{(N)}_{\mathbf{1}_{(m,L)},1}\equiv 0$, and (7.1.26) is immediate for $\ell=1$. Now fix $\ell\ge 2$ and

$T < \infty$. Substituting $\varphi(x,s) = \mathbf{1}_{(m,L)}(x)$ in representation (5.2.28) of $\mathcal{R}^{(N)}_{\varphi,\ell}$, we have

(7.1.27)
$$\mathcal{R}^{(N)}_{\mathbf{1}_{(m,L)},\ell}(T) = \sum_{j\geq j_0}\sum_{j'\geq 1} \mathbb{1}\left(\alpha^{(N)}_{j'} \leq t\right)\mathbb{1}\left(a^{(N)}_j(\alpha^{(N)}_{j'}) > m\right)\mathbb{1}\left(\beta^{(N)}_j < \alpha^{(N)}_{j'} \leq \gamma^{(N)}_j\right)$$
$$\mathbb{1}\left(\chi^{(N)}_j(\alpha^{(N)}_{j'}-) = \ell - 1\right)\mathbb{1}\left(\kappa^{(N)}_{j'} = \kappa^{(N)}_j\right).$$

For every $j \geq j_0$, consider the (random) set

$$\mathcal{J}^{(N)}_{j,\ell}(T) \doteq \left\{j' \geq 1\ : \alpha^{(N)}_{j'} \leq t, \alpha^{(N)}_{j'} \in [\beta^{(N)}_j, \gamma^{(N)}_j), \chi^{(N)}_j(\alpha^{(N)}_{j'}-) = \ell - 1, \kappa^{(N)}_{j'} = \kappa^{(N)}_j\right\},$$

of jobs j' that arrive while job j is receiving service, and is being routed to the same queue as j and has length $\ell - 1$ right before the arrival of j'. When $\mathcal{J}^{(N)}_{j,\ell}(T) \neq \emptyset$, define $j_* = j_*(j,\ell,N,T) \doteq \min \mathcal{J}^{(N)}_{j,\ell}(T)$ to be such a job with the smallest index. During the interval $[\alpha^{(N)}_{j_*}, \gamma^{(N)}_j)$, there are no departures from the queue $\kappa^{(N)}_j$ and hence the length of this queue is at least ℓ during that interval. and hence, $\mathcal{J}^{(N)}_{j,\ell}(T)$ is a singleton $\{j_*\}$. Also, for $j > E^{(N)}(T)$, $\alpha^{(N)}_j > T$ and thus, $\mathcal{J}^{(N)}_{j,\ell}(T) = \emptyset$. Therefore, (7.1.27) can be written as

$$\mathcal{R}^{(N)}_{\mathbf{1}_{(m,L)},\ell}(T) = \sum_{j\geq j_0}\sum_{j'\in\mathcal{J}^{(N)}_{j,\ell}(T)} \mathbb{1}\left(a^{(N)}_j(\alpha^{(N)}_{j'}) > m\right)$$
$$(7.1.28)\qquad = \sum_{j=j_0}^{E^{(N)}(T)} \mathbb{1}\left(\mathcal{J}^{(N)}_{j,\ell}(T) \neq \emptyset\right)\mathbb{1}\left(a^{(N)}_j(\alpha^{(N)}_{j_*}) > m\right).$$

Now we consider two possible cases depending on the value of L. If $L = \infty$ and a job j has initial age $a^{(N)}_j(0) \leq m$, then $a^{(N)}_j(t) \leq m + T$ for all $t \in [0,T]$. Therefore, (7.1.28) implies

$$(7.1.29)\quad \mathbb{E}\left[\overline{\mathcal{R}}^{(N)}_{\mathbf{1}_{(m+T,L)},\ell}(T)\right] \leq \mathbb{E}\left[\frac{1}{N}\sum_{j\geq j_0}\mathbb{1}\left(a^{(N)}_j(0) > m\right)\right] = \mathbb{E}\left[\overline{\nu}^{(N)}_1(0,(m,L))\right].$$

On the other hand, when $L < \infty$, using (7.1.28) and the fact that the age $a^{(N)}_j(\cdot)$ of each job j is non-decreasing and bounded by the total service time requirement v_j

of that job, we have

$$\mathcal{R}^{(N)}_{\mathbf{1}_{(m,L)},\ell}(T) \le \sum_{j=j_0}^{0} \mathbb{1}\left(a_j^{(N)}(0) > m\right) + \sum_{j=j_0}^{0} \mathbb{1}\left(a_j^{(N)}(0) \le m\right) \mathbb{1}\left(v_j > m\right) + \sum_{j=1}^{E^{(N)}(T)} \mathbb{1}\left(v_j > m\right). \tag{7.1.30}$$

For the second term on the right-hand side above, conditioning on $\mathcal{F}_0^{(N)}$ and using the fact that j_0, $a_j^{(N)}(0)$, and $\nu_1^{(N)}(0)$ are $\mathcal{F}_0^{(N)}$-measurable, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{j=j_0}^{0} \mathbb{1}\left(a_j^{(N)}(0) \le m\right) \mathbb{1}\left(v_j > m\right)\right] &= \mathbb{E}\left[\sum_{j=j_0}^{0} \mathbb{1}\left(a_j^{(N)}(0) \le m\right) \mathbb{E}\left[\mathbb{1}\left(v_j > m\right) | \mathcal{F}_0^{(N)}\right]\right] \\
&= \mathbb{E}\left[\sum_{j=j_0}^{0} \mathbb{1}\left(a_j^{(N)}(0) \le m\right) \frac{\overline{G}(m)}{\overline{G}(a_j^{(N)}(0))}\right] \\
&= \mathbb{E}\left[\int_{[0,m)} \frac{\overline{G}(m)}{\overline{G}(x)} \nu_1^{(N)}(0, dx)\right].
\end{aligned}$$

Therefore, taking expectation in (7.1.30) and using the independence of the initial conditions and the arrival process for the third term on the right-hand side, we have

$$\mathbb{E}\left[\overline{\mathcal{R}}^{(N)}_{\mathbf{1}_{(m,L)},\ell}(T)\right] \le \mathbb{E}\left[\overline{\nu}_1^{(N)}(0, (m, L))\right] + \mathbb{E}\left[\int_{[0,m)} \frac{\overline{G}(m)}{\overline{G}(x)} \overline{\nu}_1^{(N)}(0, dx)\right] + \overline{G}(m)\mathbb{E}\left[\overline{E}^{(N)}(T)\right]. \tag{7.1.31}$$

Taking first the supremum over N and then the limit as $m \to L$ of both sides of (7.1.29) and (7.1.31), the claim (7.1.26) for both cases follows by using (7.1.13) and (7.1.14) in Lemma 7.12, the bound (7.1.4), and the the elementary identity $\lim_{m \to L} \overline{G}(m) = 0$.

The rest of the proof is similar to the proof of Lemma 5.13 in **[61]**. Fix $\eta > 0$ and note that by the bound (7.1.8) with $\varphi = \mathbf{1}$ and $t = T$,

$$\sup_N \mathbb{E}\left[\overline{\mathcal{R}}^{(N)}_{\mathbf{1},\ell}(T)\right] < \infty,$$

and hence the constant

$$C(\eta, T) \doteq \frac{2}{\eta} \sup_N \mathbb{E}\left[\overline{\mathcal{R}}^{(N)}_{\mathbf{1},\ell}(T)\right] \tag{7.1.32}$$

is finite. Also by (7.1.26), there exists a sequence $\{m(n,\eta)\}$ with $m(n,\eta) \to L$ as $n \to \infty$ such that

$$\sup_N \mathbb{E}\left[\overline{\mathcal{R}}^{(N)}_{\mathbf{1}_{(m(n,\eta),L)},\ell}(T)\right] \leq \frac{\eta}{n2^{n+1}}. \tag{7.1.33}$$

The subset $\mathcal{K}_{\eta,T}$ of $\mathbb{M}_F([0,L) \times \mathbb{R}_+)$ defined as

$$\mathcal{K}_{\eta,T} \doteq \left\{\mu \in \mathbb{M}_F([0,L) \times \mathbb{R}_+) \ : \langle \mathbf{1}, \mu \rangle \leq C(\eta, T),\ \mu((m(n,\eta), L) \times R_+) \leq \frac{1}{n}\ \forall n \in \mathbb{N}\right\},$$

is compact because $\sup_{\mu \in \mathcal{K}_{\eta,T}} \ \mu([0,L) \times \mathbb{R}_+) \leq C(\eta, T)$ and

$$\inf_{\substack{C \subset [0,L) \times \mathbb{R}_+ \\ C \text{ compact}}} \ \sup_{\mu \in \mathcal{K}_{\eta,T}} \mu(C^c) \leq \inf_n \sup_{\mu \in \mathcal{K}_{\eta,T}} \mu\big((m(n,\eta), L) \times R_+\big) = 0.$$

Finally, by definition of $\mathcal{K}_{\eta,T}$ and using (7.1.32) and (7.1.33), we have

$$\begin{aligned}
&\mathbb{P}\left\{\overline{\mathcal{R}}^{(N)}_{\cdot,\ell}(t) \notin \mathcal{K}_\eta \text{ for some } t \in [0,T]\right\} \\
&\qquad \leq \mathbb{P}\left\{\overline{\mathcal{R}}^{(N)}_{\mathbf{1},\ell}(T) \geq C(\eta)\right\} + \sum_{n \in \mathbb{N}} \mathbb{P}\left\{\overline{\mathcal{R}}^{(N)}_{\mathbf{1}_{(m(n,\eta),L)},\ell}(T) > \frac{1}{n}\right\} \\
&\qquad \leq \sup_N \frac{\mathbb{E}\left[\overline{\mathcal{R}}^{(N)}_{\mathbf{1},\ell}(T)\right]}{C(\eta)} + \sum_{n \in \mathbb{N}} n\mathbb{E}\left[\overline{\mathcal{R}}^{(N)}_{\mathbf{1}_{(m(n,\eta),L)},\ell}(T)\right] \\
&\qquad \leq \eta.
\end{aligned}$$

This shows that $\{\overline{\mathcal{R}}^{(N)}_{\cdot,\ell}\}_{N \in \mathbb{N}}$ satisfies the compact containment condition J1, and hence, completes the proof. $\square$

LEMMA 7.19. Suppose Assumptions I, II.a, and III hold. Then, for every $\ell \geq 1$, the sequence $\{\overline{\mathcal{D}}^{(N)}_{\cdot,\ell}\}$ is tight in $\mathbb{D}_{\mathbb{M}_F([0,L)\times\mathbb{R}_+)}[0,\infty)$.

PROOF. For $\ell = 1$, the compact containment condition J1 follows Lemma 7.13. For $\ell \geq 2$, since for every non-negative measurable function φ and every $t \geq 0$, $\mathcal{D}^{(N)}_{\varphi,\ell}(t) \leq \mathcal{D}^{(N)}_{\varphi,1}(t)$, for every $\eta > 0$ and $T \geq 0$ the bound (7.1.17) holds with $\mathcal{D}^{(N)}_{\varphi,1}$ replaced by $\mathcal{D}^{(N)}_{\varphi,\ell}$ as well for the same compact set $\mathcal{K}_\eta$ defined in (7.1.18), and thus the condition J1 holds for $\overline{\mathcal{D}}^{(N)}_{\cdot,\ell}, \ell \geq 2$. The condition J2 for $\ell \geq 1$ is proved in Lemma 7.16. $\square$

Below, we summarize the results of this section.

THEOREM 7.20. Suppose Assumptions I, II.a, and III hold. Then, for every $\ell \geq 1$, the sequence

$$\{(\overline{\nu}^{(N)}, \overline{\mathcal{D}}^{(N)}_{\cdot,\ell}, \overline{\mathcal{R}}^{(N)}_{\cdot,\ell})\}_{N\in\mathbb{N}}$$

is tight in $\mathbb{D}_{\mathbb{S}}[0,\infty) \times \mathbb{D}_{\mathbb{M}_F([0,L)\times\mathbb{R}_+)}[0,\infty)^2$.

PROOF. By Proposition 7.17, for every $\ell \geq 1$, $\{\overline{\nu}^{(N)}_{\ell}\}_{N\in\mathbb{N}}$ is tight in $\mathbb{M}_F[0,L)$. Therefore, by a diagonalization argument, every subsequence of $\{\overline{\nu}^{(N)}\}_{N\in\mathbb{N}} = \{(\overline{\nu}^{(N)}_{\ell}; \ell \geq 1)\}_{N\in\mathbb{N}}$ has a further subsequence that is convergent, simultaneously for all $\ell \geq 1$. This means that the sequence $\{\overline{\nu}^{(N)}\}_{N\in\mathbb{N}} \subset \mathbb{S}$ is relatively compact with respect to the norm $d_{\mathbb{S}}$ defined in (5.1.8). Together with Lemmas 7.19 and 7.18, this proves the result. $\square$

7.2. Convergence of a Measure-Valued Representation

We have proved in Theorem 7.20 that the sequence $\{\nu^{(N)}\}_{N\in\mathbb{N}}$ is tight in $\mathbb{D}_{\mathbb{S}}[0,\infty)$, and hence, every subsequence has a further convergent subsequence. In this section, we show that all such subsequential limits are solutions to hydrodynamic equations defined in Definition 6.1. Since the these equations have a unique solution by Theorem 6.3, this provides a unique characterization for all subsequential limits, and hence, completes the proof of convergence. The limit of the routing process is characterized in Section 7.2.1, and limit of the departure process is characterized in Section 7.2.2. Finally, Theorem 7.1 is proved in Section 7.2.3.

7.2.1. Limit of the Routing Process. Recall the constant λ from Assumption I. To establish the limit of the routing process $\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}$, we will need the following preliminary result.

LEMMA 7.21. Suppose Assumption I holds.

(1) For every $0 \leq s \leq t$, the following limit holds in probability as $N \to \infty$:

$$\frac{1}{N} \int_s^t h_E^{(N)}(R^{(N)}(u))du \to (t-s)\lambda; \tag{7.2.1}$$

(2) For every $t \geq 0$,

$$C_E(t) \doteq \limsup_{N\to\infty} \mathbb{E}\left[\left(\frac{1}{N}\int_0^t h_E^{(N)}(R^{(N)}(u))du\right)^2\right] < \infty. \tag{7.2.2}$$

PROOF. By Assumption I, $E^{(N)}$ is a pure renewal process with inter-arrival distribution $G_E^{(N)}$ and backward recurrence time $R^{(N)}$. Applying Lemma 7.30 with P, G, h, r and r_0 replaced by $E^{(N)}$, $G_E^{(N)}$, $h_E^{(N)}$, $R^{(N)}$, and $R^{(N)}$, respectively, it follows from (7.A.3) and an application of Chebychev's inequality that for every $t \geq 0$,

$$\begin{aligned}
&\mathbb{P}\left\{\left|\frac{1}{N}\int_0^t h_E^{(N)}(R^{(N)}(s))ds - \lambda t\right| > 2\epsilon\right\} \\
&\qquad \leq \frac{1}{\epsilon^2}\mathbb{E}\left[\left(\frac{1}{N}\int_0^t h_E^{(N)}(R^{(N)}(s))ds - \overline{E}^{(N)}(t)\right)^2\right] + \mathbb{P}\left\{\left|\overline{E}^{(N)}(t) - \lambda t\right| > \epsilon\right\} \\
&\qquad \leq \frac{1}{N\epsilon^2}\left(\frac{4}{N} + \mathbb{E}\left[\overline{E}^{(N)}(t)\right]\right) + \mathbb{P}\left\{\left|\overline{E}^{(N)}(t) - \lambda t\right| > \epsilon\right\}
\end{aligned}$$

By (7.1.4), $\mathbb{E}[\overline{E}^{(N)}(t)]$ is finite, and $\overline{E}^{(N)}(t)$ converges in expectation, and hence in probability, to λt by Lemma 7.6. Hence, the right-hand side of display above converges to zero as $N \to \infty$. Since both sides of (7.2.1) are linear in t, (7.2.1) follows.

To establish (7.2.2), by another application of (7.A.3), we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{N}\int_0^t h_E^{(N)}(R^{(N)}(s))ds\right)^2\right] \\
&\qquad\qquad \leq \frac{2}{N^2}\mathbb{E}\left[\left(\int_0^t h_E^{(N)}(R^{(N)}(s))ds - E^{(N)}(t)\right)^2\right] + 2\mathbb{E}\left[\overline{E}^{(N)}(t)^2\right] \\
&\qquad\qquad \leq \frac{8}{N^2} + \frac{2}{N}\mathbb{E}\left[\overline{E}^{(N)}(t)\right] + 2\mathbb{E}\left[\overline{E}^{(N)}(t)^2\right],
\end{aligned}$$

Taking the limit superior as $N \to \infty$ from both sides of the inequality above and using the bound (7.1.4), (7.2.2) follows for

$$C_E(t) = 2\limsup_{N\to\infty} \mathbb{E}\left[\overline{E}^{(N)}(t)^2\right],$$

which is finite by (7.1.5). $\square$

PROPOSITION 7.22. Suppose Assumptions I, II.a, and III hold, and fix $\ell \geq 1$. If $(\overline{\mathcal{R}}^{(N)}_{\cdot,\ell}, \overline{\nu}^{(N)})$ converges to $(\mathcal{R}_{\cdot,\ell}, \nu)$ in $\mathbb{D}_{\mathbb{M}_F([0,L)\times[0,\infty))}[0,\infty) \times \mathbb{D}_{\mathbb{S}}[0,\infty)$, almost surely as $N \to \infty$, then for every $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$, $\mathcal{R}_{\varphi,\ell}$ is continuous and for every $t \geq 0$,

$$\mathcal{R}_{\varphi,\ell}(t) = \int_0^t \langle \varphi(\cdot, s), \eta_\ell(s) \rangle ds, \qquad \text{a.s.,} \tag{7.2.3}$$

where

$$\eta_\ell(t) \doteq \begin{cases} \lambda \left(1 - \langle \mathbf{1}, \nu_1(t) \rangle^d\right) \delta_0 & \text{if } \ell = 1, \\ \lambda \mathfrak{P}_d\Big(\langle \mathbf{1}, \nu_{\ell-1}(t) \rangle, \langle \mathbf{1}, \nu_\ell(t) \rangle\Big) (\nu_{\ell-1}(t) - \nu_\ell(t)) & \text{if } \ell \geq 2. \end{cases} \tag{7.2.4}$$

PROOF. Fix $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$. For every $\epsilon > 0$, using the relation $\overline{\mathcal{R}}^{(N)}_{\varphi,\ell} = \overline{B}^{(N)}_{\varphi,\ell} + \overline{\mathcal{N}}^{(N)}_{\varphi,\ell}$ in (5.3.3), we have

$$\begin{aligned} \mathbb{P}\left\{ \left| \mathcal{R}_{\varphi,\ell}(t) - \int_0^t \langle \varphi(\cdot,s), \eta_\ell(s) \rangle ds \right| > 3\epsilon \right\} \\ \leq \mathbb{P}\left\{ \left| \mathcal{R}_{\varphi,\ell}(t) - \overline{\mathcal{R}}^{(N)}_{\varphi,\ell}(t) \right| > \epsilon \right\} + \mathbb{P}\left\{ \left| \overline{\mathcal{N}}^{(N)}_{\varphi,\ell}(t) \right| > \epsilon \right\} \\ + \mathbb{P}\left\{ \left| \overline{B}^{(N)}_{\varphi,\ell}(t) - \int_0^t \langle \varphi(\cdot,s), \eta_\ell(s) \rangle ds \right| > \epsilon \right\}. \end{aligned} \tag{7.2.5}$$

By the hypothesis of this proposition, $\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}$ converges almost surely to $\mathcal{R}_{\varphi,\ell}$, in $\mathbb{D}_{\mathbb{R}}[0,\infty)$. Also, by (7.1.7), the jump sizes of $\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}$ are bounded by $\|\varphi\|_\infty$ times the jump sizes of $\overline{E}^{(N)}$, which are at most $1/N$. Therefore, the maximum jump size of $\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}$ converges to zero as $N \to \infty$, and by [**8**, Theorem 13.4], $\mathcal{R}_{\varphi,\ell}$ is almost surely continuous and $\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}$ converges to $\mathcal{R}_{\varphi,\ell}$, almost surely, uniformly on compact sets. In particular, for every $t \geq 0$, $\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}(t)$ converges almost surely, and hence in probability, to $\mathcal{R}_{\varphi,\ell}(t)$, that is,

$$\limsup_{N\to\infty} \mathbb{P}\left\{ \left| \mathcal{R}_{\varphi,\ell}(t) - \overline{\mathcal{R}}^{(N)}_{\varphi,\ell}(t) \right| > \epsilon \right\} = 0. \tag{7.2.6}$$

Moreover, by definitions (5.3.2) and (7.2.4) of $B^{(N)}_{\varphi,\ell}$ and η_ℓ, we can write

(7.2.7)

$$\begin{aligned}\overline{B}^{(N)}_{\varphi,\ell}(t) - \int_0^t \langle \varphi(\cdot,s), \eta_\ell(s)\rangle ds &= \int_0^t \left(\frac{1}{N} h^{(N)}_E(R^{(N)}(u)) - \lambda\right) f_{\varphi,\ell}(u) du \\ &\quad + \frac{1}{N}\int_0^t h^{(N)}_E(R^{(N)}(u))(f^{(N)}_{\varphi,\ell}(u) - f_{\varphi,\ell}(u))du\end{aligned}$$

with

$$f^{(N)}_{\varphi,\ell}(u) \doteq \begin{cases} \varphi(0,u)(1-\overline{S}^{(N)}_1(u)^d) & \text{if } \ell = 1, \\ \left\langle \varphi(\cdot,u), \overline{\nu}^{(N)}_{\ell-1}(u) - \overline{\nu}^{(N)}_{\ell}(u)\right\rangle \mathfrak{P}_d\left(\overline{S}^{(N)}_{\ell-1}(u), \overline{S}^{(N)}_{\ell}(u)\right) & \text{if } \ell \geq 2, \end{cases}$$

and

$$f_{\varphi,\ell}(u) \doteq \begin{cases} \varphi(0,u)(1-S_1(u)^d) & \text{if } \ell = 1, \\ \langle \varphi(\cdot,u), \nu_{\ell-1}(u) - \nu_\ell(u)\rangle \mathfrak{P}_d\left(S_{\ell-1}(u), S_\ell(u)\right) & \text{if } \ell \geq 2, \end{cases}$$

and recall $S^{(N)}_\ell = \langle \mathbf{1}, \nu^{(N)}_\ell\rangle$ and likewise, $S_\ell = \langle \mathbf{1}, \nu_\ell\rangle$. By the hypothesis of this proposition, almost surely, $\overline{\nu}^{(N)}_\ell$ converges weakly to ν_ℓ, and since $\mathbf{1}$ and φ are bounded continuous functions, almost surely, $\overline{S}^{(N)}_\ell$ and $\langle \varphi, \overline{\nu}^{(N)}_{\ell-1} - \overline{\nu}^{(N)}_\ell\rangle$ converge to S_ℓ and $\langle \varphi, \nu_{\ell-1} - \nu_\ell\rangle$, respectively, in D. By Lemma 5.15, there is at most one arrival or departure at each time, and hence, the maximum jump sizes of $\overline{S}^{(N)}_\ell$ and $\langle \varphi, \overline{\nu}^{(N)}_{\ell-1} - \overline{\nu}^{(N)}_\ell\rangle$ is bounded by $1/N$ and $\|\varphi\|_\infty/N$, respectively. Therefore, by [**8**, Theorem 13.4], S_ℓ and $\langle \varphi, \nu_{\ell-1} - \nu_\ell\rangle$ are continuous and aforementioned convergence is uniformly on compact sets, almost surely. Consequently, $f_{\varphi,\ell}$ is continuous, and $f^{(N)}_{\varphi,\ell}$ converges to $f_{\varphi,\ell}$ uniformly on compact sets, almost surely.

For $t \geq 0$ and $\delta > 0$, let $w_{f_{\varphi,\ell}}(\delta, t)$ denote the modulus of continuity of $f_{\varphi,\ell}$ (see Section 1.4), and hence, for every partition $\mathcal{P} = \{0 = t_0 < t_1 < ... < t_n = t\}$ of size δ,

$$\sup\{|f_{\varphi,\ell}(t_j) - f_{\varphi,\ell}(u)|;\ u \in (t_j, t_{j+1}), j = 0, ..., n-1\} \leq w_{f_{\varphi,\ell}}(\delta, t).$$

Therefore,

$$\mathbb{P}\left\{\left|\int_0^t \left(\frac{1}{N}h_E^{(N)}(R^{(N)}(u))-\lambda\right) f_{\varphi,\ell}(u)du\right| > 2\epsilon\right\}$$
$$\leq \mathbb{P}\left\{\sum_{j=0}^{n-1} f_{\varphi,\ell}(t_j)\left|\frac{1}{N}\int_{t_j}^{t_{j+1}} h_E^{(N)}(R^{(N)}(u))du - (t_{j+1}-t_j)\lambda\right| > \epsilon\right\}$$
$$+\,\mathbb{P}\left\{w_{f_{\varphi,\ell}}(\delta,t)\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}\left|\frac{1}{N}h_E^{(N)}(R^{(N)}(u))-\lambda\right| du > \epsilon\right\}. \tag{7.2.8}$$

By (7.2.1), we see that the first term on the right-hand side of (7.2.8) satisfies

$$\limsup_{N\to\infty}\mathbb{P}\left\{\sum_{j=0}^{n-1} f_{\varphi,\ell}(t_j)\left|\frac{1}{N}\int_{t_j}^{t_{j+1}} h_E^{(N)}(R^{(N)}(u))du - (t_{j+1}-t_j)\lambda\right| > \epsilon\right\} = 0. \tag{7.2.9}$$

For the second term, using the Markov and Cauchy-Schwartz inequalities, and (7.2.2)

$$\mathbb{P}\left\{w_{f_{\varphi,\ell}}(\delta,t)\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}\left|\frac{1}{N}h_E^{(N)}(R^{(N)}(u))-\lambda\right| du > \epsilon\right\}$$
$$\leq \frac{1}{\epsilon}\mathbb{E}\left[w_{f_{\varphi,\ell}}(\delta,t)\left(\frac{1}{N}\int_0^t h_E^{(N)}(R^{(N)}(u))du+\lambda t\right)\right]$$
$$\leq \frac{2}{\epsilon}\mathbb{E}\left[w^2_{f_{\varphi,\ell}}(\delta,t)\right]^{\frac{1}{2}}\left(\mathbb{E}\left[\left(\frac{1}{N}\int_0^t h_E^{(N)}(R^{(N)}(u))du\right)^2\right]+\lambda^2t^2\right)^{\frac{1}{2}}.$$

Taking limit superior on both sides of the display above, using the bound (7.2.2) and substitute with (7.2.9) in (7.2.8), we conclude

$$\limsup_{N\to\infty}\mathbb{P}\left\{\left|\int_0^t \left(\frac{1}{N}h_E^{(N)}(R^{(N)}(u))-\lambda\right) f_{\varphi,\ell}(u)du\right| > 2\epsilon\right\} \leq \frac{C(t)}{\epsilon}\mathbb{E}\left[w^2_{f_{\varphi,\ell}}(\delta,t)\right],$$

where $C(t) \doteq 2(C_E(t)+\lambda^2t^2)^{1/2}$. Since $f_{\varphi,\ell}$ is bounded and continuous, $w^2_{f_{\varphi,\ell}}(\delta,t)$ is bounded and converges almost surely to zero as $\delta\to 0$. Sending $\delta\to 0$ in the last display and applying the bounded convergence theorem, we see that for every $\epsilon>0$,

$$\limsup_{N\to\infty}\mathbb{P}\left\{\left|\int_0^t \left(\frac{1}{N}h_E^{(N)}(R^{(N)}(u))-\lambda\right) f_{\varphi,\ell}(u)du\right| > 2\epsilon\right\} = 0. \tag{7.2.10}$$

In a similar fashion, for every $N \in \mathbb{N}$, by the Markov and Cauchy-Schwartz inequalities, we have

$$\mathbb{P}\left\{\left|\frac{1}{N}\int_0^t h_E^{(N)}(R^{(N)}(u))(f_{\varphi,\ell}^{(N)}(u) - f_{\varphi,\ell}(u))du\right| > \epsilon\right\}$$
$$\leq \frac{1}{\epsilon}\mathbb{E}\left[\sup_{u\in[0,t]}\left|f_{\varphi,\ell}^{(N)}(u) - f_{\varphi,\ell}(u)\right|^2\right]^{\frac{1}{2}}\mathbb{E}\left[\left(\frac{1}{N}\int_0^t h_E^{(N)}(R^{(N)}(u))du\right)^2\right]^{\frac{1}{2}}.$$

Taking the limit superior as $N \to \infty$ of both sides of the last display and using (7.2.2), we have

$$\limsup_{N\to\infty}\mathbb{P}\left\{\left|\frac{1}{N}\int_0^t h_E^{(N)}(R^{(N)}(u))(f_{\varphi,\ell}^{(N)}(u) - f_{\varphi,\ell}(u))du\right| > \epsilon\right\}$$
$$= \frac{1}{\epsilon}C_E(t)^{\frac{1}{2}}\limsup_{N\to\infty}\mathbb{E}\left[\|f_{\varphi,\ell}^{(N)} - f_{\varphi,\ell}\|_t^2\right]^{\frac{1}{2}}.$$

Since $f_{\varphi,\ell}^{(N)}$ converges to $f_{\varphi,\ell}$ uniformly on compact sets, almost surely, and $f_{\varphi,\ell}^{(N)}$ and $f_{\varphi,\ell}$ are uniformly bounded by $2\|\varphi\|_\infty$, the right-hand side of the display above is zero by the bounded convergence theorem, and hence,

$$\limsup_{N\to\infty}\mathbb{P}\left\{\left|\frac{1}{N}\int_0^t h_E^{(N)}(R^{(N)}(u))(f_{\varphi,\ell}^{(N)}(u) - f_{\varphi,\ell}(u))du\right| > \epsilon\right\} = 0. \tag{7.2.11}$$

We conclude from (7.2.7), (7.2.10) and (7.2.11) that

$$\limsup_{N\to\infty}\mathbb{P}\left\{\left|\overline{B}_{\varphi,\ell}^{(N)}(t) - \int_0^t \langle\varphi(\cdot,s), \eta_\ell(s)\rangle ds\right| > \epsilon\right\} = 0. \tag{7.2.12}$$

Finally, (7.2.3) follows on sending first $N \to \infty$ on the right-hand side of (7.2.5), next incorporating relations (7.2.6), (7.1.9) of Lemma 7.9 and (7.2.12), and then sending $\epsilon \downarrow 0$. □

7.2.2. Limit of Departure Process.

PROPOSITION 7.23. Suppose Assumptions I-III hold, and fix $\ell \geq 1$. If $(\overline{\mathcal{D}}_{\cdot,\ell}^{(N)}, \overline{\nu}^{(N)})$ converges to $(\mathcal{D}_{\cdot,\ell}, \nu)$ in $\mathbb{D}_{\mathbb{M}_F([0,L)\times[0,\infty))}[0,\infty) \times \mathbb{D}_{\mathbb{S}}[0,\infty)$ almost surely, then, for every $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$, $\mathcal{D}_{\varphi,\ell}$ is continuous and for every $t \geq 0$,

$$\mathcal{D}_{\varphi,\ell}(t) = \int_0^t \langle\varphi(\cdot,s)h(\cdot), \nu_\ell(s)\rangle ds, \qquad \text{a.s.} \tag{7.2.13}$$

PROOF. Fix $t \geq 0$ and $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$. For every $N \in \mathbb{N}$, using (5.3.7) we can write

$$\begin{aligned}\mathcal{D}_{\varphi,\ell}(t) - \int_0^t \langle \varphi(\cdot,s)h(\cdot), \nu_\ell(s)\rangle ds =& \mathcal{D}_{\varphi,\ell}(t) - \overline{\mathcal{D}}^{(N)}_{\varphi,\ell}(t) + \overline{\mathcal{M}}^{(N)}_{\varphi,\ell}(t) \\ &+ \overline{A}^{(N)}_{\varphi,\ell}(t) - \int_0^t \langle \varphi(\cdot,s)h(\cdot), \nu_\ell(s)\rangle ds.\end{aligned} \tag{7.2.14}$$

By the hypothesis of this proposition, $\overline{\mathcal{D}}^{(N)}_{\varphi,\ell}$ converges to $\mathcal{D}_{\varphi,\ell}$ in $\mathbb{D}_{\mathbb{R}}[0,\infty)$, almost surely, as $N \to \infty$. Moreover, by Lemma 5.15, only one departure can occur at each time, and hence the jump size of $\overline{\mathcal{D}}^{(N)}_{\varphi,\ell}$ is bounded by $\|\varphi\|_\infty$. Therefore, by [**8**, Theorem 13.4], $\mathcal{D}_{\varphi,\ell}$ is continuous and the convergence also holds uniformly on compact sets, almost surely. In particular, $\overline{\mathcal{D}}^{(N)}_{\varphi,\ell}(t)$ converges $\mathcal{D}_{\varphi,\ell}(t)$ almost surely, and hence, in probability, that is,

$$\limsup_{N\to\infty} \mathbb{P}\left\{ \left| \mathcal{D}_{\varphi,\ell}(t) - \overline{\mathcal{D}}^{(N)}_{\varphi,\ell}(t) \right| > \epsilon \right\} = 0. \tag{7.2.15}$$

Furthermore, we claim that

$$\limsup_{N\to\infty} \mathbb{E}\left[\left| \overline{A}^{(N)}_{\varphi,\ell}(t) - \int_0^t \langle \varphi(\cdot,s)h(\cdot), \nu_\ell(s)\rangle ds \right| \right] = 0. \tag{7.2.16}$$

For $\ell = 1$, (7.2.16) follows from parts a. and b. of Lemma 7.14. Also, for $\ell \geq 2$ Since $\overline{\nu}^{(N)}(t)$ and $\nu(t)$ both lie in $\mathbb{S}$, for every $N \in \mathbb{N}$ and $t \geq 0$,

$$\langle f, \nu^{(N)}_\ell(t)\rangle \leq \langle f, \nu^{(N)}_1(t)\rangle, \qquad \langle f, \nu_\ell(t)\rangle \leq \langle f, \nu_1(t)\rangle, \quad \forall \ell \geq 1,$$

for every non-negative measurable function f. Thus, the bounds (7.1.19)-(7.1.21) hold with $\overline{\nu}^{(N)}_1$ and ν_1 replaced by $\overline{\nu}^{(N)}_\ell$ and ν_ℓ, respectively. Part b. of Lemma 7.14 shows that (7.2.16) holds for all $\ell \geq 1$, which proves the claim.

We conclude by noting that using (7.2.14), for every $\epsilon > 0$,

$$\begin{aligned}&\mathbb{P}\left\{ \left| \mathcal{D}_{\varphi,\ell}(t) - \int_0^t \langle \varphi(\cdot,s)h(\cdot), \nu_\ell(s)\rangle ds \right| > \epsilon \right\} \\ &\qquad \leq \mathbb{P}\left\{ \left| \mathcal{D}_{\varphi,\ell}(t) - \overline{\mathcal{D}}^{(N)}_{\varphi,\ell}(t) \right| > \epsilon \right\} + \mathbb{P}\left\{ \left| \overline{\mathcal{M}}^{(N)}_{\varphi,\ell}(t) \right| > \epsilon \right\} \\ &\qquad\quad + \mathbb{P}\left\{ \left| \overline{A}^{(N)}_{\varphi,\ell}(t) - \int_0^t \langle \varphi(\cdot,s)h(\cdot), \nu_\ell(s)\rangle ds \right| > \epsilon \right\}.\end{aligned}$$

The result follows on taking limit superior as $N \to \infty$ from both sides of the display above, using (7.2.15), (7.1.24) of Lemma 7.15, and (7.2.16) and then sending $\epsilon \to 0$. $\square$

7.2.3. Proof of Theorem 7.1.

PROOF OF THEOREM 7.1. By Theorem 7.20, Lemma 7.19, Lemma 7.18, and Assumption III, the sequence $\{Y^{(N)}\}_{N\in\mathbb{N}}$, defined as

$$Y^{(N)} \doteq \left(\overline{\nu}^{(N)}(0), \overline{E}^{(N)}, \overline{\nu}^{(N)}, \overline{\mathcal{D}}^{(N)}_{\cdot,\ell}, \overline{\mathcal{R}}^{(N)}_{\cdot,\ell}; \ell \geq 1\right), \tag{7.2.17}$$

is tight in

$$\mathcal{Y} \doteq \mathbb{S} \times \mathbb{D}_{\mathbb{R}}[0,\infty) \times \mathbb{D}_{\mathbb{S}}[0,\infty) \times \mathbb{D}_{\mathbb{M}_F([0,L)\times\mathbb{R}_+)}[0,\infty)^{\mathbb{N}_0} \times \mathbb{D}_{\mathbb{M}_F([0,L)\times\mathbb{R}_+)}[0,\infty)^{\mathbb{N}_0}. \tag{7.2.18}$$

Therefore, every subsequence of $\{Y^{(N)}\}$ has a further converging subsequence $\{N_k\}$, that is, there exists $\nu \in \mathbb{D}_{\mathbb{S}}[0,\infty)$ and $\mathcal{D}_{\cdot,\ell}, \mathcal{R}_{\cdot,\ell} \in \mathbb{D}_{\mathbb{M}_F([0,L)\times\mathbb{R}_+)}[0,\infty)$, $\ell \geq 1$, such that $Y^{(N)}$ converges in distribution to Y along $\{N_k\}$, where

$$Y \doteq (\nu(0), \lambda\mathbf{Id}, \nu, \mathcal{D}_{\cdot,\ell}, \mathcal{R}_{\cdot,\ell}; \ell \geq 1). \tag{7.2.19}$$

It follows from the Skorokhod representation theorem that there exists $\mathcal{Y}$-valued random elements $\widetilde{Y}^{(N)}$ and $\widetilde{Y}$, possibly on a different probability space, such that $\widetilde{Y}^{(N)} \overset{(d)}{=} Y^{(N)}$, $Y \overset{(d)}{=} \widetilde{Y}$, and $\widetilde{Y}^{(N)} \to \widetilde{Y}$ almost surely in $\mathcal{Y}$, along $\{N_k\}$. Since we are only interested in the distribution of $Y^{(N)}$, with a slight abuse of notation we identify $\widetilde{Y}^{(N)}$ and $\widetilde{Y}$ with $Y^{(N)}$ and Y, and denote the subsequence of $\{Y^{(N)}\}$ along $\{N_k\}$, again by $\{Y^{(N)}\}$. Using this convention, we have

$$Y^{(N)} \to Y, \qquad \text{in } \mathcal{Y}, \text{ a.s.} \tag{7.2.20}$$

Now we uniquely characterize the subsequential limit Y. For every $\ell \geq 1$ and $f \in \mathbb{C}_b[0,L)$, by (7.2.20), $\langle f, \overline{\nu}^{(N)}_\ell\rangle$ converge almost surely to $\langle f, \nu_\ell\rangle$ in $\mathbb{D}_{\mathbb{R}}[0,\infty)$. Since for every $N \in \mathbb{N}$, the maximum jump size of $\langle f, \overline{\nu}^{(N)}_\ell\rangle$ is bounded by $\|f\|_\infty/N$ (recall from Lemma 5.15 that there is at most one arrival or departure can occur at each time), the limit $\langle f, \nu_\ell \rangle$ is continuous, and hence ν is a continuous $\mathbb{S}$-valued process, almost surely.

Moreover, let $\mathcal{T}$ be a countable dense subset of $\mathbb{R}_+$ which contains 0 (say the diadic numbers). For every $\ell \geq 1$ and $t \in \mathcal{T}$, it follows from Proposition 7.23 with φ substituted by the constant function $\mathbf{1}$ that $\overline{D}_\ell^{(N)} = \overline{\mathcal{D}}_{\mathbf{1},\ell}^{(N)}$ converges almost surely in $\mathbb{D}_\mathbb{R}[0,\infty)$ to the continuous function $D_\ell \doteq \mathcal{D}_{\mathbf{1},\ell}$, where

$$\mathcal{D}_{\mathbf{1},\ell}^{(N)}(t) = \int_0^t \langle h, \nu_\ell(s) \rangle ds < \infty. \tag{7.2.21}$$

By Proposition 5.6, for every $\ell \geq 1$ and $N \in \mathbb{N}$, $Y^{(N)}$ satisfies (5.2.3). Taking the limit as $N \to \infty$ from both sides of (5.2.3) and using Proposition 7.22, for every $t \in \mathcal{T}$ and $\ell \geq 1$ the identity

$$\langle \mathbf{1}, \nu_\ell(t) \rangle - \langle \mathbf{1}, \nu_\ell(0) \rangle = D_{\ell+1}(t) + \int_0^t \langle \mathbf{1}, \eta_\ell(s) \rangle ds - D_\ell(t), \tag{7.2.22}$$

holds almost surely, where η is defined by (7.2.4). Moreover, the equalities (7.2.21) and (7.2.22) hold simultaneously for all $\ell \geq 1$ and $t \in \mathcal{T}$ because $\mathcal{T}$ is countable, and therefore for all $t \geq 0$ because both sides are continuous, by Proposition 7.22 and Proposition 7.23.

Furthermore, let $\mathcal{C}$ be a countable dense subset of $\mathbb{C}_c^{1,1}([0,L) \times \mathbb{R}_+)$, and fix $\varphi \in \mathcal{C}$ and $t \in \mathcal{T}$. By Proposition 5.6, for every $\ell \geq 1$ and $N \in \mathbb{N}$, $Y^{(N)}$ satisfies the equation (5.2.2). Since φ, φ_x and φ_s are all bounded continuous functions, $\langle \varphi(\cdot,t), \overline{\nu}_\ell^{(N)}(t) \rangle$ and $\int_0^t \langle \varphi_s(\cdot,s) + \varphi_x(\cdot,s), \overline{\nu}_\ell^{(N)}(s) \rangle ds$ converge almost surely to $\langle \varphi(\cdot,t), \nu_\ell(t) \rangle$ and $\int_0^t \langle \varphi_s(\cdot,s) + \varphi_x(\cdot,s), \nu_\ell(s) \rangle ds$, respectively. Also, as we have already shown, $\overline{D}_{\ell+1}^{(N)}$ converges to $D_{\ell+1}$ almost surely in $\mathbb{D}_\mathbb{R}[0,\infty)$, and therefore, the associated sequence of Stieltjes integrals $\int_{[0,t]} \varphi(0,s) d\overline{D}_{\ell+1}^{(N)}(s)$ converges to $\int_{[0,t]} \varphi(0,s) dD_{\ell+1}(s)$, almost surely. Finally, the $\overline{\mathcal{D}}_{\varphi,\ell}^{(N)}(t)$ converges to $\int_0^t \langle \varphi(\cdot,s) h(\cdot), \nu_\ell(s) \rangle ds$ by Proposition 7.22, and $\overline{\mathcal{R}}_{\varphi,\ell}^{(N)}(t)$ converge to $\int_0^t \langle \varphi(\cdot,s), \eta_\ell(s) \rangle ds$ by Proposition 7.22. Therefore, taking the limit as $N \to \infty$ of both sides of (5.2.2) we have

$$
\begin{aligned}
\langle \varphi(\cdot,t), \nu_\ell(t)\rangle =& \langle \varphi(\cdot,0), \nu_\ell(0)\rangle + \int_0^t \langle \varphi_x(\cdot,s) + \varphi_s(\cdot,s), \nu_\ell(s)\rangle ds \\
&- \int_0^t \langle \varphi(\cdot,s)h(\cdot), \nu_\ell(s)\rangle ds + \int_0^t \varphi(0,s) dD_{\ell+1}(s) \\
&+ \int_0^t \langle \varphi(\cdot,s), \eta_\ell(s)\rangle ds,
\end{aligned}
\tag{7.2.23}
$$

almost surely. The equation (7.2.23) above holds with probability one, simultaneously for all $\ell \geq 1$, $\varphi \in \mathcal{C}$ and $t \in \mathcal{T}$, because both $\mathcal{C}$ and $\mathcal{T}$ are countable. Moreover, since both sides of the equation above are continuous functions of t, the identity holds simultaneously for all $t \geq 0$, and since ν_ℓ, $\mathcal{D}_{\cdot,\ell}$ and $\mathcal{R}_{\cdot,\ell}$ are finite Radon measures, the identity holds simultaneously for all $\varphi \in \mathbb{C}_c^{1,1}([0,L) \times \mathbb{R}_+)$ using the Dominated Convergence Theorem.

Consequently, it follows from (7.2.21),(7.2.22) and (7.2.23) that for every subsequent limit Y of $\{Y^{(N)}\}$ and every $\ell \geq 1$, ν_ℓ is a solution to the generalized age equation associated with $\lambda\mathbf{Id}$ and $(\nu_\ell(0), D_{\ell+1}, \eta_\ell)$ (see Definition 6.5). By Proposition 6.4, ν is a solution to the hydrodynamic equations (6.1.3)-(6.1.7) associated to $(\lambda\mathbf{Id}, \nu(0))$, which is proved to be unique in Theorem 6.3. This provides a unique characterization of all subsequential limits of $\nu^{(N)}$, and completes the proof. □

7.3. Convergence of Performance Parameters

In this section, we establish the convergence of queue length distribution and virtual waiting time. More importantly, we show that the limit of these quantities can be characterized as functionals of the solution to the hydrodynamic PDEs. As this solution can be approximated using numerical methods, this provides a more computationally effective alternative to Monte Carlo simulations for analyzing properties of large networks. We prove the convergence of queue length distributions in Section 7.3.1, and the convergence of virtual waiting time in Section 7.3.2.

7.3.1. Convergence of Queue Length Distribution. In the section, we establish the convergence of the queue length distribution to a limit that can be characterized by the hydrodynamic PDEs. Specifically, if shows that when the initial conditions are exchangeable, then any finite set of queue at any time t are asymptotically independent and the typical queue length distribution is determined by $(S_\ell; \ell \geq 1)$, and hence we establish a "propagation of chaos" result . Recall that $a^{(N),i}(0)$ to be the initial age of job receiving service in the i^{th} queue at time 0 in the N-server network, if the queue is initially non-empty.

THEOREM 7.24. Suppose Assumptions I-III hold, and the initial conditions are exchangeable in the sense that for every N, the vector

$$\left(X^{(N),\pi(i)}(0), a^{(N),\pi(i)}(0)\mathbb{1}(X^{(N),\pi(i)}(0) > 0); i = 1, ..., N\right),$$

has the same distribution for any permutations π on the queue indices $i \in \{1, ..., N\}$. Then, for every $\ell \geq 1$ and $t \geq 0$,

$$\lim_{N\to\infty} \mathbb{P}\left\{X^{(N),1}(t) \geq \ell\right\} = S_\ell(t), \tag{7.3.1}$$

and for ever fixed $n \geq 0$ and $\ell_1, ..., \ell_n \in \mathbb{N}$,

$$\lim_{N\to\infty} \mathbb{P}\left\{X^{(N),1}(t) \geq \ell_1, ..., X^{(N),n}(t) \geq \ell_n\right\} = \prod_{m=1}^{n} S_{\ell_m}(t). \tag{7.3.2}$$

REMARK 7.25. Let $(Z_\ell; \ell \geq 1)$ is the solution to the hydrodynamic PDEs corresponding to $(\lambda\mathbf{Id}, (Z^0_\ell; \ell \geq 1))$, where

$$Z^0_\ell \doteq \langle \vartheta^r, \nu_\ell(0)\rangle, \qquad r \geq 0.$$

Using the relation $Z_\ell(t, 0) = \langle \mathbf{1}, \nu_\ell(t)\rangle$ established in Proposition (6.10) between the solutions to the measure-valued hydrodynamic equations and the PDEs, equation (7.3.1) can be written as

$$\lim_{N\to\infty} \mathbb{P}\left\{X^{(N),1}(t) \geq \ell\right\} = Z_\ell(t, 0). \tag{7.3.3}$$

and equation (7.3.2) can be written as

$$\lim_{N\to\infty} \mathbb{P}\left\{X^{(N),1}(t) \geq \ell_1, ..., X^{(N),K}(t) \geq \ell_K\right\} = \prod_{k=1}^{K} Z_{\ell_k}(t,0). \tag{7.3.4}$$

PROOF OF THEOREM 7.24. Since the routing algorithm is symmetric with respect to queue indices, the queue lengths and age distributions remain exchangeable for all finite times $t \geq 0$. In particular,

$$\left(X^{(N),i}(t); i = 1, ..., N\right) \stackrel{(d)}{=} \left(X^{(N),\pi(i)}(t); i = 1, ..., N\right). \tag{7.3.5}$$

By definition of $S_\ell^{(N)}(t)$, we have

$$\begin{aligned}
\mathbb{E}\left[\overline{S}_\ell^{(N)}(t)\right] &= \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N} \mathbb{1}\left(X^{(N),i}(t) \geq \ell\right)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N} \mathbb{P}\left\{X^{(N),i}(t) \geq \ell\right\} \\
&= \mathbb{P}\left\{X^{(N),1}(t) \geq \ell\right\}, \qquad (7.3.6)
\end{aligned}$$

where the last equation due to (7.3.5). By Theorem 7.1, and since the solution ν to the hydrodynamic equations is continuous, $\overline{\nu}^{(N)}$ converges to ν uniformly on compact sets, and in particular, for every $t \geq 0$, $\overline{\nu}^{(N)}(t)$ converges to ν in $\mathbb{S}$, in distribution. Hence, by the continuous mapping theorem, $\overline{S}_\ell^{(N)}(t)$ converges to $S_\ell(t)$ in distribution. Since $\sup_N \overline{S}_\ell^{(N)}(t)$ is bounded by 1, the convergence also holds in expectation using the bounded convergence theorem, and (7.3.1) follows from taking the limit as $N \to \infty$ of both sides of (7.3.6).

Now to prove the second claim, note that

(7.3.7)

$$\begin{aligned}
\mathbb{E}\left[\prod_{m=1}^{n} \overline{S}_{\ell_m}^{(N)}(t)\right] =& \frac{1}{N^m}\mathbb{E}\left[\sum_{i_1=1}^{N} ... \sum_{i_n=1}^{N} \mathbb{1}\left(X^{(N),i_1}(t) \geq \ell_1\right) ... \mathbb{1}\left(X^{(N),i_n}(t) \geq \ell_n\right)\right] \\
=& \frac{1}{N^n}\sum_{i_1=1}^{N} ... \sum_{i_n=1}^{N} \mathbb{P}\left\{X^{(N),i_1}(t) \geq \ell_1, X^{(N),i_n}(t) \geq \ell_n\right\} \\
=& \mathbb{P}\left\{X^{(N),1}(t) \geq \ell_1, ..., X^{(N),n}(t) \geq \ell_n\right\},
\end{aligned}$$

where the last equality is again due to (7.3.5). As discussed above, $\overline{\nu}^{(N)}(t)$ converges to $\nu(t)$ in $\mathbb{S}$ in distribution, and hence, using the continuous mapping theorem, $\prod_{m=1}^{n} \overline{S}_{\ell_m}^{(N)}(t)$ converges to $\prod_{m=1}^{n} S_{\ell_m}(t)$ in distribution. Also, since $\sup_N \prod_{m=1}^{n} \overline{S}_{\ell_m}^{(N)}(t)$ is bounded by 1, the convergence also holds in expectation, using the bounded convergence theorem. Taking the limit as $N \to \infty$ of both sides of (7.3.7), (7.3.2) follows. □

7.3.2. Convergence of the Mean Virtual Waiting Time. To show that the solution to the hydrodynamic PDEs can capture the limit of performance measures other than queue length distribution, we establish the convergence of the mean virtual waiting time. The virtual waiting time at any time t is the time that a virtual job that hypothetically arrives at t has to wait in order to receive service. As another application of Theorem 7.1, our next result shows that for large N, the hydrodynamic PDE also provides an approximation to the mean virtual waiting time in an N-server network. Again, the result of this section is depicted for the case $d = 2$, and the extension to general d is straightforward.

THEOREM 7.26. Suppose Suppose Assumptions I-IV hold, and let $(Z_\ell; \ell \geq 1)$ be the unique solution to the hydrodynamic PDE associated to $(\lambda \mathbf{Id}, Z_\ell^0; \ell \geq 1)$. If, in addition, $a^{(N),i}(0) < T_0$ for some $T_0 < \infty$ and every $i = 1, ..., N$, then

$$\lim_{N\to\infty} \mathbb{E}\left[W^{(N)}(t)\right] = \sum_{\ell \geq 2} Z_\ell(t,0)^2 + \sum_{\ell \geq 1} \left[Z_\ell(t,0) + Z_{\ell+1}(t,0)\right] \times \int_0^\infty \left[Z_\ell(t,r) - Z_{\ell+1}(t,r)\right] dr. \tag{7.3.8}$$

REMARK 7.27. The uniform boundedness assumption on the initial ages is reasonable for transient analysis since it will be satisfied by any network that started empty a finite time interval ago. However, the assumption can be relaxed, as illustrated in Figure 2(b) in Section 8.1.2.3. We impose it only to simplify the proofs.

REMARK 7.28. We can show that in fact, as $N \to \infty$, the sequence of virtual waiting time distributions (and not just their means) converge to a limit distribution whose characteristic function can be expressed as a functional of the solution to the hydrodynamic PDE. We do not present the details of the proof, but it is similar to that of Theorem 7.26.

7.3.2.1. *Routing Probabilities.* For every server index i and queue length ℓ, define the routing probability $p(i,\ell;t)$ to be the conditional probability, given the state of the network at time t, that the load-balancing algorithm would route a hypothetical job arriving at time t to server i, when its queue length is ℓ. Defining the vector $\mathbf{X}^{(N)}(t)$ of all queue lengths and ages,

$$\mathbf{X}^{(N)}(t) \doteq \left(X^{(N),i}(t), a^{(N),i}(t); i = 1, ..., N\right). \tag{7.3.9}$$

and denoting by $\kappa(t)$ the (random) index to which the virtual job arriving at time t is routed, the routing probability $p(i,\ell;t)$ is defined by

$$p(i,\ell,t) \doteq \mathbb{1}\left(X^{(N),i}(t) = \ell\right) \mathbb{P}\left\{\kappa(t) = i | \mathbf{X}^{(N)}(t)\right\}. \tag{7.3.10}$$

Now we compute the routing probabilities $p(i,\ell;t)$ for $\ell \geq 1$ and under the $SQ(2)$ algorithm described in Section 5.1.1. If κ_1 and κ_2 are indices of queues chosen independently and uniformly at random, then $\kappa(t)$ is the index associated with the shorter queue (with ties being broken uniformly at random). The job is routed to a server of queue length exactly ℓ if and only if both κ_1 and κ_2 have queue lengths at least ℓ, and at least one of them has queue length ℓ. Given the vector of all queue lengths in $\mathbf{X}^{(N)}(t)$, this happens with probability $(\overline{S}_\ell^{(N)}(t))^2 - (\overline{S}_{\ell+1}^{(N)}(t))^2$. Since all servers with queue length equal to ℓ are equally likely to be chosen and there are $S_\ell^{(N)}(t) - S_{\ell+1}^{(N)}(t)$ of them, for $\ell \geq 1$ we have

$$p(i,\ell;t) = \mathbb{1}\left(X^{(N),i}(t) = \ell\right) \frac{1}{N}\left(\overline{S}_\ell^{(N)}(t) + \overline{S}_{\ell+1}^{(N)}(t)\right) \tag{7.3.11}$$

REMARK 7.29. Note that the form of the routing probability for the $SQ(d)$ algorithm is clearly the same as in the exponential case. To apply our framework to other load balancing algorithms, one would have to replace the expression in (7.3.11) with the routing probabilities associated with that algorithm.

PROOF OF THEOREM 7.26. Suppose that the virtual job arriving at time t is routed to queue i, that is $\kappa(t) = i$. If the server i is idle (i.e., $X^{(N),i}(t) = 0$), the virtual waiting time is zero; otherwise if $X^{(N),i}(t) = \ell$ for some $\ell \geq 1$, the virtual waiting time $W^{(N)}(t)$ is the sum of service times v_j of jobs waiting in queue i plus the residual time $b^{(N),i}(t)$ of the job in service at server i at time t, that is,

$$\mathbb{1}\left(\kappa(t) = i\right) W^{(N)}(t) = \sum_{j=1}^{\ell-1} v_j + b^{(N),i}(t).$$

Summing over all possible queue indices i and queue lengths ℓ, we have

$$W^{(N)}(t) = W_1^{(N)}(t) + W_2^{(N)}(t), \tag{7.3.12}$$

where

$$W_1^{(N)}(t) \doteq \sum_{i=1}^{N} \sum_{\ell \geq 1} \mathbb{1}\left(\kappa(t) = i, X^{(N),i}(t) = \ell\right) \sum_{j=1}^{\ell-1} v_j, \tag{7.3.13}$$

and

$$W_2^{(N)}(t) \doteq \sum_{i=1}^{N} \sum_{\ell \geq 1} \mathbb{1}\left(\kappa(t) = i, X^{(N),i}(t) = \ell\right) b^{(N),i}(t). \tag{7.3.14}$$

To compute the expectation of $W_1^{(N)}(t)$, note that by Assumption II, the service times v_j of jobs that are still waiting in queues satisfy $\mathbb{E}[v_j] = 1$ and are independent of all queue lengths and ages at time t, as well as $\kappa(t)$. Therefore, taking the conditional expectation given $\mathbf{X}^{(N)}(t)$ of both sides of (7.3.13) and using the definition (7.3.10) of $p(i, \ell, t)$, we have

$$\begin{aligned}
\mathbb{E}\left[W_1^{(N)}(t) | \mathbf{X}^{(N)}(t)\right] &= \sum_{i=1}^{N} \sum_{\ell \geq 1} \mathbb{P}\left\{\kappa(t) = i, X^{(N),i}(t) = \ell \middle| \mathbf{X}^{(N)}(t)\right\} \sum_{j=1}^{\ell-1} \mathbb{E}\left[v_j\right] \\
&= \sum_{\ell \geq 1} (\ell - 1) \sum_{i=1}^{N} \mathbb{1}\left(X^{(N),i}(t) = \ell\right) p(i, \ell; t).
\end{aligned}$$

Taking expectations of both sides of the last equation, substituting $p(i, \ell; t)$ from (7.3.11), recalling that the number of servers with queue length equal to ℓ is

$S_\ell^{(N)}(t) - S_{\ell+1}^{(N)}(t)$, we see that

$$\begin{aligned}
\mathbb{E}[W_1^{(N)}(t)] &= \frac{1}{N}\mathbb{E}\left[\sum_{\ell\geq 1}(\ell-1)\left(\overline{S}_\ell^{(N)}(t)+\overline{S}_{\ell+1}^{(N)}(t)\right)\sum_{i=1}^{N}\mathbb{1}\left(X^{(N),i}(t)=\ell\right)\right] \\
&= \frac{1}{N}\sum_{\ell\geq 1}(\ell-1)\mathbb{E}\left[\left(\overline{S}_\ell^{(N)}(t)+\overline{S}_{\ell+1}^{(N)}(t)\right)\left(S_\ell^{(N)}(t)-S_{\ell+1}^{(N)}(t)\right)\right] \\
&= \sum_{\ell\geq 1}(\ell-1)\mathbb{E}\left[\left(\overline{S}_\ell^{(N)}(t)\right)^2-\left(\overline{S}_{\ell+1}^{(N)}(t)\right)^2\right] \\
&= \sum_{\ell\geq 2}\mathbb{E}\left[\left(\overline{S}_\ell^{(N)}(t)\right)^2\right]. \qquad (7.3.15)
\end{aligned}$$

To compute the expectation of $W_2^{(N)}(t)$, note that according to the $SQ(2)$ algorithm, the random queue indices κ_1 and κ_2 are independent of all other random variables, and therefore, given $\mathbf{X}^{(N)}(t)$, $\kappa(t)$ is independent of all residual service times. Therefore, taking conditional expectations of both sides of (7.3.14) and invoking the definition (7.3.10) of $p(i,\ell;t)$ again, we have

(7.3.16)

$$\mathbb{E}\left[W_2^{(N)}(t)|\mathbf{X}^{(N)}(t)\right] = \sum_{\ell\geq 1}\sum_{i=1}^{N}\mathbb{1}\left(X^{(N),i}(t)=\ell\right)p(i,\ell;t)\mathbb{E}\left[b^{(N),i}(t)|\mathbf{X}^{(N)}(t)\right].$$

The residual service time $b^{(N),i}(t)$ of the job that begins being processed by server i at time t satisfies

$$b^{(N),i}(t) = v_{J(i;t)} - a^{(N),i}(t),$$

where $J(i;t)$ is the index of the job begin processes in server i at time t. It follows from the i.i.d assumption on the service times and their independence from the arrival process (Assumption II) that the residual service time $b^{(N),i}(t)$ depends on the state variable $\mathbf{X}^{(N)}(t)$ only through the age $a^{(N),i}(t)$. A complete rigorous justification of this intuitive assertion is rather long and technical, and hence will be presented elsewhere. Hence using equation (3.2) of Section V.3 in [**3**] in the second equality below, for every $r\geq 0$, we have

$$\mathbb{P}\left\{b^{(N),i}(t)>r|\mathbf{X}^{(N)}(t)\right\} = \mathbb{P}\left\{b^{(N),i}(t)>r|a^{(N),i}(t)\right\} = \frac{\overline{G}(a^{(N),i}(t)+r)}{\overline{G}(a^{(N),i}(t))},$$

and therefore,

$$\mathbb{E}\left[b^{(N),i}(t)|\mathbf{X}^{(N)}(t)\right] = \int_0^\infty \frac{\overline{G}(a^{(N),i}(t)+r)}{\overline{G}(a^{(N),i}(t))}\,dr. \tag{7.3.17}$$

Now, substituting equations (7.3.11) and (7.3.17) in (7.3.16), taking expectations of both sides, and using definition of $\nu_\ell^{(N)}$, we can see that

(7.3.18)

$$\begin{aligned}\mathbb{E}[W_2^{(N)}(t)] &= \sum_{\ell\geq 1}\frac{1}{N}\mathbb{E}\left[\left(\overline{S}_\ell^{(N)}(t)+\overline{S}_{\ell+1}^{(N)}(t)\right)\int_0^\infty\sum_{i=1}^N \mathbb{1}\left(X^{(N),i}(t)=\ell\right)\frac{\overline{G}(a^{(N),i}(t)+r)}{\overline{G}(a^{(N),i}(t))}dx\right]\\ &= \sum_{\ell\geq 1}\mathbb{E}\left[\left(\overline{S}_\ell^{(N)}(t)+\overline{S}_{\ell+1}^{(N)}(t)\right)\int_0^\infty\langle\vartheta^r,\overline{\nu}_\ell^{(N)}(t)-\overline{\nu}_{\ell+1}^{(N)}(t)\rangle dr\right].\end{aligned}$$

Finally, taking the limit as $N\to\infty$ in (7.3.18), we obtain

$$\begin{aligned}\lim_{N\to\infty}\mathbb{E}\left[W_1^{(N)}(t)\right] &= \sum_{\ell\geq 2}\lim_{N\to\infty}\mathbb{E}\left[\overline{S}_\ell^{(N)}(t)^2\right]\\ &= \sum_{\ell\geq 2}\left(S_\ell(t)\right)^2.\end{aligned} \tag{7.3.19}$$

The exchange of limit and summation in the first equality is justified by Assumption III.c, the bound $\overline{S}_\ell^{(N)}(t)\leq 1$ and the dominated convergence theorem, while the second equality follows from Theorem 7.1 and the bounded convergence theorem. Moreover, by the uniform boundedness assumption on the ages imposed in Theorem (7.26), $a^{(N),i}(t) < T_0+t$. Therefore, since $\int_0^\infty \overline{G}(r)dr = 1$ because the service time has mean 1,

$$\begin{aligned}\int_0^t\langle\vartheta^r,\overline{\nu}_\ell^{(N)}(t)\rangle dr &\leq \frac{1}{N}\int_0^t\sum_{i=1}^N\frac{\overline{G}(r)}{\overline{G}(T_0+t)}dr\\ &\leq \frac{1}{\overline{G}(T_0+t)}\int_0^t\overline{G}(r)dr\\ &\leq \frac{1}{\overline{G}(T_0+t)}.\end{aligned} \tag{7.3.20}$$

Therefore, taking the limit as $N \to \infty$ of both sides of (7.3.18), we have

$$\lim_{N\to\infty} \mathbb{E}\left[W_2^{(N)}(t)\right] = \sum_{\ell\geq 1} \lim_{N\to\infty} \mathbb{E}\left[\left(\overline{S}_\ell^{(N)}(t) + \overline{S}_{\ell+1}^{(N)}(t)\right) \int_0^\infty \langle \vartheta^r, \overline{\nu}_\ell^{(N)}(t) - \overline{\nu}_{\ell+1}^{(N)}(t)\rangle dr\right]$$

$$= \sum_{\ell\geq 1} \mathbb{E}\left[\left(S_\ell(t) + S_{\ell+1}(t)\right) \int_0^\infty \langle \vartheta^r, \nu_\ell(t) - \nu_{\ell+1}(t)\rangle dr\right]. \tag{7.3.21}$$

The exchange of limit and summation in the first equality is justified by Assumption III.c, $\overline{Z}_\ell^{(N)}(t,r) \leq \overline{Z}_\ell^{(N)}(t,0) \leq 1$ and the dominated convergence theorem, while the second equality holds due to Theorem 7.1, (7.3.20) and the bounded convergence theorem. It follows from (7.3.19) and (7.3.21) that

$$\begin{aligned}\lim_{N\to\infty} \mathbb{E}\left[W^{(N)}(t)\right] &= \sum_{\ell\geq 2} \left(S_\ell(t)\right)^2 \\ &\quad + \sum_{\ell\geq 1} \mathbb{E}\left[\left(S_\ell(t) + S_{\ell+1}(t)\right) \int_0^\infty \langle \vartheta^r, \nu_\ell(t) - \nu_\ell(t)\rangle dr\right].\end{aligned}$$

The result follows from the relation $Z_\ell(t,r) = \langle \vartheta^r, \nu_\ell(t)\rangle$ established in Proposition (6.10) between the solutions to the measure-valued hydrodynamic equations and the PDEs.

□

7.A. A Bound for Renewal Processes

Fix $r_0 \in \mathbb{R}_+$ and let $P(t)$ be a delayed renewal process with inter-arrival times $\{u_n; n \geq 1\}$ with common distribution G and delay u_0 with distribution G_{r_0}, where

$$\mathbb{P}\left\{u_0 \leq x\right\} = G_{r_0}(x) \doteq \frac{G(x + r_0) - G(a)}{1 - G(r_0)}. \tag{7.A.1}$$

Assume G has density g, denote $\overline{G} \doteq 1 - G$ and let h be the hazard rate corresponding the distribution G. Also, let $r(t)$ denote the backward recurrence time of the renewal process P. By convention, $r(t) = r_0 + t$ for $t < u_0$, and in particular, $r(0) = r_0$.

LEMMA 7.30. Let P be the delayed renewal process described above. Then, for every $t \geq 0$,

$$\mathbb{E}\left[\int_0^t h(r(s))ds\right] < \infty, \tag{7.A.2}$$

and

$$\mathbb{E}\left[\left(\int_0^t h(r(s))ds - P(t)\right)^2\right] \leq 4 + \mathbb{E}\left[P(t)\right]. \tag{7.A.3}$$

PROOF. Define the epoch times $\{t_j; j \geq 0\}$ as $t_0 = u_0$ and $t_j = t_{j-1} + u_j$ for $j \geq 1$. Then, we have

$$\int_0^t h(r(s))ds = \int_0^{t_0} h(r(s))ds + \sum_{n=1}^{P(t)} \int_{t_{n-1}}^{t_n} h(r(s))ds - \int_t^{t_{P(t)}} h(r(s))ds$$

$$= \int_{r_0}^{r_0+u_0} h(v)dv + \sum_{n=1}^{P(t)} \int_0^{u_n} h(v)dv - \int_{r(t)}^{u_{P(t)}} h(v)dv. \tag{7.A.4}$$

Defining the random variables $y_n, n \in \mathbb{N}$ as $y_n \doteq \int_0^{u_n} h(v)dv$, the second term on the right-hand side of (7.A.4) can be written as $\sum_{n=1}^{P(t)} y_n$. Since the renewal times $\{u_n; n \geq 1\}$ are i.i.d, the sequence $\{y_n\}_{n\in\mathbb{N}}$ is also i.i.d with

$$\mathbb{E}\left[y_1\right] = \int_0^\infty \left(\int_0^s h(v)dv\right) g(s)ds = \int_0^\infty \frac{g(s)}{\overline{G}(s)} \left(\int_s^\infty g(v)dv\right) ds = 1,$$

and

$$\begin{aligned}
\mathbb{E}\left[(y_1)^2\right] &= \int_0^\infty \left(\int_0^s h(v)dv\right)^2 g(s)ds \\
&= \int_0^\infty \left(\log(\overline{G}(s))\right)^2 g(s)ds \\
&= \int_0^1 \left(\log(s)\right)^2 ds \\
&= 2.
\end{aligned}$$

Thus, mean and variance of y_n are both equal to 1. Now, define the discrete-time filtration $\{\mathcal{G}_n; n \geq 0\}$ by $\mathcal{G}_n = \sigma(u_j; j = 0, ..., n)$. Note that u_n, and hence, y_n are $\mathcal{G}_n$-measurable. Also, since the inter-arrival times are independent, $y_{n+1}, y_{n+2}, ...$ are independent of $\mathcal{G}_n$. Finally, the random variable $P(t)$ satisfies $\mathbb{E}\left[P(t)\right] \leq U(t) < \infty$

where U is the renewal measure corresponding to the distribution G. Moreover, since t_{n-1} and t_n are $\mathcal{G}_n$-measurable for $n \geq 1$,

$$\{P(t) = n\} = \{t_{n-1} \leq t < t_n\} \in \mathcal{G}_n,$$

and hence, $P(t)$ is an integrable $\{\mathcal{G}_n\}$-stopping time. Therefore, by Wald's lemma (see e.g. [**3**, Proposition A 10.2]),

$$\mathbb{E}\left[\sum_{n=1}^{P(t)} \int_0^{u_n} h(v)dv\right] = \mathbb{E}\left[\sum_{n=1}^{P(t)} y_n\right] = \mathbb{E}\left[P(t)\right]\mathbb{E}\left[y_1\right] = \mathbb{E}\left[P(t)\right] < \infty, \tag{7.A.5}$$

and

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{n=1}^{P(t)} \int_0^{u_n} h(v)dv - P(t)\right)^2\right] &= \mathbb{E}\left[\left(\sum_{n=1}^{P(t)} y_n - P(t)\right)^2\right]\\ &= \mathbb{E}\left[P(t)\right]\mathrm{Var}(y_1)\\ &= \mathbb{E}\left[P(t)\right]. \qquad (7.A.6)\end{aligned}$$

Now for the first term on the right-hand side of (7.A.4), recalling the distribution (7.A.1) of u_0,

$$\begin{aligned}\mathbb{E}\left[\int_{r_0}^{r_0+u_0} h(v)dv\right] &= \frac{1}{\overline{G}(r_0)}\int_0^\infty \left(\int_{r_0}^{r_0+u} h(v)dv\right) g(r_0+u)du\\ &= \frac{1}{\overline{G}(r_0)}\int_0^\infty \left(\log(\overline{G}(r_0)) - \log(\overline{G}(r_0+u))\right) g(r_0+u)du\\ &= \frac{\log(\overline{G}(r_0))}{\overline{G}(r_0)}\int_{r_0}^\infty g(u)du - \frac{1}{\overline{G}(r_0)}\int_0^{\overline{G}(r_0)} log(v)dv\\ &= 1, \qquad (7.A.7)\end{aligned}$$

and

(7.A.8)

$$\begin{aligned}
&\mathbb{E}\left[\left(\int_{r_0}^{r_0+u_0} h(v)dv\right)^2\right] \\
&\quad = \frac{1}{\overline{G}(r_0)}\int_0^\infty \left(\int_{r_0}^{r_0+u} h(v)dv\right)^2 g(r_0+u)du \\
&\quad = \frac{1}{\overline{G}(r_0)}\int_0^\infty \left(\log(\overline{G}(r_0)) - \log(\overline{G}(r_0+u))\right)^2 g(r_0+u)du \\
&\quad = \log(\overline{G}(r_0))^2 - 2\frac{\log(\overline{G}(r_0))}{\overline{G}(r_0)}\int_0^{\overline{G}(r_0)} log(v)dv + \frac{1}{\overline{G}(r_0)}\int_0^{\overline{G}(r_0)}(\log(v))^2 dv \\
&\quad = \log(\overline{G}(r_0))^2 - 2\log(\overline{G}(r_0))(\log(\overline{G}(r_0)) - 1) + (\log(\overline{G}(r_0))^2 \\
&\qquad - 2\log(\overline{G}(r_0)) + 2) \\
&\quad = 2.
\end{aligned}$$

For the last term, since $r(t) \geq 0$,

$$\mathbb{E}\left[\int_{r(t)}^{u_{P(t)}} h(v)dv\right] \leq \mathbb{E}\left[y_{P(t)}\right] = 1 \tag{7.A.9}$$

and

$$\mathbb{E}\left[\left(\int_{r(t)}^{u_{P(t)}} h(v)dv\right)^2\right] \leq \mathbb{E}\left[(y_{P(t)})^2\right] = 2. \tag{7.A.10}$$

Finally, taking expectation from both sides of (7.A.4) and substituting the bounds (7.A.5), (7.A.7), and (7.A.9), (7.A.2) follows. Moreover, by another use of (7.A.4) and triangle inequality, we have

$$\begin{aligned}
\mathbb{E}\left[\left(\int_0^t h(r(s))ds - P(t)\right)^2\right] \leq{}& \mathbb{E}\left[\left(\int_{r_0}^{r_0+u_0} h(v)dv\right)^2\right] \\
&+ \mathbb{E}\left[\left(\sum_{n=1}^{P(t)}\int_0^{u_n} h(v)dv - P(t)\right)^2\right] \\
&+ \mathbb{E}\left[\left(\int_{r(t)}^{u_{P(t)}} h(v)dv\right)^2\right],
\end{aligned}$$

and (7.A.3) follows using the bounds (7.A.6), (7.A.8) and (7.A.10). $\square$

CHAPTER 8

Numerical Results and Engineering Insights

In this chapter we discuss the application of our result in understanding the behavior of stochastic networks in practice. First in Section 8.1, we devise and use a stable numerical scheme to approximate the solution to the hydrodynamic PDEs, which provides a computationally efficient alternative to Monte Carlo simulations that could be useful for studying the performance and design of large networks. We show that the obtained numerical approximations provides a good match with simulations. Then in Section 8.2, we illustrate how this approximation can be used to shed insight into transient phenomena of practical relevance.

Throughout this chapter, we stick to the case $d = 2$.

8.1. Numerical Approximation and Validation

In this section we validate the results in Theorem 7.24 and Theorem 7.26 using Monte Carlo (MC) simulations. First in Section 8.1.1, we present a stable scheme for numerically approximating the solution to the hydrodynamic PDE. Then in Section 8.1.2, we compare the results obtained from Monte Carlo simulation and numerical solutions to PDE for a variety of different service distributions and in each case, we observe a close match.

8.1.1. A Numerical Approximation Scheme. Note that the hydrodynamic PDE is comprised of a countable number of integro-differential equations, corresponding to each Z_ℓ. Therefore, in order to numerically solve the hydrodynamic PDE (6.2.1)-(6.2.4), we truncate the state space and only solve for $Z_\ell(t,r)$, $\ell = 1, \dots, L_0$, $0 \leq t \leq T_0$, $0 \leq r \leq R_0$, for suitable $L_0 \in \mathbb{N}$ and $R_0, T_0 < \infty$. Next, we discretize r and t on uniform meshes: $\hat{Z}_\ell(t_m, r_n) \doteq Z_\ell(m\delta, n\delta)$ with $0 \leq m \leq \lfloor T_0/\delta \rfloor$ and $0 \leq n \leq \lfloor R_0/\delta \rfloor$. Then we solve the equations (6.2.1)-(6.2.4) numerically by the finite difference method and the explicit forward Euler scheme[**47**]. That is, for fixed $t_m > 0$ and $r_n > 0$, we apply the following approximations

$$
\begin{aligned}
\hat{Z}_1(t_m+\delta, r_n) = \hat{Z}_1(t_m, r_n+\delta) &- \int_{t_m}^{t_m+\delta} \overline{G}(t_m+\delta+r_n-u)\partial_r \hat{Z}_2(u,0)du \\
&+ \int_t^{t+\delta} \lambda(u)\overline{G}(t_m+\delta+r_n-u)\left(1-\hat{Z}_1(u,0)^2\right)du \\
&\approx \hat{Z}_1(t_m, r_n+\delta) - \overline{G}(r_n)\left(\hat{Z}_2(t_m,\delta)-\hat{Z}_2(t_m,0)\right) \\
&+ \overline{G}(r_n)\left(1-\hat{Z}_1(t_m,0)^2\right)\int_t^{t+\delta}\lambda(u)du.
\end{aligned}
$$

and for $\ell \geq 2$,

$$
\begin{aligned}
\hat{Z}_\ell(t_m+\delta, r_n) = \hat{Z}_\ell(t_m, r_n+\delta) &- \int_{t_m}^{t_m+\delta} \overline{G}(t_m+\delta+r_n-u)\partial_r \hat{Z}_{\ell+1}(u,0)du \\
&+ \int_{t_m}^{t_m+\delta} \lambda(u)\left(\hat{Z}_{\ell-1}(u,0)+\hat{Z}_\ell(u,0)\right) \\
&\quad \times \left(\hat{Z}_{\ell-1}(u, t_m+\delta+r_n-u) - \hat{Z}_\ell(u, t_m+\delta+r_n-u)\right)du \\
&\approx \hat{Z}_\ell(t_m, r_n+\delta) - \overline{G}(r_n)\left(\hat{Z}_{\ell+1}(t_m,\delta)-\hat{Z}_{\ell+1}(t_m,0)\right) \\
&+ \left(\hat{Z}_{\ell-1}(t_m,0)+\hat{Z}_\ell(t_m,0)\right)\left(\hat{Z}_{\ell-1}(t_m,r_n)-\hat{Z}_\ell(t_m,r_n)\right) \\
&\quad \times \int_{t_m}^{t_m+\delta}\lambda(u)du.
\end{aligned}
$$

The numerical scheme above has been stabilized by applying the upwind scheme [**23**] to the derivatives with respect to r, otherwise, the numerics will blow up in a short time interval. Using these approximations, we can update $\hat{Z}_\ell\left(t_m+\delta,\cdot\right)$ from $\hat{Z}_\ell(t_m,\cdot)$.

8.1.2. Validation of Results. We now compare the empirical queue length distribution and mean virtual waiting time obtained from the Monte Carlo simulation of an N-server network, with the corresponding limit quantities (7.3.3) and (7.3.8), as predicted by the numerical approximation of the PDE.

We consider a sequence of networks indexed by the number of servers N, with a Poisson arrival process with rate $\lambda = 0.5$, and the following initial conditions. Each server has initially one job with initial age equal to zero, that is $X^{(N),i}(0) = 1$

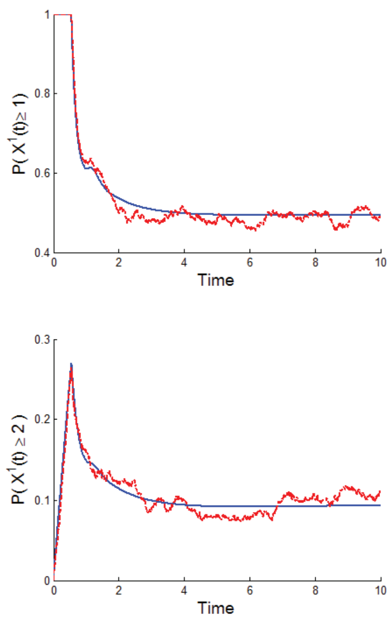
and $a^{(N),i}(0) = 0$ for all $i = 1, ..., N$. Note that this sequence of initial conditions satisfies Assumption III, and converges to the initial condition $(Z^0_\ell(\cdot); \ell \geq 1)$ for the hydrodynamic PDE, where for $r \geq 0$,

$$Z^0_1(r) = \overline{G}(r), \qquad Z^0_\ell(r) = 0, \quad \ell \geq 2.$$

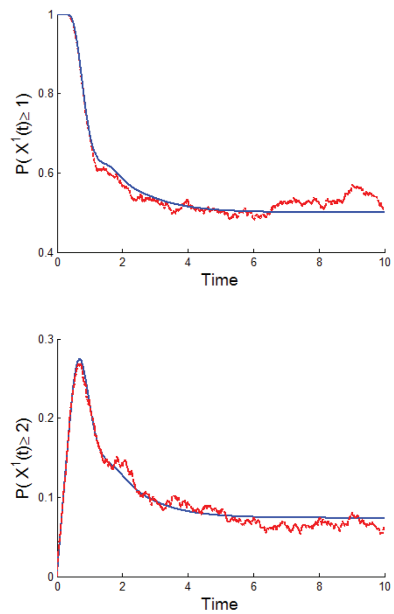
8.1.2.1. *Queue Length Distribution.* First we compute the probability that a typical (fixed) queue has length at least ℓ at time t, for $t \in [0, 10]$ in a network of $N = 1000$ servers, using Monte Carlo simulation with 1000 realizations. Then we compare this probability with the quantity obtained by numerically solving the PDE using the method described in Section 8.1.1, with $L_0 = 6$, $R_0 = 20$ and $\delta = 0.001$. We make this comparison for a variety of (unit mean) service time distribution, including Pareto (with shape parameter $\beta = 2.25$), Log-Normal (with shape parameter $\sigma = 0.33$), Gamma (with shape parameter $k = 2$) and Hyper-Exponential (with parameters $\lambda_1 = 0.5$ and $\lambda_2 = 2$), which is a special case of a Phase-type distribution.

The results are illustrated in Figure 1 for $\ell = 1$ and $\ell = 2$, and a close match is observed in all cases. In our setting, the run time for the Monte Carlo method is approximately $2 - 3$ hours, which is orders of magnitude longer than the time taken to numerically approximate the PDE, which is around $7 - 9$ seconds.

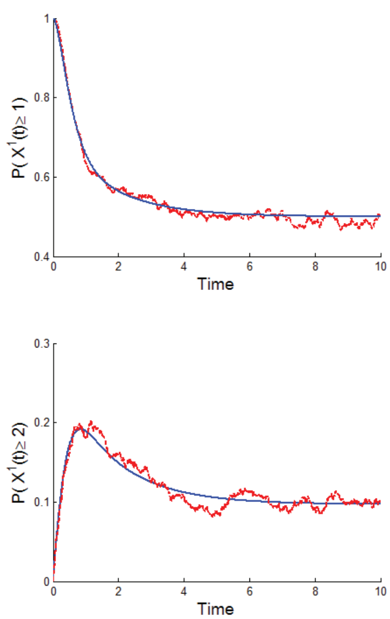
8.1.2.2. *Waiting Time.* Next, we compare the mean virtual waiting time in the same setting described above, but with the arrival rate now set to $\lambda = 0.7$. We measure the mean virtual waiting from a Monte Carlo simulation as follows: at any time t, we determine which server a virtual arriving job would have been routed to by choosing two servers uniformly at random, and picking the one with shorter queue length, and then observe the waiting time in that server. The average is taken over 2000 realizations. We compare this to the following approximation of



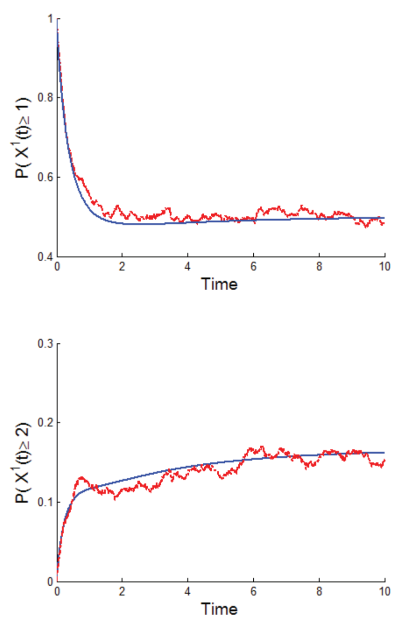
(a) Pareto



(b) Log-Normal



(c) Gamma



(d) Hyper-Exp.

FIGURE 1. Comparison of the estimate for $P(X^{(N),1}(t) \geq \ell)$ obtained from MC simulation (in red) versus the numerical approximation of $Z_\ell(t,0)$ (in blue) during $t \in [0,10]$ for $\ell = 1$ (top row) and $\ell = 2$ (bottom row).

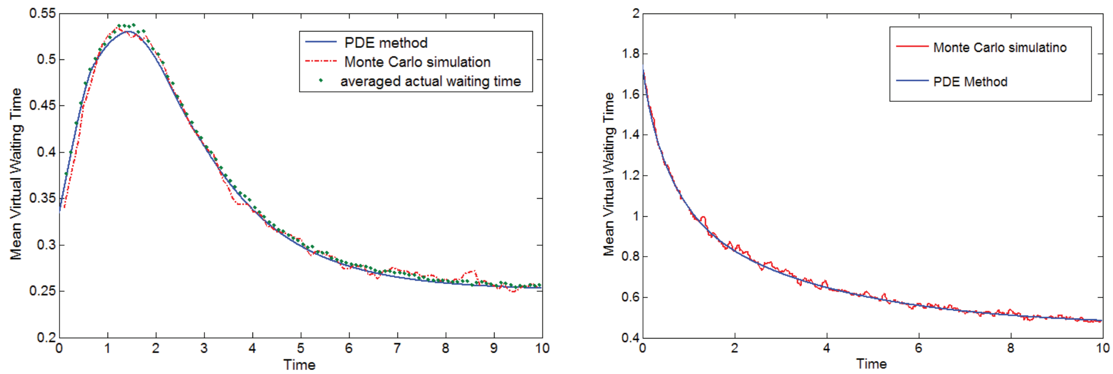


FIGURE 2. **a.** Mean virtual sojourn times from MC (red) and PDE (blue) as well as the averaged actual sojourn times from MC (green) for the Pareto distribution with $\beta = 3$. **b.** Mean virtual waiting times from MC (red) and PDE (blue) for Gamma service distribution and unbounded initial ages.

the limit provided by Theorem 7.26:

$$\mathbb{E}\left[W^{(N)}(t)\right] \approx \sum_{\ell=2}^{L_0} \hat{Z}_\ell(t,0)^2 + \sum_{\ell=1}^{L_0-1} \left[\hat{Z}_\ell(t,0) + \hat{Z}_{\ell+1}(t,0)\right] \times \sum_{j=0}^{\lfloor R_0/\delta \rfloor} \left[\hat{Z}_\ell(t,r_j) - \hat{Z}_{\ell+1}(t,r_j)\right]\delta. \tag{8.1.1}$$

where $\{\hat{Z}_\ell; \ell \geq 1\}$ is the numerical solution of the PDE described in Section 8.1.1.

The result of the comparison for the Pareto distribution (with mean set to 1 and shape parameter $\beta = 3$) and is plotted in Figure 2(a). We observe good agreement between the two curves. Furthermore, recall that the actual waiting time of a job is the difference between the arrival and service entry times of that job. We also plot the average of the average of the actual waiting time. At each time t, the latter quantity is defined to be the sum of the waiting times of all jobs arrived in that time slot in the mesh, divided by the number of jobs that arrived in that time slot. We observe that the mean virtual waiting time is a good approximation to the average actual waiting time as well.

8.1.2.3. *More General Initial Conditions.* Finally, to illustrate that the assumption $a^{(N),i}(0) \leq T_0$ on initial ages is not necessary, we validate our result for another sequence of networks with the following initial condition: there are initially two jobs in each queue, and the initial ages are all independent and distributed according to $\overline{G}(x)dx$ (which is the stationary age distribution in a renewal process with inter-arrival distribution G). The corresponding initial condition for the PDE are given by

$$Z_\ell^0(r) = \int_r^\infty \overline{G}(x)dx, \quad \ell = 1, 2, \qquad\qquad Z_\ell^0(r) = 0, \quad \ell \geq 3.$$

We validate the approximation given in (8.1.1) using Monte Carlo simulations for a network with $N = 1000$ servers and a Gamma service time distribution (with shape parameter $k = 2$). The result is depicted in Figure 2(b).

8.2. Engineering Insights

In practice, transient Quality of Service (QoS) parameters are of particular interest in many applications. However, to the best of our knowledge, prior to this work, load-balancing networks with general service distributions have only been studied in steady state and under constant Poisson arrivals [**12**]. We illustrate how our PDE approximation can be used to shed insight into transient phenomena of practical relevance. In Section 8.2.1, we study the time it takes for a congested network to get rid of a backlog of jobs, and in Section 8.2.2 we analyze the performance of a load balancing network with time-varying, periodic, Poisson arrivals, consisting of intermittent "high" and "low" periods.

8.2.1. Initial Backlog. Given a network that is congested with a backlog of jobs, the system administrator would like to know how long it would take for the system to get rid of the backlog and for the QoS to return closer to the normal operating point. In this section, we consider a network with a large initial backlog, and hence a large initial virtual waiting time, and study the time it takes for the system to unload to the extent that the mean virtual waiting time reaches half of its initial value, which we refer to as "relaxation time". We investigate the effect of service distribution statistics on the relaxation time. In each case, the goal is to illustrate how the PDE approximation may be used to help uncover interesting (and possibly unexpected) network phenomena.

In the presence of general service distributions, in order to capture congestion in the initial distribution of the network, one has to specify not only the distribution of the number of jobs in the different queues, but also the distribution of ages of jobs being served at these queues. There are a number of different configurations that could reflect a congested initial state. In order to generate initial conditions that may naturally occur in practice, we consider an initial state that corresponds to the distribution resulting from a network that has been experiencing a higher than normal arrival rate for a period of time. Specifically, we first consider a network that starts empty and runs with nominal arrival rate $\lambda = 0.6$ for $T_0 = 10$ units of time, and then experienced an arrival rate that is about 8 times the nominal arrival rate, namely $\lambda = 5$, for a period of $T_b = 2$ units prior to zero. The distribution of this network at time $T_0 + T_b$ then represents a congested state for the network and is used as the initial condition for our PDE approximations. We then assume that at time 0, the mean arrival rate reverts back to its nominal value $\lambda = 0.6$ and we study the relaxation time, namely the time it takes for the mean virtual waiting time to reach half its initial value.

In Figure 3(a), we plot the evolution in time of the mean virtual waiting time for (nominal) arrival rate $\lambda = 0.6$, both when the service distribution is Pareto with unit mean and parameter $\beta = 1.25$ (heavy-tailed) and Pareto with parameter $\beta = 2.50$ (light-tailed). The curves are obtained using the numerical scheme to solve the PDE with cutoff values $L_0 = 12$ and $R_0 = 20$ and step size $\delta = 0.001$. An interesting observation is that the heavy-tailed case clearly has a smaller relaxation time than the light-tailed case. This phenomenon may seem particularly surprising in light of the steady-state result of Bramson et. al in **[12]**, which shows that for Pareto

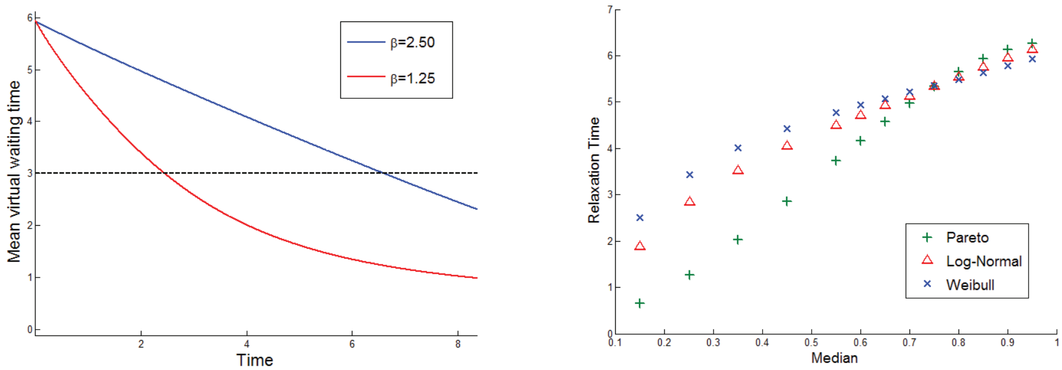


FIGURE 3. **a.** Mean virtual waiting time for network with initial backlog and Pareto service distribution with $\beta = 1.25$ (red) and $\beta = 2.50$ (blue), using the PDE method. **b.** Relaxation time vs. Median for Pareto, Log-Normal and Weibull service time distributions.

service time distributions with unit mean, the tail of the limit steady state queue-length distribution has a double exponential decay when $\beta > 2$ (light-tailed), in contrast to the case when $\beta < 2$ (heavy-tailed), when it has only a power law decay. Although the tails of the limit steady-state distribution do not represent the limit of the tails of the steady-state distribution in the N-server system (since the $N \to \infty$ and tail decay limits are typically not interchangeable), the results in **[12]** suggest that from the point of view of equilibrium queue length, the light-tailed Pareto distribution shows better performance.

The transient phenomenon observed above also persists when the initial condition is instead chosen to be a large number of jobs uniformly distributed amongst different queues, all starting with zero ages, which may roughly correspond to congestion caused due to a sudden large spurt of job arrivals into the network.

A possible heuristic explanation for the contrasting behavior observed in the transient case is that when a Pareto service distribution is fixed to have mean equal to one, the median of the distribution decreases with a decrease in β. As a result, when the heavier the tail of the service distribution, the greater the fraction of initially backlogged (and newly arriving) jobs with smaller service requirements, and the smaller the fraction of jobs with very long service times. This helps servers in a large number of queues to get rid of their backlog and reduce their queue lengths faster, whereas the jobs with long service times lead to large queue lengths in only a small fraction of servers. The latter jobs do not increase the mean virtual waiting time significantly because a new potential arrival will avoid the few long queues with high probability under the $SQ(2)$ load-balancing policy. This argument suggests that the same trend should hold for different classes of service time distributions. In Figure 3(b), we plot the relaxation times obtained via the PDE approximation for the Pareto, Weibull and Log-Normal service time distributions for different median values, and observe that the relaxation time decreases as the median decreases (and variance increases) for all these distribution families.

8.2.2. Periodic Arrivals and Effective Arrival Rate. In contrast to prior work, our method also allows us to study the $SQ(d)$ network under time-varying arrivals. In many real-world applications, the arrival process is periodic (over period of a day, for example,) comprised of peak and off-peak periods. In this scenario, we examine the effect of the time- inhomogeneity of the arrival on the performance of the network.

As an illustration, we consider a family of periodic arrival rate functions $\lambda(\cdot) \in L^1_{\text{loc}}(0,\infty)$, with period T, parameterized by the average arrival rate $\overline{\lambda}$ over the period and a constant $\Delta > 0$, which can be viewed as a burstiness parameter. The arrival rate takes the value $\overline{\lambda} + \Delta$ in the first-half of the period and the value $\overline{\lambda} - \Delta$ in the second half of the period, corresponding to peak and off-peak periods. The larger the Δ, the more dramatic the difference between the peak and off-peak arrival rates; the particular case $\Delta = 0$ corresponds to the case of constant arrival rate. Using our PDE approximation, for fixed $\overline{\lambda}$, we study the effect of the burstiness parameter Δ on the mean virtual waiting time of the network.

As one would expect, the burstiness has a negative impact on the network performance as it results in an increase in the mean virtual waiting time. However, the PDE approximation also quantifies this effect, and allows comparisons across different service time distributions. To this end, for a fixed average arrival rate and different values of the burstiness parameter Δ, we compute the mean virtual waiting time averaged over one period, denoted by $\overline{W}(\overline{\lambda}, \Delta)$, and find the *constant rate* Poisson arrival that results in the same averaged virtual waiting time. We denote this constant rate by $\lambda_{\text{eff}} = \lambda_{\text{eff}}(\overline{\lambda}, \Delta)$, and refer to it as the "effective arrival rate" corresponding to $\overline{\lambda}$ and Δ, which hence has the property

$$\overline{W}(\overline{\lambda}, \Delta) = \overline{W}(\lambda_{\text{eff}}, 0).$$

We compare the effect of traffic burstiness on the performance of the network under a heavy tail and a light tail Pareto service distribution. In Figure 4, we plot λ_{eff} as a function of Δ for $\overline{\lambda} = 0.7$ and Pareto service time distributions with shape parameters $\beta = 1.5$ (heavy tail, infinite variance) and $\beta = 3.0$ (light tail, finite variance.) As illustrated in the figure, the burstiness of the arrival rate has a greater effect on the network when the service time has finite variance ($\beta = 3.0$) rather than infinite variance ($\beta = 1.5$). In other words, as far as the average mean virtual waiting time is concerned, the network with the heavy-tailed service time shows less increase in λ_{eff} than the light-tailed service time. This is in line with our observation of backlogged networks that heavy tails seem to have a less deleterious effect on the transient behavior of the network than in equilibrium.

It is worth mentioning that quantitative insights into the network such as those illustrated in Figure 4, may require the solution of inverse problems, such as the computation of λ_{eff} for various values of Δ, which would be incredibly time-consuming, if not infeasible using Monte Carlo simulation.

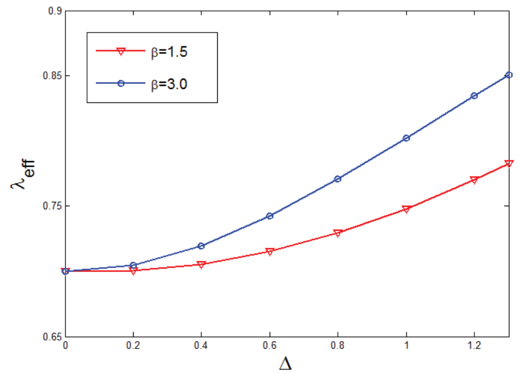


FIGURE 4. Effective arrival rate $\boldsymbol{\lambda}_{\text{eff}}$ vs. burstiness Δ for Pareto service distributions with $\beta = 1.5$ and $\beta = 3$.

Bibliography

[1] R. A. Adams. *Sobolev Spaces*. Pure and applied mathematics. Academic Press, 1975.

[2] Z. Aksin, M. Armony, and V. Mehrotra. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6):665–688, 2007.

[3] S. Asmussen. *Applied Probability and Queues*. Springer-Verlag, 2nd edition edition, 2003.

[4] R. Atar, H. Kaspi, and N. Shimkin. Fluid limits for many-server systems with reneging under a priority policy. *Math. Oper. Res.*, 39(3):672–696, 2014.

[5] Yossi Azar, Andrei Z. Broder, Anna R. Karlin, and Eli Upfal. Balanced allocations. *SIAM J. Comput.*, 29(1):180–200, September 1999.

[6] Y. Bakhtin and J. C. Mattingly. Stationary solutions of stochastic differential equations with memory and stochastic partial differential equations. *Commun. Contemp. Math.*, 7(5):553–582, 2005.

[7] D. Berger, B. Daily, P. Dunn, D. Gamarnik, R. Levi, W. C. Levine, W.S. Sandberg, and T Schoenmeyr. A model for understanding the impacts of demand and capacity on waiting time to enter a congested recovery room. *Anesthesiology*, 110(6):1293 – 1304, 2009.

[8] P. Billingsley. *Convergence of Probability Measures*. John Wiley, New York, 1968.

[9] M. Bramson. Stability of join the shortest queue networks. *Ann. Appl. Probab.*, 21(4):1568–1625, 08 2011.

[10] Maury Bramson, Yi Lu, and Balaji Prabhakar. Randomized load balancing with general service time distributions. *SIGMETRICS Perform. Eval. Rev.*, 38(1):275–286, June 2010.

[11] Maury Bramson, Yi Lu, and Balaji Prabhakar. Asymptotic independence of queues under randomized load balancing. *Queueing Systems*, 71(3):247–292, 2012.

[12] Maury Bramson, Yi Lu, and Balaji Prabhakar. Decay of tails at equilibrium for FIFO join the shortest queue networks. *The Annals of Applied Probability*, 23(5):1841–1878, 10 2013.

[13] L. Breiman. *Probability*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1968.

[14] P. Bremaud. *Point Processes and Queues: Martingale Dynamics: Martingale Dynamics*. Advances in Physical Geochemistry. Springer, 1981.

[15] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer New York, 2010.

[16] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center. *Journal of the American Statistical Association*, 100(469):36–50, 2005.

[17] A. Budhiraja and C. Lee. Stationary distribution convergence for generalized Jackson networks in heavy traffic. *Math. Oper. Res.*, 34(1):45–56, 2009.

[18] T.A. Burton. *Volterra Integral And Differential Equations*. Mathematics in Science and Engineering Series. Academic Press, 2005.

[19] T. Cayirli and E. Veral. Outpatient scheduling in health care: A review of literature. *Production and Operations Management*, 12(4):519–549, 2003.

[20] H. Chen and D. Yao. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Applications of mathematics : stochastic modelling and applied probability. Springer, 2001.

[21] S. Chen, Y. Sun, U. C. Kozat, L. Huang, P. Sinha, G. Liang, X. Liu, and N. B. Shroff. When queueing meets coding: Optimal-latency data retrieving scheme in storage clouds. In *INFOCOM, 2014 Proceedings IEEE*, pages 1042–1050, April 2014.

[22] J. W. Cohen and O. J. Boxma. A survey of the evolution of queueing theory. *Statistica Neerlandica*, 39(2):143–158, 1985.

[23] Richard Courant, Eugene Isaacson, and Mina Rees. On the solution of nonlinear hyperbolic differential equations by finite differences. *Communications on Pure and Applied Mathematics*, 5(3):243–255, 1952.

[24] G. Da Prato and J. Zabczyk. *Ergodicity for Infinite Dimensional Systems*. Cambridge Monographs on Particle Physics, Nuclear Physics and Cosmology. Cambridge University Press, 1996.

[25] J.G. Dai, A.B. Dieker, and X. Gao. Validity of heavy-traffic steady-state approximations in many-server queues with abandonment. *Queueing Systems*, 78(1):1–29, 2014.

[26] D. J. Daley and N. R. Mohan. Asymptotic behaviour of the variance of renewal processes and random walks. *The Annals of Probability*, 6(3):pp. 516–521, 1978.

[27] L. Decreusefond and P. Moyal. A functional central limit theorem for the $M/GI/\infty$ queue. *Annals of Applied Probability*, 18(6):2156–2178, 2008.

[28] E. Dibenedetto. *Real Analysis*. Birkhäuser Advanced Texts Basler Lehrbücher Series. Birkhauser Verlag GmbH, 2002.

[29] B. Doytchinov, J. Lehoczky, and S. Shreve. Real-time queues in heavy traffic with earliest-deadline-first queue discipline. *Ann. Appl. Probab.*, 11(2):332–378, 2001.

[30] W. E, J. C. Mattingly, and Y. Sinai. Gibbsian dynamics and ergodicity for the stochastically forced Navier-Stokes equation. *Comm. Math. Phys.*, 224:83–106, 2001.

[31] A. K. Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Elektrotkeknikeren*, 13, 1917.

[32] A. K. Erlang. *On the rational determination of the number of circuits*. The Copenhagen Telephone Company, Copenhagen, 1948.

[33] S.N. Ethier and T.G. Kurtz. *Markov processes: characterization and convergence*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, 1986.

[34] L.C. Evans. *Partial Differential Equations*. Graduate Studies in Mathematics. American Math. Soc., 1998.

[35] V.F. Farias, C.C. Moallemi, and B. Prabhakar. Load balancing with migration penalties. In *Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on*, pages 558–562, Sept 2005.

[36] S. Foss and N. Chernova. On the stability of a partially accessible multi-station queue with state-dependent routing. *Queueing Systems*, 29(1):55–73, 1998.

[37] D. Gamarnik and D. Goldberg. Steady-state $GI/GI/n$ queue in the Halfin-Whitt regime. *Ann. Appl. Probab.*, 23(6):2382–2419, 2013.

[38] D. Gamarnik and P. Momcilovic. Steady-state analysis of a multi-server queueing system in the QED regime. *Adv. Appl. Probab.*, 40(2):548–577, 2008.

[39] D. Gamarnik and A. Zeevi. Validity of heavy traffic steady-state approximations in generalized jackson networks. *Ann. Appl. Prob.*, 16(1):56–90, 2006.

[40] David Gamarnik and Alexander L. Stolyar. Multiclass multiserver queueing system in the Halfin-Whitt heavy traffic regime: asymptotics of the stationary distribution. *Queueing Systems*, 71(1-2):25–51, 2012.

[41] N. Gans, G. Koole, and A. Mandelbaum. Commissioned paper: Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, April 2003.

[42] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227, 2002.

[43] D. Goldberg. On the steady-state probability of delay and large negative deviations for the GI/GI/n queue in the Halfin-Whitt regime. 2013.

[44] Carl Graham. Chaoticity on path space for a queueing network with selection of the shortest queue among several. *Journal of Applied Probability*, 37(1):198–211, 03 2000.

[45] H. C. Gromoll. Diffusion approximation for a processor sharing queue in heavy traffic. *Ann. Appl. Probab.*, 14(2):555–611, 05 2004.

[46] H. C. Gromoll, Ł. Kruk, and A. L. Puha. Diffusion limits for shortest remaining processing time queues. *Stoch. Syst.*, 1(1):1–16, 2011.

[47] C. Grossmann, H.G. Roos, and M. Stynes. *Numerical Treatment of Partial Differential Equations*. Universitext. Springer Berlin Heidelberg, 2007.

[48] I. Gurvich. Validity of heavy-traffic steady-state approximations in multiclass queueing networks: The case of queue-ratio disciplines. *Math. Oper. Res.*, 39(1):121–162, 2014.

[49] M. Hairer. Exponential mixing properties of stochastic PDEs through asymptotic coupling. *Probab. Theory Related Fields*, 124:345380, 2002.

[50] M. Hairer, J. C. Mattingly, and M. Scheutzow. Asymptotic coupling and a general form of Harris theorem with applications to stochastic delay equations. *Probability Theory and Related Fields*, 149(1-2):223–259, 2011.

[51] S. Halfin and W. Whitt. Heavy-traffic limit theorems for queues with many exponential servers. *Oper. Res.*, 29:567–588, 1981.

[52] M. J. Harrison. *Brownian Models of Performance and Control*. Cambridge University Press, 2013.

[53] R. Jackson. Some applications of queueing theory in operational research ($1953-1989$). *IMA Journal of Management Mathematics*, 2(2):127–140, 1989.

[54] D. L. Jagerman. Some properties of the erlang loss function. *Bell System Technical Journal*, 53(3):525–551, 1974.

[55] A. Jakubowski. On the Skorokhod topology. *Ann. Inst. H. Poincaré Probab. Statist.*, 22(3):263–285, 1986.

[56] P. Jelenkovic, A. Mandelbaum, and P. Momcilovic. Heavy traffic limits of queues with many deterministic servers. *Queueing Systems*, 47:53–69, 2004.

[57] W. Kang and K. Ramanan. Fluid limits of many-server queues with reneging. *Ann. Appl. Probab.*, 20:2204–2260, 2010.

[58] W. Kang and K. Ramanan. Asymptotic approximations for the stationary distributions of many-server queues. *Ann. Appl. Probab.*, 22(2):477–521, 2012.

[59] I. Karatzas and S. Shreve. *Brownian Motion and Stochastic Calculus*. Graduate Texts in Mathematics Series. Springer-Verlag GmbH, 1991.

[60] K. Kardassakis. Load balancing in stochastic networks: Algorithms, analysis, and game theory. Undergraduate Honors Thesis, Brown University, 2014.

[61] H. Kaspi and K. Ramanan. Law of large numbers limits for many-server queues. *Ann. Appl. Probab.*, 21(1):33–114, 2011.

[62] H. Kaspi and K. Ramanan. SPDE limits of many server queues. *Annals of Applied Probabability*, Vol. 23(No. 1):145–229, 2013.

[63] T. Katsuda. State-space collapse in stationarity and its application to a multiclass single-server queue in heavy traffic. *Queueing Systems*, 65(3):237–273, 2010.

[64] N. Kazamaki and T. Sekiguchi. Uniform integrability of continuous exponential martingales. *Tohoku Math. J.*, Vol. 35(No. 2):289–301, 1983.

[65] J. Kiefer and J. Wolfowitz. On the theory of queues with many servers. *Transactions of the American Mathematical Society*, 78(1):1–18, 1955.

[66] F. C. Klebaner. *Introduction to Stochastic Calculus with Applications*. Introduction to Stochastic Calculus with Applications. Imperial College Press, 2005.

[67] P. Kolesar. Stalking the endangered CAT: A queueing analysis of congestion at Automatic Teller Machines. *Interfaces*, 14(6):16–26, 1984.

[68] I. P. Kornfeld, S. V. Fomin, and Ya. G. Sinai. *Ergodic theory*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, 1982.

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

[70] Ł. Kruk, J. Lehoczky, K. Ramanan, and S. Shreve. *Double Skorokhod map and reneging real-time queues*, volume 4 of *IMS Collections*, pages 169–193. 2008. *Preprint*, to appear in a Festschrift Volume for Tom Kurtz.

[71] Ł. Kruk, J. Lehoczky, K. Ramanan, and S. Shreve. Heavy traffic analysis for EDF queues with reneging. *Ann. Appl. Probab.*, 21(2):484–545, 2011.

[72] G. Liang and U. C. Kozat. TOFEC: achieving optimal throughput-delay trade-off of cloud storage using erasure codes. In *INFOCOM, 2014 Proceedings IEEE*, pages 826–834, April 2014.

[73] T. Lindvall. *Lectures on the Coupling Method*. Dover Books on Mathematics Series. Dover, 1992.

[74] Malwina J. Luczak and Colin McDiarmid. On the maximum queue length in the supermarket model. *The Annals of Probability*, 34(2):493–527, 03 2006.

[75] Malwina J. Luczak and James Norris. Strong approximation for the supermarket model. *The Annals of Applied Probability*, 15(3):2038–2061, 08 2005.

[76] A. Mandelbaum and P. Momcilovic. Queues with many servers: The virtual waiting-time process in the QED regime. *Math. Oper. Res.*, 33(3):561–586, 2008.

[77] V. Marianov and D. Serra. Location models for airline hubs behaving as M/D/c queues. *Computers & Operations Research*, 30(7):983 – 1003, 2003.

[78] B. Maslowski and J. Seidler. Invariant measures for nonlinear SPDEs: Uniqueness and stability. *Archivum Math*, 34, 1998.

[79] J. C. Mattingly. On recent progress for the stochastic Navier-Stokes equations. *Journées Equations aux dérivées partielles*, pages 1–52, 2003.

[80] S. Mischler. Lecture notes of an introduction to evolution PDEs, November 2014.

[81] M. Mitzenmacher. Analyses of load stealing models based on families of differential equations. *Theory of Computing Systems*, 34(1):77–98, 2000.

[82] Michael Mitzenmacher. The power of two choices in randomized load balancing. *IEEE Trans. Parallel Distrib. Syst.*, 12(10):1094–1104, October 2001.

[83] G. Pang and W. Whitt. Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems*, 65(4):325–364, 2010.

[84] F. Pollaczek. U ber zwei formeln aus der theorie des wartens vor schaltergruppen. *Elektrische Nachrichtentechnik*, 8:256–268, 1931.

[85] A. A. Puhalskii and J. Reed. On many-server queues in heavy traffic. *Ann. Appl. Probab.*, 20(1):129–195, 2010.

[86] A. A. Puhalskii and M. I. Reiman. The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Advances in Applied Probability*, 32(2):pp. 564–595, 2000.

[87] J. Reed. The G/GI/N queue in the Halfin-Whitt regime. *Ann. Appl. Probab.*, 19(6):2211–2269, 2009.

[88] J. Reed and R. Talreja. Distribution-valued heavy-traffic limits for the GI/GI/∞ queue. *Annals of Applied Probability*, 25(3):1420–1474, 2015.

[89] L. C. G. Rogers and D. Williams. *Diffusions, Markov processes, and Martingales. Vol. 1*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000.

[90] L. C. G. Rogers and D. Williams. *Diffusions, Markov processes, and Martingales. Vol. 2*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000.

[91] M. Scheutzow. Chaining techniques and their application to stochastic flows. In J. Blath, P. Mörters, and M. Scheutzow, editors, *Trends in Stochastic Analysis*, pages 35–63. Cambridge University Press., 2009.

[92] A. Selvi and M. Rosenshine. A queueing system for airport buses. *Transportation Research Part B: Methodological*, 17(6):427 – 434, 1983.

[93] A. V. Skorokhod. *Integration in Hilbert space*. Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer-Verlag, 1974.

[94] A. L. Stolyar. Diffusion-scale tightness of invariant distributions of a large-scale flexible service system. *Adv. in Appl. Probab.*, 47(1):251–269, 03 2015.

[95] W. Szczotka. Tightness of the stationary waiting time in heavy traffic. *Adv. in Appl. Probab.*, 31(3):788–794, 1999.

[96] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich. A queueing system with a choice of the shorter of two queues—an asymptotic approach. *Problemy Peredachi Informatsii*, 32(1):20–34, 1996.

[97] D. F. Walnut. *An Introduction to Wavelet Analysis*. Applied and Numerical Harmonic Analysis. Birkhäuser Boston, 2013.

[98] J. B. Walsh. An introduction to stochastic partial differential equations. In P.L. Hennequin, editor, *École d'été de Probabilités de Saint Flour XIV - 1984*, volume 1180 of *Lecture Notes in Mathematics*, pages 265–439. Springer Berlin Heidelberg, 1986.

[99] W. Whitt. Approximations for the GI/G/m queue. *Production and Operations Management*, 2(2):114–161, 1993.

[100] W. Whitt. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2002.

[101] W. Whitt. Heavy-traffic limits for the G/H_2^* /n/m queue. *Math. Oper. Res*, 30:1–27, 2005.

[102] H. Q. Ye and D. D. Yao. Diffusion limit of a two-class network: Stationary distributions and interchange of limits. *SIGMETRICS Perform. Eval. Rev.*, 38(2):18–20, 2010.

[103] S. Zeltyn and A. Mandelbaum. Call centers with impatient customers: Many-server asymptotics of the M/M/n + G queue. *Queueing Systems*, 51(3):361–402, 2005.